Ferreira *et al. EURASIP Journal on Image and Video Processing* (2016) 2016:28
DOI 10.1186/s13640-016-0131-8

EURASIP Journal on Image and Video Processing

REVIEW Open Access

Towards key-frame extraction methods for 3D video: a review

Lino Ferreira[1,2,4*], Luis A. da Silva Cruz[2,3] and Pedro Assuncao[1,4]

Abstract

The increasing rate of creation and use of 3D video content leads to a pressing need for methods capable of lowering the cost of 3D video searching, browsing and indexing operations, with improved content selection performance. Video summarisation methods specifically tailored for 3D video content fulfil these requirements. This paper presents a review of the state-of-the-art of a crucial component of 3D video summarisation algorithms: the key-frame extraction methods. The methods reviewed cover 3D video key-frame extraction as well as shot boundary detection methods specific for use in 3D video. The performance metrics used to evaluate the key-frame extraction methods and the summaries derived from those key-frames are presented and discussed. The applications of these methods are also presented and discussed, followed by an exposition about current research challenges on 3D video summarisation methods.

Keywords: 3D key-frames extraction, 3D video summarisation, Shot boundary detection

1 Review

In the last years, new features have been implemented in video applications and terminal equipments due to users demand, who are always seeking for new viewing experiences more interactive and immersive, such as those provided by 3D video. This new visual experience is created by depth information that is part of 3D video and absent in classic 2D video. The inclusion of depth information in video signals is not a recent innovation, but the interest in this type of content and aspects related to it, such as acquisition, analysis, coding, transmission and visualisation, have been increasing recently [1, 2]. Lately, 3D video has been attracting attention from industry, namely content producers, equipment providers, distributors and from the research community mostly on account of the improvements in Quality of Experience that it provides to viewers [3], as well as due to the new business opportunities presented by this emerging multimedia format.

In the past, video repositories were relatively small so that indexing and retrieval operations were easy to perform. More recently, the massification of 3D video and its applications have resulted in the generation of huge amounts of data, increasing the need for methods that can efficiently index, search, browse and summarise the relevant information with minimum human intervention. Furthermore, 3D video description and management is also required to enable quick presentation of the most important information in a user-friendly manner [4, 5]. Video summarisation is a video-content representation method that can fulfil these requirements. In contrast to summarisation of 2D video, which has been the subject of a significant amount of research, 3D video summarisation is still a relatively unexplored research problem which deserves more attention.

A video summary is a short version of a full-length video that preserves the essential visual and semantic information of the original unabridged content. In the video summarisation process, a subset of key-frames or a set of shorter video sub-sequences (with or without audio) are chosen to represent the most important segments of the original video content according to predefined criteria [4]. This video content representation can be used in the promotion of movies, TV channels or other entertainment services. Video summarisation can also be used for

*Correspondence: lino.ferreira@ipleiria.pt
[1]Instituto de Telecomunicações (Leiria), Campus 2 Morro do Lena—Alto do Vieiro, 2411-901 Leiria, Portugal
[2]Dep. de Engenharia Electrotécnica e de Computadores, Universidade de Coimbra, Pólo II da UC, 3030-290 Coimbra, Portugal
Full list of author information is available at the end of the article

Springer Open

© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

content adaptation operations in constrained communication environments where bandwidth, storage capacity, decoding power or visualisation time is limited [6].

The literature defines two types of video summaries, namely those based on key-frames and those comprised of video skims [7]. A video summary based on key-frames is made up of a set of relevant frames selected from the video shots obtained from the original video. This type of summary is static, since the key-frames, being temporally distant and non-uniformly distributed, do not enable adequate rendering/reproduction of the original temporal evolution of the video content. Here, the video content is displayed in a quick and compact way for browsing and navigation purposes, without complying with timing or synchronisation requirements. Video skims are usually built by extracting the most relevant temporal segments (with or without audio) from the source sequence. After the extraction, all temporal segments are concatenated into sequential video with much shorter length than the source sequence. The methods used for computation of key-frames and video skims summaries are quite distinct, but these two types of representations for video content can be transformed from one to the other. A video skim can be generated from a key-frame summary by adding frames or segments that include the key-frames, while a key-frame summary can be created from a video skim by uniform sampling or selecting the most representative frame from each video skim segment [4].

In regard to 3D video content, a detailed study of the existing scientific literature reveals that comprehensive comparative studies of 3D video summarisation methods are missing. To help filling this gap, this paper presents a review of 3D video summarisation methods based on key-frames. This overview of the current state-of-the-art is mainly focused on the methods and features that are used to generate and evaluate 3D video summaries and not so much on the limitations or performance of specific methods. Since experimental set-ups, 3D video formats and features used for summarisation are considerably different from one computational method to another, a fair comparative analysis of the results, advantages and shortcomings of all methods is almost impossible. This paper also identifies open issues to be investigated in the area of 3D key-frame extraction for summarisation.

The remainder of the paper is organised as follows. Section 1.1 presents the existing 3D video representation formats and relevant features for the purpose of summarisation; then, in Section 1.3, the generic framework normally used in 3D key-frame extraction methods is presented, after which Section 1.4 reviews the most important shot boundary detection (SBD) methods for 3D video. Then, Section 1.5 characterises the relevant methods used in 3D key-frame extraction for summarisation while Section 1.6 addresses common methods used for presentation of key-frames. Section 1.7 describes performance evaluation methods suitable for 3D video summaries based on key-frames, and Section 1.8 describes some applications of this kind of summaries. Section 1.9 discusses the prospects and challenges of the 3D key-frame extraction methods, and finally Section 2 concludes the paper.

1.1 3D video representation formats

In this review article, '3D video', is defined as a representation format which differs from 2D video by the inclusion of information that allows viewers to perceive depth. This depth information can be conveyed either indirectly via two or more views of the scene (e.g. left and right views) or explicitly through either depth maps of geometric representation of connected 3D points and surfaces.

The most common formats used used to represent 3D visual scenes include natural video and/or geometric representations.

- *Stereoscopic video* is composed of two slightly shifted video views of the same scene, where one corresponds to what would be observed by a left eye and the other by the right eye of a human observer. Since these are two views of the same scene, the corresponding images are related by the binocular disparity, which refers to the difference in the image plane coordinates of similar features captured in two stereo images. The scene depth is perceived from the disparity when using stereoscopic displays and can also be computed for different types of computer vision applications (e.g. measuring distances in 3D navigation).
- *Multiview video (MVV)* is composed of more than two video views shifted in the vertical and/or horizontal position. Typically, MVV acquisition is done using an array of synchronised cameras with some spatial arrangement, which capture the visual scene from different viewpoints. The MVV format is useful for applications supported by autostereoscopic displays with or without head tracking, which render a denser set of 3D views that are displayed through lenticular and parallax barriers. With this type of display, viewers are able to see the portrayed scene from different angles by moving the head along a horizontal plane. A typical application of this video format is freeview navigation where users are given the option of freely choosing the preferred viewpoint of the scene.
- *Video-plus-depth (V+D)* is composed of a video signal (texture) and respective depth map. Each value of the depth map represents the distance of the object to the camera for the corresponding pixel position. Typically, the depth information is quantised with 8 bits, where the closest point is represented with value

255 and the most distant point is represented with 0. Additional virtual views (i.e. not captured) of the same scene imaged can be synthesised from the V+D original information by using 3D warping transformations. Several different applications and services can benefit from the V+D format, due to its inherent backward compatibility with 2D video systems and higher compression efficiency achievable when compared to stereoscopic video. For instance, 3DTV services can be seamlessly deployed while maintaining compatibility with legacy 2D video services.

- *Multiview video-plus-depth (MVD)* is composed of video and depth maps for more than two views of the specific scene. The depth information can be computed from different views or captured directly using time-of-flight (ToF) sensors. MVD can be used to support dense multiview autotereoscopic display in a relatively efficient manner. From a relatively small set of different views and corresponding depth maps, a much larger set of views can be synthesised at the display side, avoiding coding and transmission of a great deal of data while enabling smooth transitions between viewpoints. Several emerging applications such as free viewpoint video and free viewpoint TV will use the MVD format due to its compact representation of 3D visual information. Mixed-reality applications and gaming are also important application fields for MVD.
- *3D computer graphics* use a geometry-based representation, where the scene is described by a set of connected 3D points (or vertices), with associated texture/colour mapped onto them. The data content of this format can be organised into geometry, appearance and scene information [8]. The geometry includes the 3D position of vertices and polygons (e.g. triangles) that are constructed by joining these vertices. The appearance is an optional attribute which associates some properties (e.g. colour, texture coordinates) to the geometry data. Finally, the scene information includes the layout of a 3D scene with reference to the camera (or view), the light source and description of other 3D models if they are present in the scene. 3D computer graphics can provide better immersive and interactive experience than conventional 2D video, since the user is provided more freedom to interact with the content and get a realistic feeling of 'being there'. Relevant applications can be found in quite different fields, such as medicine, structural engineering, automobile industry, architecture and entertainment.
- *Plenoptic video* is composed of a very large of the number of views (e.g. hundreds or thousands) captured simultaneously. This multiple view acquisition process can be interpreted as a partial sampling of the plenoptic function [9], which represents not only spatial or temporal information but also angular information of about the captured light rays, i.e. captures a segment of the whole observable scene represented by a light field. In practice a 3D plenoptic image is captured by a normal image sensor placed behind an array of uniformly spaced semi-spherical micro-lenses. Each micro-lens works as an individual low resolution camera that captures the scene from an angle (viewpoint) slightly different from that of its neighbours. Plenoptic video, also known as light field video, is an emerging visual data representation with known applications in computational photography, microscopy, visual inspection and medical imaging among others.

1.2 3D video features for summarisation

The scene depth is the additional information that is either implicitly or explicitly conveyed by 3D video formats. Therefore, depth is also the signal component that mostly contributes to distinguish 3D video summarisation methods from those used for 2D video. One of the first works combining depth with features of 2D video to summarise 3D video was done by Doulamis et al. [10]. The authors proposed an algorithm jointly operating on both the depth map and the left channel image to obtain a feature vector for use in video segmentation and key-frame extraction methods. The feature vector including segment size, location, colour and depth. Another important feature is the depth variance associated with the temporal activity, which was used by Ferreira et al. in [11] for temporal segmentation of 3D video. The average stereo disparity per frame and temporal features such as image difference and histogram difference are computed and combined in feature vectors which in turn are used by a clustering algorithm to partitioning 3D video in temporal segments. Another work using frame intensity histogram distributions as features and the Jensen-Shannon difference to measure frame difference in feature space is presented in [12]. This is used to segment a video clip into shots, and then to choose key-frames in each one. More recently, Papachristou et al. in [13] also segmented 3D video using low-level features obtained from disparity, colour and texture descriptors computed from histograms and wavelet moments. An improved three-dimensional local sparse motion scale invariant feature transform descriptor is used in [14], for RGB-D videos, based on grey pyramid, depth pyramid and optical flow pyramids that are built for both colour frames and depth maps. Point features are determined with a SIFT descriptor are combined with depth information of the point as well as optical-flow derived motion information. Although this work is focused on gesture recognition, the features and similarity

measures may also be used in key-frame extraction methods. This work presented in [15], uses vectors containing moments (mean, standard deviation, skewness and kurtosis) of signature profiles of blocks with variable size for the luminance and disparity frames. A descriptor of frame moments was developed for summarising stereoscopic video through key-frame extraction and also to produce stereoscopic video skims. Yanwei et al. in [16] proposed a multiview video summarisation method which uses low-level and high-level features. The low-level features are based on visual attribute of video, as colour histogram, edge histogram and wavelet while for high-level features the authors use the Viola-Jones as face detector in each frame.

Geometric features are also relevant for temporal segmentation and extraction of key-frames in 3D visual information. For instance, Assa et al. in [17] proposed a method to produce an action synopsis of skeletal animation sequences for presenting motion in still images. The method selects key-frames based on skeleton joints and their associated attributes (joint positions, joint angles, joint velocities, and joint angular velocities). In [18], the authors also use geometric features, such as the number and location of vertices of surface, to produce video summaries of animation sequences. Other geometric features used by Jianfeng et al. in [19] to summarise 3D video are features vectors formed by the histograms of the vertices in the spherical coordinate system. A different type of features rely on 3D shape descriptors. For instance, Yamasaki et al. in [20] used the shape of 3D models to split the video in different motion/pose temporal segments. Relevant shape features, such as shape histograms, shape distribution, spin image and spherical harmonics were studied in [21], where the performance of shape similarity metrics is evaluated for applications in 3D video sequences of people. Since similarity measures are of utmost importance in summarisation this is a relevant work in the context of this paper. Another type of features used in [22] for key-frame extraction is based on deformation analysis of animating meshs while the vertices positions in mesh models and motion intensity were used by Xu et al. in [23] in temporal segmentation of 3D video.

1.3 3D key-frame extraction framework

Summarisation of 3D video follows a generic processing chain that is extended from 2D video by considering the inherent depth and geometric information as relevant feature contributors for selecting the dominating content in 3D moving scenes. A possible approach is based on clustering by grouping similar frames according to some similarity measure [24], without any prior processing or feature extraction. However, a more generic and systematic approach that better suits the problem of 3D video summarisation follows the three-step framework of Fig. 1,

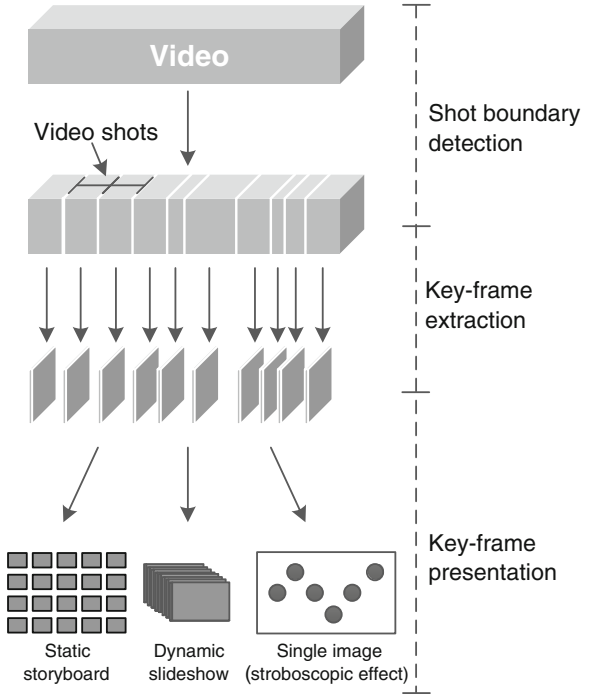

Fig. 1 A conceptual framework for key-frame summarisation

where the entire video sequence is first divided into video shots based on scene transitions using an SBD method, followed by a key-frame extraction method applied to each video shot to extract the most representative frames, based on the specific properties of the video content and similarity measures. Finally, the extracted key-frames are either presented to the viewers or stored, following some predefined presentation structure.

Following the conceptual framework shown in Fig. 1, the input video is segmented into video shots, mostly based on spatio-temporal criteria, but other criteria can be used such as based on motion [20, 25] or the combination of the temporal and depth features [11]. More details can be found in Section 1.4. After this segmentation, one or more key-frames are extracted from each video shot according to user-defined parameters, or based on specific requirements (in Fig. 1, only one key-frame is extracted). The most relevant key-frame extraction methods are presented in Section 1.5. Once the key-frames are extracted, they need to be presented in an organised manner for easy viewing during video browsing or navigation. In this framework, three key-frame presentation methods are described, static storyboard, dynamic slideshow and single image based on stroboscopic effect, but other methods can be found in the literature (see Section 1.6). The key-frame presentation methods are independent of the key-frame extraction operation and thus the same key-frame summary can be presented to viewers in different ways.

1.4 Shot boundary detection

In the recent past, development of SBD methods for 2D video received a lot of the attention from the video processing research community. However, very few works have investigated the SBD problem in the context of 3D video, especially taking into account depth information. Relevant surveys of video SBD methods with specific application in 2D video can be found in the literature [26–28]. In this section, we briefly introduce the main concepts behind these methods for 2D video. Then, the most promising and better performing SBD methods used for 3D key-frame extraction are explained in detail.

A video segment can be decomposed into a hierarchical structure of scenes, video shots and frames, with the linear video first divided into video scenes, which may comprise one or more video shots (set of correlated frames). A video scene is defined as a set of frames which is continuous and temporally and spatially cohesive [29], while a video shot may also be defined by camera operations, such as zoom and pan. Thus, the video shot is the fundamental unit in the content structure of a video sequence. Since its size is variable, the identification of start and end of the video shots is done using specific SBD methods.

Figure 2 presents a generic framework of a SBD method. While this framework is similar for both 2D and 3D video, the actual algorithms used for each type of content are not the same due to the difference in their relevant features. Firstly, the relevant visual features are computed, in general forming feature vectors for each video frame as described in Section 1.2. In the second step, the visual features of consecutive frames are compared using specific similarity measures some decision criteria are used to identify shot boundaries. The decision methods used to find shot boundaries can be based on static thresholds (as in Fig. 2), adaptive thresholds (thresholds depend on the statistics of the visual features used), B-splines fittings [30], support vector machines (SVM) [31] and K-means clustering [11]. The detection accuracy of SBD methods is improved by combining several visual features [32].

Video shot boundaries can be classified into two types: abrupt shot boundary (ASB) (as in Fig. 2) and gradual shot boundary (GSB), according to a certain classification of scene transition, which in general is related to content variation over time. This is common in 2D and 3D video, despite the fact that scene transition in 3D video may include depth changes besides the visual content itself. In ASB, the scene transition occurs over very few frames, usually a single frame defines the boundary. In the case of GSB, the transition takes place gradually over a short span of frames. The most common gradual transitions are fade-ins, fade-outs, dissolves and wipes [26–28]. A common problem in SBD is the correct discrimination between camera operations and object motion that originate the gradual transitions, since the temporal variation of the frame content can be of the same order of magnitude and take place over the same number of frames. This similarity of visual effects caused by camera operations and object motion can induce false detections of gradual shot boundaries. This problem is aggravated for video sequences with intense motion.

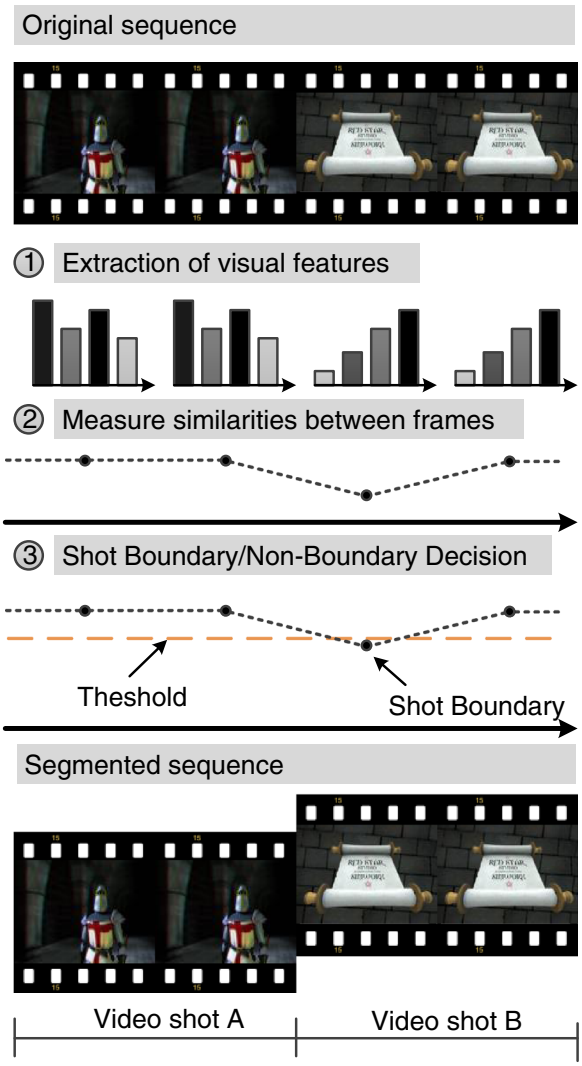

Fig. 2 A generic diagram of SBD framework

1.4.1 SBD methods

Doulamis et al. in [10] proposed a key-frame extraction method for stereo video which includes a SBD method. Here, the entire video sequence is divided into video shots using an algorithm based on the analysis of DC coefficients of compressed videos, following the solution proposed in [33]. More recently, Papachristou et al. in [13] presented a framework for stereoscopic video shot classification, that uses a well-known method designed for 2D video to segment the original stereoscopic video into shots

[34]. However, this method was applied only to the colour channels of the videos to be summarised. Ferreira et al. [11] proposed an algorithm to detect 3D shot boundaries (3DSB) based on a joint depth-temporal criterion. The absolute frame difference and sum of absolute luminance histogram difference are used as the relevant measures in the temporal dimension, while in the depth dimension, the variance of depth in each frame is used. A K-means clustering algorithm that does not require training and does not use thresholds is applied to choose the 3DSB transition frames. Ferreira's method is independent of the video content and can be applied to 2D or 3D video shot boundary identification. In the case of the 2D video, absolute frame differences and sum of absolute luminance histogram difference are used.

Some methods target segmentation of 3D mesh sequences using properties of 3D objects as the shape and motion/action (e.g. human body motion, raise hands) to detect the shot boundaries. Yamasaki et al. [20] proposed a temporal segmentation method for 3D video recordings of dances, which is based on motion speed, i.e. when a dancer/person changes motion type or direction, the motion becomes small during some short period and in some cases it is even paused for some instants, according to the type of dance. To seek the points where motion speed becomes small the authors used an iterative close point algorithm proposed in [35] which is employed in the 3D space (spherical coordinates). In contrast to conventional approaches based on thresholds, the authors devised a video segmentation scheme appropriate for different types of dance. In this scheme, each local minimum is compared with local maxima occurring before ($lmax_{bef}$) and after ($lmax_{aft}$) the local minima. When $lmax_{bef}$ and $lmax_{aft}$ are 1.1 times larger than the local minimum, a temporal segmentation point is declared to occur at the minimum location. Since the decision rule is not based on absolute values and thresholds, rather on relative values of extrema, it is more robust to data variation (like type of dance) and no empirically derived decision thresholds are used.

Another method which uses the motion speed of the 3D objects was presented by Xu et al. [23]. To reduce computation time of motion information the authors used the point distance (DP) instead of vertices position in Cartesian coordinates. DP is defined as Euclidean distance between one fixed point and all 3D objects' vertices coordinates of each frame. Figure 3 shows the point distance for 2 frames of Batter's sequence. Before determination of scene transitions, the histogram of point distance of each frame is calculated.

To detect abrupt and gradual transitions of 3D video, the Euclidean distance between the histograms of point distance and three thresholds are used, where the threshold values were derived empirically.

Fig. 3 Point distance of the frames #38 and #39 of Batter sequence. Grey values means the point distance from (0,0,0) [23]

Ionescu et al. [36] used a histogram-based algorithm specially tuned for animated films to detect ASB. From GSB only fades and dissolves are detected, since they are the most common gradual transition. The GSB detection is done using a pixel-level statistical approach proposed by [37]. The authors proposed the Short Colour Change (SCC) detection algorithm to reduce the cut detection false positives. The SCC is the effect that accompanies short term frame colour changes, caused by explosion, lightning and flash-like visual effects. More recently, Slama et al. [38] proposed a method based on the motion speed to split a 3D video sequence into segments characterised by homogeneous human body movements (e.g. walk, run, and sprint). However, the author only considers as significant video shot transition indicators changes of type of movement. Here, video shots with small differences from previous shots and small number of frames are avoided. The motion segmentation used in this work is based on the fact that when humans modify the motion type or direction, the motion magnitude decreases significantly. Thus, finding the local minimum of motion speed can be used to detect the break point where human body movements changes and consequently to segment the entire video into shots.

1.4.2 Evaluation metrics

Three well-known performance indicators are used in the evaluation of the SBD methods for 2D video: recall rate (R), precision rate (P) [39] and accuracy measure F1 [40]. The computation of these values is based on the comparison of manual segmentation (ground-truth) and computed segmentation. If a ground-truth is available these metrics can be applied to 3D video SBD methods.

Recall rate is defined as the ratio between the number of shot boundaries detected by an algorithm and the total number of boundaries in the ground-truth dataset (see Eq. (1)). Precision rate, computed according to Eq. (2), is defined as the ratio between the number of shot boundaries detected by an algorithm and the sum of this value

with the number of false positives. F1 is a measure that combines P and R, see Eq. (3).

$$R = \frac{D}{D + D_M} \tag{1}$$

$$P = \frac{D}{D + D_F} \tag{2}$$

$$F1 = \frac{2RP}{R + P} \tag{3}$$

where D is the number of shot boundaries correctly detected by the algorithm, D_M is the number of missed boundaries and D_F is the number of false detections. For good performance, the recall and precision rates should have values close to 1. The best performance is reached when F1 is equal to 1, while the worst occurs at 0.

The recall rate, precision rate and measure F1 were used to evaluate the performance of temporal segmentation methods for 3D video in [11, 23, 38], while Yamasaki et al. [20] only used recall and precision rates in the evaluation process. Although, these 3D SBD methods used the same evaluation metrics, the comparison of the results and performance obtained from such SBD methods is not possible because different datasets were used.

1.4.3 Discussion

Since the major difference between 2D and 3D video is the implicit or explicit availability of depth information, the visual features used in the SBD methods for 3D video must take depth into account, i.e. the temporal segmentation must also consider depth information in order to use depth discontinuities in shot detection. Until now, most research works on SDB for 3D video, did not use the depth information in the detection process. For example, Doulamis et al. in [10] proposed a key-frame extraction method for stereo video which includes a SBD method. However, this algorithm does not take into account the depth information of the stereo video and it is only applied to one view of the stereo sequence, for instance the left view. Another drawback of Doulamis' work is the lack of performance evaluation of the proposed temporal segmentation method. A method to segment stereo video was proposed in [13], but the proposed procedure does not take depth into account either.

In [20, 23, 38], the authors proposed SDB methods for 3D video, which are only applicable to 3D mesh models and require modifications to be used with most common pixel-based 3D video formats, like stereo or video-plus-depth. Finally, Ferreira et al. [11] proposed a method which uses the depth and temporal information for automatic detection of 3D video shots from the 3D video sequence that uses a K-means clustering algorithm to locate the boundaries. This algorithm has the advantage of not using any explicit thresholds or training procedure.

A common problem with the 2D video SBD methods described in the literature is the lack of common comparison grounds, as few works use the same dataset to test the methods proposed and evaluate their performance. This is a serious problem as it limits the number of comparisons that can be made to compare the different SBD methods. For the 3D case, the lack of comparative analyses is even more severe, due to the reduced number of SBD methods developed so far specifically for this type of visual information. The few works that have been proposed for SBD in 3D video usually use the Recall and Precision rate to evaluate performance, but the lack of benchmark 3D video sequences with ground-truth shot segmentations severely limit the number and types of performance evaluations that can be made. As mentioned above, the evaluation metrics presented in Section 1.4.2 are based on comparison between manual and computed segmentation. Therefore, besides being very important to have common test datasets, the development of universal and objective measures, which are specific for SBD and can be applied in different content domains and 3D video formats is highly recommended and desired.

1.5 3D key-frame extraction

In this section, we briefly introduce the main concepts behind key-frame extraction methods for 2D video and describe key-frame extraction methods for 3D video. The key-frame extraction methods under review are grouped into seven categories: non-optimised, clustering, minimum correlation, minimum reconstruction error (MRE), curve simplification, matrix factorisation and other methods.

1.5.1 Non-optimised methods

The simplest method for 3D key-frame summarisation is uniform sampling (UnS). This method selects key-frames at regular time-intervals (see Fig. 4a), e.g. selecting one video frame every minute to be a key-frame. This will result in a set of key-frames evenly distributed throughout the video. However, the selected key-frames might not contain meaningful or pertinent visual content or there may be two or more similar key-frames. For instance, the selected key-frame might show a bad image (e.g. unfocused) or no key-frame exists for some video shots, thus failing to effectively represent the video content.

Another simple and computationally efficient frame selection method is position sampling (PoS). In PoS, once the boundaries of a video shot are detected, the method selects frames according to their position in the video shot, and e.g. the first, or the last or the middle frame of the video shot (see Fig. 4b) can be chosen as key-frames. Thus, the size of key-frame summary corresponds to the number of video shots of the entire video. In some summarisation applications, one key-frame per video shot

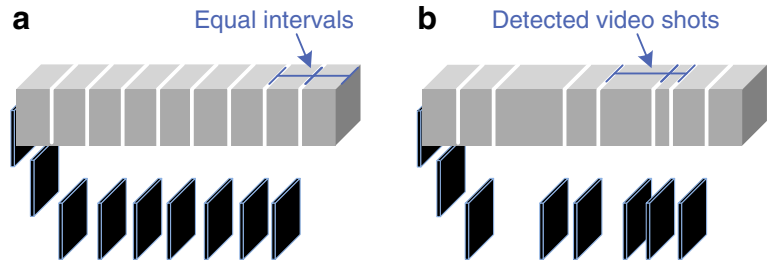

Fig. 4 a UnS method: uniform sampling at equal intervals. **b** PoS method: selecting the first frame of each video shot

is not enough, and the PoS method can be adapted to this need by allowing the selection of multiple frames at fixed positions within the video shot. For 3D video, UnS and PoS are used mostly as references for comparisons with other methods, as in [18, 41, 42]. Ionescu et al. [36] selected as key-frames the frames in the middle of the video shot to reduce temporal redundancy and computation cost of his animation movies summarisation method. Yanwei et al. [16] used the middle frame of each video skim segment to represent this summary in a storyboard. In general the above non-optimised methods may be used in both 2D and 3D video with minor adaptations.

1.5.2 Clustering

Clustering can be used to partition a large set of data into groups minimising intra-group variability and maximising inter-group separation. After partitioning, the data in each cluster have similar features. The partitioning can be based on the similarities or distances between the data points where each datum represents a vector of features of a frame. These points are grouped into clusters based on feature similarity and one or more points from each cluster are selected to represent the cluster, usually the points closest to the cluster centre. The representative points of the clusters can be used as key-frames for the entire video sequence. A significant number of clustering methods reported in the literature use colour histograms as the descriptive features, and the clustering is performed using distance functions such as Euclidean distances or histogram intersection measures. These methods are very popular due to its good clustering robustness and the simplicity inherent to computing colour histograms [43, 44]. Other features can also be used in clustering-based methods. For example, Ferreira et al. in [11] used temporal and depth features with a clustering algorithm to segment 3D video sequences into 3D video shots.

K-means is one of the simplest algorithms used to solve the clustering problem. This clustering algorithm can be applied to extract key-frames from short video sequences or shots, but its application to longer video sequences must be done with care taking into account the large processing time and memory requirements of the algorithm. To reduce the number of frames used by the clustering algorithm some authors pre-sample the original video, as proposed in [24]. The quality of the summaries may not be affected by this operation but the sampling rate must be chosen carefully. Although K-means is a popular and well-known clustering algorithm it has some limitations such as the need to pre-establish the number of clusters desired priori and the fact that the sequential order of the key-frames may not be preserved. Huang et al. [18] used the K-means clustering algorithm for extracting a set of 3D key-frames to be compared with the output of their key-frame extraction method.

1.5.3 Curve simplification

In the curve simplification method, each frame of the video sequence can be treated as a point in multidimensional feature space. The points are then connected in sequential order through an interpolating trajectory curve. The method then searches for a set of points which best represent the curve shape. Binary curve splitting algorithm [45] and discrete contour evolution [46, 47] are two curve simplification algorithms used in the key-frame extraction methods. Curve simplification-based algorithms preserve the sequential information of the video sequence during key-frame extraction; however, the search for the best representation curve has high computational complexity. The curve simplification method proposed in [48] was used by Huang et al. [18] in the evaluation process of the 3D key-frame extraction method they proposed.

1.5.4 Minimum correlation

Minimum correlation based methods extract a set key-frames such that the inter-key-frame correlation is minimal, i.e. it extracts the key-frames that are more dissimilar from each other. Formally, the optimal key-frame extraction based on minimum correlation can be defined as

$$K = \arg \min_{l_0, l_1, .., l_{n-1}} \text{Corr}\left(f_{l_0}, f_{l_1}, \ldots, f_{l_{n-1}}\right) \quad (4)$$

where Corr(.) is a correlation measure, n is the frame number of original sequence $\mathbf{F}$, l_i is the frame index

in **F** and K is the set of resulting key-frames with m frames, $K = \{f_{l_0}, f_{l_1}, \ldots, f_{l_{m-1}}\}$. Different algorithms can be used to find the optimal solution, such as logarithmic and stochastic search or a genetic algorithm [4]. The key-frame extraction method for stereoscopic video based on minimum correlation has been presented first by Doulamis et al. in [10]. Here, colour and depth information are combined to summarise stereo video sequences. After the segmentation of the entire video sequence, a shot feature vector is constructed based on size, location, colour and depth of each shot. To limit the number of shot candidates, a shot selection method based on similarly between shots is applied. Finally, the stereo key-frames are extracted from each of the most representative shots. The extraction is achieved by minimising a cross correlation criterion and uses a genetic algorithm [49]. Since this approach selects frames by minimisation of cross correlation, they are not similar to each other in terms of colour and depth.

1.5.5 Minimum reconstruction error

In MRE based methods, the extraction of the key-frames is based on the minimisation of the difference between the original video sequence/shot and the sequence reconstructed from the key-frames. A frame interpolation function $\mathcal{I}(t, K)$ is used to compute the frame at time t, of the reconstructed sequence, from a set of key-frames K. The frame-copy method can be used to reconstruct the video sequence/shot (i.e. performing zero-order interpolation), but more sophisticated methods like motion-compensated interpolation might be used as proposed in [50]. The reconstruction error $\mathcal{E}(\mathbf{F}, K)$ is defined as

$$\mathcal{E}(\mathbf{F}, K) = \frac{1}{n} \sum_{i=0}^{n-1} d\left(f_i, \mathcal{I}(i, K)\right) \tag{5}$$

where $d(.)$ is the difference between two frames, **F** is video sequence/shot with n frames, $\mathbf{F} = \{f_0, f_1, \ldots, f_{n-1}\}$, where f_i is the ith frame.

The key-frame ratio $R(K)$ defines the ratio between the number of frames in the set K, m and the total number of frames in video sequence/shot **F**, n, i.e. $R(K) = m/n$. Given a key-frame ratio constraint R_m, the optimum set of key-frames K^* is the one that minimises the reconstruction error, i.e.

$$K^* = \arg\min_{K \in \mathbf{F}} \mathcal{E}(\mathbf{F}, K) \quad s.t.\ R(K) \leq R_m \tag{6}$$

Thus, the MRE is defined by

$$MRE = \mathcal{E}(\mathbf{F}, K^*) \tag{7}$$

For example, given a shot **F** with $n = 10$ frames and a key-frame ratio $R(K) = 0.2$, this algorithm extracts at most 2 frames as key-frames, i.e. $m = 2$.

Xu et al. in [19] presented a key-frame extraction method to summarise sequences of 3D mesh models, wherein the number and location of key-frames are found through a rate-distortion optimisation process. As in all shot-based methods, in Xu's method shot detection is performed before key-frame extraction. Here, the SBD is based on the motion activity of a human body in dancing and sports videos. The motion activity is measured by the Euclidean distance between feature vectors of neighbouring 3D frames. The feature vectors are derived from three histograms (one for each spherical coordinate r, θ and ϕ) of all vertices of the 3D frames. Before the computation of spherical histograms, the Cartesian coordinates of vertices are transformed to the spherical coordinates. One of the three histograms is computed by splitting the range of the data in equal size bins. Then, the number of points from the data set that fall into each bin is counted. After shot detection, the key-frames are extracted in each shot. The key-frame extraction method is based on a rate-distortion trade-off expressed by a Lagrangian cost function, $\text{cost}(\text{Shot}_k) = \text{Distortion}(\text{Shot}_k) + \lambda \text{Rate}(\text{Shot}_k)$ where *Rate* is the number of key-frames in a shot and *Distortion* is the Euclidean distance between feature vectors.

Huang et al. [51] also presented a key-frame extraction method for 3D video based on rate-distortion optimisation, where *Rate* and *Distortion* definitions are similar to those used in [19]. However this method is not based on shot identification, since it produces 3D key-frame summaries without requiring prior video shot detection. The key-frame summary sought should minimise a Conciseness cost function, which is a weighted sum of the *Rate* and *Distortion* functions defined in the work. A graph-based method for extracting the key-frames is used, such that the key-frames selection is based on the shortest path in the graph that is constructed from a self-similarity map. The spherical histogram of the 3D frames is used to compute the self-similarity map.

More recently, Ferreira et al. [42] proposed a shot-based key-frame extraction method based on rate-distortion optimisation for 2D and 3D video. For each video-shot, a corresponding set of key-frames is chosen via dynamic programming by minimising the distortion between the original video shot and the one reconstructed from the set of key-frames. The distortion metric comprises not only information about frame difference, but also the visual relevance of different image regions as estimated by and aggregated saliency map, which combines three saliency feature maps computed from spatial, temporal and depth information.

1.5.6 Matrix factorisation

Another class of methods use matrix factorisation techniques to extract frames from a video sequence. Matrix factorisation (MF) techniques are based on approximating

a high dimension matrix **A** (original data) by a product of two or more lower dimension matrices. The **A** matrix can be composed of different features of the video or image, e.g. Gong and Liu [52] used the colour histograms to represent video frames while, Cooper et al. [53] computed the MF of the similarity matrix into essential structural components (lower dimension matrices). In addition to dimension reduction, the MF techniques allow reducing significantly the processing time and memory used during the operation. The MF techniques found in these key-frame extraction methods include singular value decomposition (SVD) and non-negative matrix factorisation.

Gong and Liu [52] proposed a key-frame extraction method based on SVD. To reduce the number of frames to be computed before the SVD, only a subset is taken from the input video at a pre-defined sample rate. Then, colour histograms (RGB) are used to create a frame-feature matrix **A** of the pre-selected frames. Next, the SVD is performed on matrix **A** to obtain an orthonormal matrix **V** in which each column vector represents one frame in the defined feature space. Then a set of key-frames are identified by clustering the projected coefficients. According to user's request, the output can be a set of key-frames (one of each cluster) or a video skim with a user specified time duration. To construct the set of key-frames, the frames that are closest to the centres of the clusters are selected as key-frames. Non-negative similarity matrix factorisation based on low-order discrete cosine transforms [53] and sliding-window SVD [54] are other approaches for key-frame extraction based on matrix factorisation.

In [18], Huang et al. proposed a method to be used with 3D video to represent an animation sequence with a set of key-frames. Given an animation sequence with n frames and m vertices of a surface in each frame, an $n \times m$ matrix **A** is built with the vertices coordinates. This matrix **A** is then approximately factorised into a weight $n \times k$ matrix **W** and a key-frame $k \times m$ matrix **H**, where k is the predefined number of key-frames. As k is selected to be smaller than n and m, this decomposition results in compact version of the original data $\mathbf{A} \approx \mathbf{WH}$. An iterative least square minimisation procedure is used to compute the weights and extract the key-frames. This procedure is driven by user-defined parameters such as a number of key-frames and an error threshold. Lee et al. [22] introduced a deformation-driven genetic algorithm to search good representative animation key-frames. Once the key-frames are extracted, similar to [18], the animation is reconstructed by a linear combination of the extracted key-frames for better approximation. To evaluate the performance of the proposed method, the authors compare it with Huang's method proposed in [18].

1.5.7 Other methods

The methods described in this section could not be easily classified into the preceding categories, mostly on account of the diversity of approaches followed in solving the key-frame extraction problem. As such, and given their importance, they are described all together in this section.

Assa et al. proposed a method to create an action synopsis image composed of key poses (human body motion) based on the analysis through motion curve. The method integrates several key-frames into a single still image or a small number of images to illustrate the action. Currently, it is applied in 3D animation sequences and 2D video as documented in [17].

Lee et al. [25] proposed a method to select key-frames from 3D animation video using the depth information of the animation. The extracted key-frames are used to compose a single image summary. The entire video sequence is divided into temporal segments based on the motion of the slowest moving objects, and then a summarisation method is applied to the segments. The depth information and the respective gradient (computed with depth values of each frame) is used to compute the importance of each frame. A single image summary composed of several foreground visual objects is built based on the importance of each frame. The authors proposed a threshold based approach to control the visual complexity (number of foreground objects) of the single image summary (one for each video sequence), as it is showed in Fig. 5. By using this approach, the number of video frames to be analysed is reduced, but in some cases the method can miss important information contained in the temporal segments.

Jin et al. [41] proposed a key-frame extraction method for animation sequences (skeletal and mesh animations). The method uses animation saliency computed from the

Fig. 5 Single image key-frame presentation method [25]

original data to aid the selection of the key-frames that can be used to reconstruct the original animation with smaller error. Usually, an animation sequence is characterised by a large amount of information. For computational efficiency, the animation sequence is projected to a lower-dimensional space where all frames of the sequence are represented as points of curves defined in the new lower-dimensional space. Then, the curves in the lower-dimensional space are sampled and these sampled points are used to compute the Gaussian curvature values. Next, the points with the largest curvature value are selected as candidate key-frames. Finally, a key-frame refinement method is employed to minimise an error function which incorporates visual saliency information. The aim of a visual saliency is to identify the regions of an image which attract higher human visual attention. Lee Huang et al. [55] expanded this idea to 3D video and computed mesh saliency for use in a mesh simplification algorithm that preserves much information of the original input. More recently, visual saliency has also been used in 3D key-frame extraction, in the method proposed by Ferreira et al. in [42].

Yanwei et al. [16] proposed a multiview summarisation method for non-synchronised views, including 4 of them covering 360°, which results in small inter-view correlation, thus more difficult to compute similarity measures. In this method, each view is segmented into video shots and general solution combines features of different shots and uses a graph model for the correlations between shots. Due to the correlation among multi-view shots, the graph has complicated connectivity, which makes summarisation very challenging. For that purpose, random walks are used to do shot clustering and then the final summary is generated by a multi-objective optimisation process based on different user requirements, such as the number of shots, summary length and information coverage. The output of Yanwei's method is a multiview storyboard, condensing spatial and temporal information.

1.5.8 Discussion

The problem of key-frame extraction for 3D video has been presented first by Doulamis et al. in [10] who proposed a method combining colour and depth information to summarise stereo video sequences. Papachristou et al. in [13] developed a video shot classification framework for stereoscopic video, in which the key-frame extraction method used is based on mutual information. Even though the framework was proposed for stereoscopic video, the key-frame extraction method only uses one view of the stereoscopic video. Until now, only some specific 3D video formats were considered by the existing key-frame extraction methods. Stereoscopic video was used in [10, 42], V+D is used by Ferreira et al. in [42] and 3D computer graphics format in [17–19, 22, 25, 51]. Thus, further room exists for research on efficient key-frame extraction methods that can be applied to other 3D video formats, such as MVV, MVD and holoscopic video.

Most 3D key-frame extraction methods cited in this paper were developed for specific content and 3D format and only four of them include comparisons with similar methods [18, 22, 41, 51]. In [18], curve simplification, UnS and clustering methods were utilised as reference methods for performance evaluation and comparison of the proposed matrix factorisation methods. The authors showed that the method based on matrix factorisation extracts more representative key-frames in comparison with the other three competing methods [22, 41, 51]. However, the algorithm is very slow with quadratic running time complexity. In [22], the proposed method based on genetic algorithm is compared with Huang's method [18] in terms of the PSNR and computational complexity. The former is very efficient in terms of computation time when compared to the latter but qualitywise (average PSNR) it is slightly worse. However, Huang's method [18] is slightly better when comparing maximum and minimum PSNR.

Peng Huang et al. in [51] confront their key-frame extraction method with the method used in [19] and the results show improved performance for all 3D video sequences used. Jin et al. in [41] compare the proposed method with the UnS and Principal Component Analysis methods [56]. The results show that the proposed method achieves much better reconstruction of skeletal and mesh animation than the other methods under analysis.

As mentioned before, most of the key-frame extraction methods for 3D video, rely on a previous SBD step. However, the methods just described, from [18, 22, 41, 51], do not perform any pre-analysis of the video signal to identify shots and their boundaries. The quality of key-frame summaries obtained using these approaches can be negatively affected when accurate shot segmentation is not available. Another important issue is the definition of the number of key-frames need to represent the original sequence. This number depends on user requirements and on the content of the video to be summarised and its choice frequently involves a trade-off between the quality and efficiency of the key-frame summary.

1.6 Key-frame presentation

Once the key-frames are extracted, they need to be presented in an organised manner to facilitate video browsing and navigation operations by the user. The video presentation methods aim to show the key-frames in some meaningful way allowing the user to grasp the content of a video without watching it from beginning to end [4]. The most common methods for key-frame presentation are the static storyboard, dynamic slideshow and single image, see Fig. 1.

Static storyboard presents a set of miniaturised key-frames spatially tiled in chronological order, allowing a quick browsing and viewing of the original video sequence. This presentation method was used with 3D video in [10, 18, 19, 22, 41, 51]. The second method is the dynamic slideshow that presents the key-frames one by one on the screen, which allows browsing over the whole video sequence. Other presentation method is the single image, which morphs parts of different key-frames in chronological to produce a single image. Normally, in this presentation type the background and foreground objects (time shifted) are aggregated in single image, as exemplified in Fig. 6. In this figure, the foreground is the children who plays in bars of a playground. Here, we can see 3 positions of the children in the bars which corresponds to 3 key-frames of video sequence.

Qing et al. [12] proposed a generic method for extracting key-frames in which the Jensen-Shannon divergence is used to measure the difference between video frames to segment the video into shots and to choose key-frames in each shot. The authors also proposed a 3D visualisation tool, used to display key-frames and the useful information related to the process of key-frame selection. More recently, Nguyen et al. [57] proposed the Video Summagator. This method provides a 3D visualisation of a video cube of static and dynamic video summaries. Assa et al. proposed a method to create an action synopsis image from a 3D animation sequence or 2D video [17]. Lee et al. also proposed a method to summarise a 3D animation into a single image based on depth information [25].

In [58] a 3D interface (3D-Ring and 3D-Globe) was proposed as an alternative to the 2D grid presentation for interactive item-search in visual content databases, see Fig. 7. Even though this system was designed to be used with a large database it can also be applied to visualise key-frames summaries of 2D and 3D video.

Fig. 6 Video synopsis proposed [43]

Fig. 7 a 3D-Ring interface, **b** 3D-Globe interface and **c** 2D grid presentation (figures based on [58])

1.6.1 Discussion

Most of the 3D key-frame extraction methods proposed in the literature until now are focused on the extraction rather than in the presentation of key-frame sets to the viewers. So far only Assa et al. and Lee et al. proposed in [17] and [25] two presentation solutions distinct from the static storyboard used in association with most of 3D key-frame extraction methods [10, 18, 19, 22, 41, 51]. In this scenario, with only two presentation solutions, it is foreseeable that the development of new 3D video and image display devices will lead to the creation of new methods to display 3D video summaries or key-frame collages providing the user with more immersive and more meaningful ways to observe these types of time-condensed video representations.

1.7 Quality evaluation of 3D key-frame summaries

One of the most important topics in video summarisation algorithmic development is the evaluation of the key-frame extraction methods. In this section, we present current key-frame summary evaluation methods and some

related issues. These methods are aggregated into three groups: result description, subjective and objective methods, as it was proposed in [4].

1.7.1 Result description

This is the most common and simple form of evaluation key-frame extraction methods since it does not require a reference, either for objective or subjective comparison with other methods. Usually, it is used to explain and describe the advantages of some method compared with others based on presentation or/and description of the key-frames extracted (visual comparison), as in [18, 19, 22, 25, 41, 51]. This type of evaluation can also be used to discuss the influence of specific parameters or features of the method and also the influence of the content in the key-frame set, as in [10, 19]. In some works, this type of evaluation method is complemented with objective and/or subjective methods as in [19, 25]. However, the result description method has some limitations, such as the reduced number of methods which can be compared at same time, i.e. it is inadequate to compare key-frame summaries of a large number of video sequences or methods. Another drawback is the subjectivity inherent to this type of evaluation, since the underling comparisons results are usually user-dependent and so prone to inter and intra observer fluctuations.

1.7.2 Subjective methods

Subjective methods rely on the independent opinion of a panel of users judging the quality of the generated key-frame video summaries according to a known methodology. In this type evaluation, a panel of viewers are asked to observe both the summaries and the original sequence and then respond to questions related to some evaluation criteria, (e.g. 'Was the key-frame summary useful?', 'Was the key-frame summary coherent?') or if each key-frame is 'good', 'fair', or 'poor' according to the original video sequence.

The experiments can include a set of absolute evaluations and/or a set of relative evaluations, in which two key-frame summaries are presented and compared. Usually, the summary visualisation and rating steps are repeated for each video in the evaluation set by each viewer. During the evaluation of the key-frame summaries, it is also required taking into account the external factors which can influence the ratings of the summaries, such as the attention and fatigue specially when there are long evaluation sessions with many video summaries. In addition to these factors, the experiments must follow standard recommended protocols prepared specifically for subjective assessment of video quality [59].

Subjective evaluation methods were used in [16, 60–63]. In [60], subjective assessment was used to grade the single key-frame representations as 'good', 'bad' or 'neutral' for each video shot and also give appreciations on the number of key-frames with possible grades being 'good', 'too many' and 'too few' in the case of multiple key-frames per shot. In [61, 63], the quality of the key-frame summary is evaluated by asking users to give a mark between 0 to 100 for three criteria, 'informativeness', 'enjoyability' and 'rank' after watching the original sequences and the respective key-frames summaries. Ejaz et al. [62] used subjective evaluations to compare the proposed method with four prominent key-frame extraction methods: open video project (OV) [45], Delaunay triangulation (DT) [64], STIll and MOving Video Storyboard (STIMO) [65] and Video SUMMarisation (VSUMM) [24]. In this case, the evaluation is based on mean opinion scores (MOS) and viewers are asked to rate the quality of the key-frame summary on scale of 0 (minimum value) to 5 (maximum value) after watching the original sequences and the respective summaries generated by all the methods.

In [16] subjective assessments were also used to evaluate multiview video summaries. The aim is to grade the 'enjoyability', 'informativeness' and 'usefulness' of the video summary. Here, three questions were put to the viewer to evaluate the method: Q_1: 'How about the enjoyability of the video summary?' Q_2: 'Do you think the information encoded in the summary is reliable compared to the original multiview videos' and Q_3: 'Will you prefer the summary to original multiview videos if stored in your computer?'. In reply to the questions Q_1 and Q_2, the viewers assigned a score between 0 (minimum value) to 5 (maximum value) and for Q_3 the viewers only need to respond with 'yes' or 'no'. From all 3D key-frame extraction methods reviewed, only [16, 17, 25] used subjective evaluations.

1.7.3 Objective methods

Although subjective evaluation provides a better representation of the human perception than objective evaluation, it is not suitable for practical implementations due to the time required to conduct the opinion collection campaigns. Objective evaluation methods are reproducible and can be specified analytically, and since they are automatable can be used to rate the proposed method on large number of videos of variable genres and formats. These methods can be applied to all types of video formats without requiring the services of video experts and can be performed rapidly and automatically if suitable quality measures are available. Besides being faster, simpler and easily replicable, this type of method is more economical than the subjective evaluation.

The works reviewed in this article, which use objective quality evaluation, employ several quality measures originally developed for 2D video, but can be also applied to 3D video, after being modified to take into account the specific features of 3D visual information. The shot

reconstruction degree (SRD) distortion measure [66] and the fidelity measure (Fm) defined in [67] follow two different approaches. Fidelity measure employs a global strategy, while SRD uses a local evaluation of the key-frames. To judge the conciseness of a key-frame summary a measure of the Compression Ratio (CR) is used [68]. If a ground-truth summary is available the Comparison of User Summaries (CUS) [24], Recall rate, Precision rate and accuracy measure (F1) measures can be used. These measures compare the computed summaries with those manually built by users. More details on these measures are presented in the next sub-sections.

Shot reconstruction degree: SRD measures the capability of a set of key-frames to represent the original video sequence/shot. Assuming a video shot $\mathbf{F} = \{f_0, f_1, \ldots, f_{n-1}\}$ of n frames and $K = \{f_{l_0}, f_{l_1}, \ldots, f_{l_{m-1}}\}$ a set of m key-frames selected from **F**, the reconstructed scene shot $F' = \{f'_0, f'_1, \ldots, f'_{n-1}\}$ is obtained from the K set by using some type of frame interpolation. The SRD measure is defined as

$$\mathrm{SRD}(\mathbf{F}, F') = \frac{1}{n}\sum_{k=0}^{n-1} \mathcal{S}im\left(f_k, f'_k\right) \tag{8}$$

where n is the size of the original video sequence/shot **F** and $\mathcal{S}im(.)$ is the similarity between two video frames. In Liu et al. [66], the similarity measure chosen was peak signal-to-noise ratio (PSNR), but other similarity metrics that include 3D features can also be used in the evaluation of 3D key-frame summaries. A K key-frame summary is a good representation of the original **F** when the magnitude of its SRD is high.

Fidelity Fm is computed as the maximum of the minimal distances between the set of key-frames K and each frame of the original **F**, i.e. a Semi-Hausdorff distance d_{sh}. Let **F** be a video sequence/shot containing n frames, and the set $K = \{f_{l_0}, f_{l_1}, \ldots, f_{l_{m-1}}\}$ of m frames, selected from **F**. The distance between the set K and a generic frame f_k s.t. $0 \leq k \leq n-1$ belonging to **F** can be calculated as follows.

$$d_{\min}(f_k, K) = \min_j \left\{ d\left(f_k, f_{l_j}\right) \right\}; j = 0, 1, \ldots, m-1 \tag{9}$$

Then the semi-Hausdorff distance d_{sh} between K and **F** is defined as

$$d_{sh}(\mathbf{F}, K) = \max_k \{d_{\min}(f_k, K)\}; k = 0, 1, \ldots, n-1 \tag{10}$$

The Fidelity measure is defined as

$$Fm(\mathbf{F}, K) = \mathrm{MaxDiff} - d_{sh}(\mathbf{F}, K) \tag{11}$$

where *MaxDiff* is the largest possible value that the frame difference measure can assume. The function $d(f_a, f_b)$ measures the difference between two video frames a and b. The majority of the existing dissimilarity measures can be used for $d(,)$, such as the L_1-norm (city block distance), L_2-norm (Euclidean distance) and L_n-norm [67]. As it was mentioned before, the Fm measure can be used for 3D video with the necessary changes in the $d(,)$ distance. Whenever Fm is high, this means that the selected key-frames provide an accurate representation of the whole **F**.

Compression ratio: A video summary should not contain too many key-frames since the aim of the summarisation process is to allow viewers to quickly grasp the content of a video sequence. For this reason it is important to quantify the conciseness of the key-frame summary. The conciseness is the length of the key-frame video summary in relation to the original video segment length and can be measured a compression ratio, defined as the relative amount of 'savings' provided by the summary representation:

$$CR(\mathbf{F}) = 1 - \frac{m}{n} \tag{12}$$

where m and n are the number of frames in the key-frame set K and the original video sequence **F** respectively. Generally, high compression ratio is desirable for a compact video summary [68].

Comparison of user summaries (CUS): CUS is a quantitative measure based on the comparison of summaries built manually by users and computed summaries. It was proposed by Avila et al. in [24]. The user summaries are taken as reference, i.e. the ground-truth, and the comparison between the summaries is based on specific metrics. The colour histogram is used for comparing key-frames from different video summaries, while the distance between them is measured using the Manhattan distance. Two key-frames are similar if the Manhattan distance of their colour histograms is below a predetermined threshold δ. In [24], this threshold value was set to 0.5. Two evaluation metrics, accuracy rate CUS_A and error rate CUS_E, are used to measure the quality of the computed summaries. They are defined as follows:

$$\mathrm{CUS}_A = \frac{n_{\text{match}}}{n_{\text{US}}} \quad \mathrm{CUS}_E = \frac{n_{\text{no-match}}}{n_{\text{US}}} \tag{13}$$

where n_{match} and $n_{\text{no-match}}$ are, respectively, the number of matching and non-matching key-frames between the computed and the user generated summary and n_{US} is the total number of key-frames in the summary. CUS_A varies between 0 and 1, where $\mathrm{CUS}_A = 0$ is the worst value indicating that none of the key-frames from the computed summary matches those of the user summary. A value of $\mathrm{CUS}_A = 1$ is the best case and indicates that all key-frames from both summaries perfectly match each other. A null value for CUS_E indicates a perfect match between both summaries.

Computational complexity: Another relevant performance metric taken into account in the evaluation of key-frame extraction methods is the computational complexity, which is usually equated with the time spent to construct a key-frame summary. This metric was used in [24, 62, 63, 68, 69] for 2D video summaries. In 3D key-frame extraction methods, the computational complexity metric is only used by Lee et al. in [22], where the computational complexity of Lee's and Huang's et al. [18] methods are compared.

Other methods: Other methods and measures were used for objective evaluation of 3D key-frames summaries. In [19, 51] a rate-distortion curve is used, modelling a monotonic relationship between rate and distortion, with increases of the former leading to decreases of the latter. In the case of [18], the root mean square error (RMSE) distance between the original and reconstructed animation was used as the objective quality measure (with an inverse relationship in this case). This measurement is the same as in [70] and [71]. Lee et al. [22] used PSNR to measure the reconstruction distortion. Jin et al. in [41] measure reconstruction error of the animation from the extracted key-frames, using average of degrees of freedom (DOF) of reconstruction error magnitude.

1.7.4 Discussion

Conciseness, coverage, context and coherence are desirable attributes in any key-frame summary. Some of these attributes are mostly subjective as is the case of the context and coherence. Conciseness is related to the length of the key-summary, while the coverage evaluation is based on comparison between computed key-frames summary and ground-truth summary, expressed by the recall rate, precision rate, CUS_A and CUS_E.

Most evaluation metrics reviewed above were developed for 2D video. However, some of them, such as Fm and SRD, have also been extended to evaluate 3D video summaries after some adaptation. This is the casse of the 3D key-frame extraction method presented by Ferreira et al. in [42], where the Fm and SRD metrics were used. To measure the Recall rate, Precision rate, CUS_A, CUS_E, computational complexity and compression ratio in 3D video summarisation, no adaptation is needed.

The key-frame extraction methods are often application-dependent (e.g. summarisation of sports videos, news, home movies, entertainment videos and more recently for 3D animation) and the evaluation metrics must be adapted to the intended use. A good summary quality evaluation framework must be based on a hybrid evaluation scheme which includes the strengths of subjective and objective methods and also the advantages of result description evaluations.

1.8 Applications

In this section, some applications of 3D key-frame extraction methods and some aspects related to these applications are presented. These applications are grouped into five categories: video browsing, video retrieval, content description, animation synthesis and others.

1.8.1 Video browsing

The video browsing and associated problem has been investigated by the research community for decades, [72]. However, the growing use of 3D video and the specific characteristics of this type of visual information make 3D video browsing a more interesting and challenging problem. The access to databases or other collection of videos could be eased by the use the key-frame extraction methods to abstract/resume long video sequences in the repository of interest. With this kind of abridged video representation, a viewer can quickly find the desired video in a large database. For example, once an interesting topic has been identified through display of the key-frames, a simple operation as a click on the respective key-frame can initiate video playback of the original content at that particular instant. Many video browsing methods have been proposed for 2D video [72]. However, to the best of the authors' knowledge, in the case of 3D video there are no works reported in the literature.

1.8.2 Video retrieval

In contrast to video browsing, where viewers often just browse interactively through video summaries in order to explore their content, in video retrieval the viewers search for certain visual objects (e.g. objects, people and scenes) in a video database. In this type of retrieval processes, viewers are typically expected to know exactly what they are looking for. Therefore, it is crucial to implement appropriate search mechanisms for different types of queries provided by distinct viewers and with particular interests. The matching between the viewers' interests (queries) and the database content can be made with recourse to textual or image based descriptions or combinations of both. Some 2D video search and retrieval applications have combined video browsing and retrieval in the same platform [72]. In the case of 3D video this problem is still open for research. Finally, it is worth to point out that work done on 3D object recognition techniques which can also be used in retrieval, as published in [73–75].

1.8.3 Content description

Vretos et al. [76] presented a way of using the audio-visual description profile (AVDP) of the MPEG-7 standard for 3D video content description. The description of key-frames is contemplated in the AVDP profile through the MediaSourceDecompositionDS (i.e. MediaSourceDecompositionDS is used in the AVDP context to

decompose an audiovisual segment into the constituent audio and video channels). Thus this content description scheme, allows that 3D key-frames can be used for fast browsing and condensed representation of query results of 3D video search tasks. Other application of key-frames to content description was proposed by Sano et al. [77]. Here, the authors proposed and discussed how the AVDP profile of the MPEG-7 can be applied to multiview 3D video content [56].

1.8.4 Animation synthesis

Blanz et al. [78] proposed a morphable 3D face model by transforming the shape and texture of example into a new 3D model representation. According to this modelling approach, new or similar faces and expressions can be created by forming linear combinations of the 3D face models. A similar concept to the proposed in [78] can be applied to generate 3D models [79] or to synthesise new motion from captured motion data [80]. Animation synthesis based on key-frames [81] using the same concept has been presented in [78–80], to interpolate frames between two key-frames. However, the quality of the interpolated frames is dependent on the inter-key-frame distance and on the interpolation method used.

1.8.5 Others

Assa et al. [17] proposed the use of action synopsis images as icons (personal computer desktop and folders) and thumbnails of the 3D animation. Assa et al. also proposed an automatic or semi-automatic generation method to create comic strips and storyboards for 3D animation. Lee et al. [25] presented a method to create a single image summary of a 2D or 3D animation, which can be used in the same application as Assa's work. Halit et al. [56] proposed a tool for thumbnail generation from motion animation sequences. Several authors, as [82–85] have used key-frame extraction methods in the 2D-to-3D video conversion.

1.9 Prospects and challenges

Although some significant work has been done in the 3D video summarisation domain, many issues are still open and deserve further research, especially in the following areas.

1.9.1 SBD and key-frame extraction methods

The selection of the features used by shot boundary and key-frame extraction methods is still an open research problem, because these features depend on the application, video content and representation format. For instance, in fast-motion scenes edge information is not the best choice to detect shot boundaries due to motion-induced blur. Thus, it may be better to automatically find the useful features based on some assumptions about the video-content.

The majority of key-frame extraction methods published in the literature use low-level features and content sampling approaches to identify the relevant frames that should be included in the key-frame summary. Recently, the inclusion of perceptual metrics in the SBD and key-frame methods are gaining some space. Recently and in the context of 2D video, some key-frame extraction methods based on visual attention models have emerged as, [60–63, 86]. However, for 3D video only two solutions are available [41, 42]. Hence, key-frame extraction in 3D video still poses relevant research problems to be investigated and efficiently solved.

Another open challenge is the combination of the visual features with additional information, such as audio features, text captions and content description, for use in shot boundary detection and selection of the optimal frames in 3D video. In the current literature, there is also a lack of summarisation methods based on key-frames or video skims, for the most recent 3D video formats such as MVD and plenoptic video. Another topic open to further research is the application of scalable summarisation to 3D formats [87]. Despite the fact that several previous works addressed scalable summarisation for 2D video, e.g. [88, 89], such methods were not extended to 3D and multiview, which leads to open research questions.

1.9.2 Evaluation

In the past evaluation frameworks for 2D key-frame summarisation methods were proposed in [90, 91]. More recently, Avila et al. [24] also proposed another evaluation setup, wherein the original video and the key-frame summaries of several methods are available for downloading, together with the results of several key-frame extraction methods for 2D video. Unfortunately for the case of 3D video, there is not as yet any similar framework, where key-frame summaries and the respective original sequences are available for research use.

The number and diversity of evaluation metrics (objective, result description and subjective) used to compare state-of-the-art key-frame extraction methods make their comparative assessment a difficult task. Therefore, the development of metrics which can be used in the evaluation of key-frames summaries in different domains and 3D video formats is a very important area of video-summarisation related research. Furthermore the focus of the evaluation process must be application-dependent. For instance, in browsing applications, the time spent by the user to search or browse for a particular video is the most important factor, but on the other hand, in detection events, the evaluation metric must focus on the successful detection of these events.

One other problem that arises in the evaluation process is the replication of results of previous works, as

some works are not described with enough details to allow independent implementation or the input data is unavailable or else it is not easy to use due to data format incompatibilities or lack of information about their representation format. Thus, the best way to test and compare key-frame extraction methods for 2D and 3D is to build publicly accessible repositories containing test kits, made up of executable or web-executable versions of the methods and the test sequences.

1.9.3 Presentation

Another challenging topic in the research of 3D key-frame summarisation is the design of an efficient and intuitive visualisation interface that allows easy navigation and visualisation of the key-frame summaries. These applications should be independent of the terminal capabilities (display dimension, processing and battery power), i.e. should be usable on small screen devices such as smartphones as well as on ultra-high-definition displays. In addition, the visualisation interface should be independent from the key-frame summarisation method, to allow the visualisation of different formats of 3D key-frames video summaries, such as stereoscopic video or video-plus-depth and also 2D video in the same framework. The interface should be capable of dealing with the most common key-frame visualisation methods such as, static storyboard, dynamic slideshow and hierarchically arranged viewing. In particular, the most recent 3D interface for searching and viewing images or video in large databases, 3D-Ring and 3D-Globe, are interesting solutions which must be taken into account in the definition of new key-frame visualisation methods [58].

1.9.4 Video summary coding

In the past, the problem of scalable coding of video summaries was addressed in [88, 92–94]. In [92] the authors propose a hierarchical frame selection scheme which considers semantic relevance in video sequences at different levels computed from compressed wavelet-based scalable video. In [93], a method to generate video summaries from scalable video streams based on motion information is presented; while in [94], the authors propose to partition a video summary into summarisation units related by the prediction structure and independently decodable. Ferreira et al. in [88] proposed a method to encode an arbitrary video summary using dynamic GOP structures in scalable streams. The scalable stream obtained was fully compatible with the scalable extension of the H.264/AVC standard. However, all approaches were proposed for 2D video and used older generation video coding methods. The application of video summary coding to the 3D video format and the use of the most recent video coding, such as HEVC, should also be explored to find efficient coding tools for such purpose.

2 Conclusions

In this paper, we have presented a review of 3D key-frame extraction methods covering the major results published in recent journal issues and conference proceedings. Different state-of-the-art methods for key-frame extraction and evaluation metrics were presented and examined. The most important presentation methods for key-frame summaries were also discussed.

Various suggestions for the development of future 3D video summarisation methods are made, particularly oriented for future research on 3D key-frame extraction methods and potential benefits that may be attained from further research based on visual attention models. So far, 3D video key-frame extraction methods based on visual attention have not been deeply researched, so this is an interesting point to be explored. More research effort should also be put on methods for performance evaluation of key-frame extraction algorithms. The current plethora of different objective and subjective evaluation methods, most of them not easily comparable between each other, motivates a research goal towards unified and comparable methods for performance evaluation and benchmarking of 3D video summaries.

One other important and interesting research topic is the design and implementation of methods and tools to present 3D key-frame summaries. It is clear that the way a key-frame set is presented to viewers influence the time and effort they have to devote to interpret the summarised visual data. Finally, efficient coding of video summaries also leads to research problems which are still open for further research, since no specific solutions for 3D video are currently available.

Acknowledgements

This work was supported by the R&D Unit UID/EEA/ 50008/2013, Project 3DVQM and PhD Grant SFRH/ BD/37510/2007, co-funded by FEDER-PT2020, FCT/ MEC, Portugal.

Authors' contributions

LF read and summarised some of the scientific articles reviewed and drafted the manuscript. LC and PA read and summarised some of the scientific articles reviewed, wrote some sections and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Instituto de Telecomunicações (Leiria), Campus 2 Morro do Lena—Alto do Vieiro, 2411-901 Leiria, Portugal. [2] Dep. de Engenharia Electrotécnica e de Computadores, Universidade de Coimbra, Pólo II da UC, 3030-290 Coimbra, Portugal. [3] Instituto de Telecomunicações (Coimbra), Dep. de Engenharia Electrotécnica e de Computadores, Pólo II da UC, 3030-290 Coimbra, Portugal. [4] Instituto Politécnico de Leiria/ESTG, Campus 2 Morro do Lena—Alto do Vieiro, 2411-901 Leiria, Portugal.

Received: 25 January 2016 Accepted: 12 September 2016
Published online: 29 September 2016

References

1. A Smolic, K Mueller, N Stefanoski, J Ostermann, A Gotchev, GB Akar, G Triantafyllidis, A Koz, Coding algorithms for 3DTV—a survey. Circ Syst Video Technol IEEE Trans. **17**(11), 1606–1621 (2007)
2. P Merkle, K Müller, T Wiegand, 3D video: acquisition, coding, and display. Consum Electron Transac. **56**(2), 946–950 (2010)
3. S Chikkerur, V Sundaram, M Reisslein, LJ Karam, Objective video quality assessment methods: A classification, review, and performance comparison. Broadcasting IEEE Transac. **57**(2), 165–182 (2011)
4. BT Truong, S Venkatesh, Video abstraction: a systematic review and classification. ACM Trans Multimedia Comput. Commun. Appl. (TOMCCAP). **3**(1), 3 (2007)
5. W Hu, N Xie, L Li, X Zeng, S Maybank, A survey on visual content-based video indexing and retrieval. Syst Man Cybern. Part C: Appl Reviews, IEEE Trans. **41**(6), 797–819 (2011)
6. AG Money, H Agius, Video summarisation: a conceptual framework and survey of the state of the art. J Vis Commun Image Represent. **19**(2), 121–143 (2008)
7. Y Li, S-H Lee, C-H Yeh, C-CJ Kuo, Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. IEEE Signal Process Mag. **23**(2), 79–89 (2006)
8. K McHenry, P Bajcsy, An Overview of 3D Data Content, File Formats and Viewers. Technical Report NCSA-ISDA08-002 (2008). http://207.245.165.87/applied-research/papers/overview-3d-data.pdf
9. EH Adelson, JR Bergen, in *The plenoptic function and the elements of early vision*, ed. by M Landy, JA Movshon (MIT Press, Cambridge, 1991), pp. 3–20
10. ND Doulamis, AD Doulamis, YS Avrithis, KS Ntalianis, SD Kollias, Efficient summarization of stereoscopic video sequences. Circ Syst Video Technol IEEE Transac. **10**(4), 501–517 (2000)
11. L Ferreira, P Assuncao, LA da Silva Cruz, in *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop On*. 3D video shot boundary detection based on clustering of depth-temporal features (IEEE, Veszprem, 2013), pp. 1–6
12. Q Xu, P Wang, B Long, M Sbert, M Feixas, R Scopigno, in *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference On*. Selection and 3D visualization of video key frames, (Istanbul, 2010), pp. 52–59
13. K Papachristou, A Tefas, N Nikolaidis, I Pitas, in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop On*. Stereoscopic video shot classification based on weighted linear discriminant analysis, (Reims, 2014), pp. 1–6
14. J Lin, X Ruan, N Yu, R Wei, in *The 27th Chinese Control and Decision Conference (2015 CCDC)*. One-shot learning gesture recognition based on improved 3D SMoSIFT feature descriptor from RGB-D videos, (Qingdao, 2015), pp. 4911–4916
15. I Mademlis, N Nikolaidis, I Pitas, in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. Stereoscopic video description for key-frame extraction in movie summarization, (Nice, 2015), pp. 819–823
16. Y Fu, Y Guo, Y Zhu, F Liu, C Song, Z-H Zhou, Multi-view video summarization. Multimedia IEEE Transac. **12**(7), 717–729 (2010)
17. J Assa, Y Caspi, D Cohen-Or, Action synopsis: pose selection and illustration. ACM Trans Graphics (TOG). **24**(3), 667–676 (2005)
18. K-S Huang, C-F Chang, Y-Y Hsu, S-N Yang, Key probe: a technique for animation keyframe extraction. The Visual Computer. **21**(8–10), 532–541 (2005)
19. J Xu, T Yamasaki, K Aizawa, Summarization of 3D video by rate-distortion trade-off. IEICE Transactions on Information and Systems. **90**(9), 1430–1438 (2007)
20. T Yamasaki, K Aizawa, Motion segmentation and retrieval for 3D video based on modified shape distribution. EURASIP J. Appl. Signal Process. **2007**(1), 211–211 (2007)
21. P Huang, A Hilton, J Starck, Shape similarity for 3D video sequences of people. Int J Comput Vision. **89**(2), 362–381 (2010)
22. T-Y Lee, C-H Lin, Y-S Wang, T-G Chen, Animation key-frame extraction and simplification using deformation analysis. Circ Syst Video Technol IEEE Trans. **18**(4), 478–486 (2008)
23. J Xu, T Yamasaki, K Aizawa, Temporal segmentation of 3-D video by histogram-based feature vectors. Circ Syst Video Technol IEEE Trans. **19**(6), 870–881 (2009)
24. SEF de Avila, APB ao Lopes, A da Luz Jr., A de Albuquerque Araúo, VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn Lett. **32**(1), 56–68 (2011). Image Processing, Computer Vision and Pattern Recognition in Latin America
25. H-J Lee, HJ Shin, J-J Choi, Single image summarization of 3D animation using depth images. Comput Animat Virtual Worlds. **23**(3–4), 417–424 (2012)
26. Y-J Zhang, *Advances in image and video segmentation*. (IRM Press, USA, 2006)
27. J Yuan, H Wang, L Xiao, W Zheng, J Li, F Lin, B Zhang, A formal study of shot boundary detection. Circ Syst Video Technol IEEE Trans. **17**(2), 168–186 (2007)
28. AF Smeaton, P Over, AR Doherty, Video shot boundary detection: seven years of TRECVid activity. Comput Vision Image Underst. **114**(4), 411–418 (2010). Special issue on Image and Video Retrieval Evaluation
29. C Cotsaces, N Nikolaidis, I Pitas, Video shot detection and condensed representation. A review. Signal Process Mag IEEE. **23**(2), 28–37 (2006)
30. J Nam, AH Tewfik, Detection of gradual transitions in video sequences using B-spline interpolation. Multimed IEEE Trans. **7**(4), 667–679 (2005)
31. S Lian, Automatic video temporal segmentation based on multiple features. Soft Computing. **15**, 469–482 (2011)
32. P Sidiropoulos, V Mezaris, I Kompatsiaris, H Meinedo, M Bugalho, I Trancoso, Temporal video segmentation to scenes using high-level audiovisual features. Circ Syst Video Technol IEEE Trans. **21**(8), 1163–1177 (2011)
33. B-L Yeo, B Liu, Rapid scene analysis on compressed video. Circ Syst Video Technol IEEE Transactions on. **5**(6), 533–544 (1995)
34. Z Cernekova, I Pitas, C Nikou, Information theory-based shot cut/fade detection and video summarization. Circ Syst Video Technol IEEE Transactions on. **16**(1), 82–91 (2006)
35. PJ Besl, ND McKay, A method for registration of 3-D shapes. Pattern Anal Mach Intel IEEE Trans. **14**(2), 239–256 (1992)
36. B Ionescu, D Coquin, P Lambert, V Buzuloiu, A fuzzy color-based approach for understanding animated movies content in the indexing task. Eurasip J Image Video Process. **10**(2008), 1–17 (2008)
37. WAC Fernando, CN Canagarajah, DR Bull, in *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference On*. Fade and dissolve detection in uncompressed and compressed video sequences, vol. 3, (Kobe, 1999), pp. 299–303
38. R Slama, H Wannous, M Daoudi, 3D human motion analysis framework for shape similarity and retrieval. Image Vision Comput. **32**(2), 131–154 (2014)
39. U Gargi, R Kasturi, SH Strayer, Performance characterization of video-shot-change detection methods. Circ Syst Video Technol IEEE Transactions on. **10**(1), 1–13 (2000)
40. Y Yang, X Liu, A re-examination of text categorization methods, 42–49 (1999)
41. C Jin, T Fevens, S Mudur, Optimized keyframe extraction for 3D character animations. Comput Animat Virtual Worlds. **23**(6), 559–568 (2012)
42. L Ferreira, LA da Silva Cruz, P Assuncao, A generic framework for optimal 2D/3D key-frame extraction driven by aggregated saliency maps. Signal Processing: Image Commun. **39 Part A**, 98–110 (2015)
43. A Rav-Acha, Y Pritch, S Peleg, in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference On*. Making a long video short: dynamic video synopsis, vol. 1, (2006), pp. 435–441
44. R Xu, ID Wunsch, Survey of clustering algorithms. Neural Netw IEEE Trans. **16**(3), 645–678 (2005)
45. D DeMenthon, V Kobla, D Doermann, in *Proceedings of the Sixth ACM International Conference on Multimedia*. MULTIMEDIA '98. Video summarization by curve simplification (ACM, New York, 1998), pp. 211–218
46. LJ Latecki, D de Wildt, J Hu, in *Multimedia Signal Processing, 2001 IEEE Fourth Workshop On*. Extraction of key frames from videos by optimal color composition matching and polygon simplification, (Cannes, 2001), pp. 245–250
47. J Calic, E Izuierdo, in *Information Technology: Coding and Computing, 2002. Proceedings. International Conference On*. Efficient key-frame extraction and video analysis, (Las Vegas, 2002), pp. 28–33
48. IS Lim, D Thalmann, in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*. Key-posture extraction out of human motion data, vol. 2, (2001), pp. 1167–1169
49. ND Doulamis, AD Doulamis, Y Avrithis, SD Kollias, in *Multimedia Signal Processing, 1999 IEEE 3rd Workshop On*. A stochastic framework for optimal

key frame extraction from MPEG video databases, (Copenhagen, 1999), pp. 141–146
50. B-D Choi, J-W Han, C-S Kim, S-J Ko, Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation. Circ Syst Video Technol IEEE Trans. **17**(4), 407–416 (2007)
51. P Huang, A Hilton, J Starck, in *3DPVT '08: Proceedings of the Fourth International Symposium on 3D Data Processing, Visualization and Transmission*. Automatic 3D video summarization: key frame extraction from self-similarity (IEEE Computer Society, Washington, DC, 2008)
52. Y Gong, X Liu, in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference On*. Video summarization using singular value decomposition, vol. 2, (Hilton Head Island, 2000), pp. 174–180
53. M Cooper, J Foote, in *Multimedia Signal Processing, 2002 IEEE Workshop On*. Summarizing video using non-negative similarity matrix factorization, (St. Thomas, 2002), pp. 25–28
54. W Abd-Almageed, in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference On*. Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing, (San Diego, 2008), pp. 3200–3203
55. CH Lee, A Varshney, DW Jacobs, Mesh saliency. ACM Trans. Graph. **24**(3), 659–666 (2005)
56. C Halit, T Capin, Multiscale motion saliency for keyframe extraction from motion capture sequences. Comput Animat Virtual Worlds. **22**(1), 3–14 (2011)
57. C Nguyen, Y Niu, F Liu, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Video summagator: an interface for video summarization and navigation (ACM, New York, 2012), pp. 647–650
58. K Schoeffmann, D Ahlstrom, MA Hudelist, 3D interfaces to improve the performance of visual known-item search. Multimed IEEE Trans. **16**(7), 1942–1951 (2014)
59. ITU-R Recommendations 500-13. Methodology for the Subjective Assessment of the Quality of Television Pictures (International Telecommunication Union, Geneva, 2012)
60. Y-F Ma, X-S Hua, L Lu, H-J Zhang, A generic framework of user attention model and its application in video summarization. Multimedia IEEE Trans. **7**(5), 907–919 (2005)
61. J Peng, Q Xiao-Lin, Keyframe-based video summary using visual attention clues. IEEE Multimedia. **17**(2), 64–73 (2010)
62. N Ejaz, I Mehmood, S Wook Baik, Efficient visual attention based framework for extracting key frames from videos. Signal Process Image Commun. **28**(1), 34–44 (2013)
63. N Ejaz, I Mehmood, SW Baik, Feature aggregation based visual attention model for video summarization. Comput Electr Eng. **40**(3), 993–1005 (2014)
64. P Mundur, Y Rao, Y Yesha, Keyframe-based video summarization using Delaunay clustering. Int. J. Digit. Libr. **6**(2), 219–232 (2006)
65. M Furini, F Geraci, M Montangero, M Pellegrini, Stimo: still and moving video storyboard for the web scenario. Multimed Tools Appl. **46**(1), 47–69 (2010)
66. T-Y Liu, X-D Zhang, J Feng, K-T Lo, Shot reconstruction degree: a novel criterion for key frame selection. Pattern Recogn Lett. **25**, 1451–1457 (2004)
67. HS Chang, S Sull, SU Lee, Efficient video indexing scheme for content-based retrieval. Circ Syst Video Technol IEEE Trans. **9**(8), 1269–1279 (1999)
68. G Ciocca, R Schettini, An innovative algorithm for key frame extraction in video summarization. J Real-Time Image Process. **1**(1), 69–88 (2006)
69. Q-G Ji, Z-D Fang, Z-H Xie, Z-M Lu, Video abstraction based on the visual attention model and online clustering. Signal Process Image Commun. **28**(3), 241–253 (2013)
70. A Khodakovsky, P Schröder, W Sweldens, in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '00*. Progressive geometry compression (ACM Press/Addison-Wesley Publishing Co., New York, 2000), pp. 271–278
71. HM Briceño, PV Sander, L McMillan, S Gortler, H Hoppe, in *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '03*. Geometry videos: a new representation for 3D animations (Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2003), pp. 136–146
72. K Schoeffmann, F Hopfgartner, O Marques, L Boeszoermenyi, JM Jose, Video browsing interfaces and applications: a review. J Photonics Energy. **1**, 018004–01800435 (2010)
73. B Bustos, DA Keim, D Saupe, T Schreck, Vranić, Feature-based similarity search in 3D object databases. ACM Comput Surv. **37**(4), 345–387 (2005)
74. T Napoléon, H Sahbi, From 2D silhouettes to 3D object retrieval: contributions and benchmarking. J Image Video Process, 1 (2010)
75. M Savelonas, I Pratikakis, K Sfikas, An overview of partial 3D object retrieval methodologies. Multimed Tools Appl,. **74**, 1–26 (2014)
76. N Vretos, N Nikolaidis, I Pitas, in *3DTV-Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*. The use of audio-visual description profile in 3D video content description, (Zurich, 2012), pp. 1–4
77. M Sano, W Bailer, A Messina, J-P Evain, M Matton, in *IVMSP Workshop, 2013 IEEE 11th*. The MPEG-7 audiovisual description profile (AVDP) and its application to multi-view video, (Seoul, 2013), pp. 1–4
78. V Blanz, T Vetter, in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '99*. A morphable model for the synthesis of 3D faces (ACM Press/Addison-Wesley Publishing Co., New York, 1999), pp. 187–194
79. C Shelton, Morphable surface models. Int J Comput Vis. **38**(1), 75–91 (2000)
80. J Xu, T Yamasaki, K Aizawa, in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. Motion editing in 3D video database, (2006), pp. 472–479
81. R Parent, *Computer animation: algorithms and techniques*. (Morgan-Kaufmann, USA, 2012)
82. X Cao, Z Li, Q Dai, Semi-automatic 2D-to-3D conversion using disparity propagation. Broadcasting, IEEE Transactions on. **57**(2), 491–499 (2011)
83. W-N Lie, C-Y Chen, W-C Chen, 2D to 3D video conversion with key-frame depth propagation and trilateral filtering. Electronics Letters. **47**(5), 319–321 (2011)
84. D Wang, J Liu, J Sun, W Liu, Y Li, in *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium On*. A novel key-frame extraction method for semi-automatic 2D-to-3D video conversion, (Seoul, 2012), pp. 1–5
85. K Ju, H Xiong, in *Proc. SPIE*. A semi-automatic 2D-to-3D video conversion with adaptive key-frame selection, vol. 9273 (SPIE, Beijing, 2014), pp. 92730M1–92730M8
86. J-L Lai, Y Yi, Key frame extraction based on visual attention model. J Vis Commun Image Represent. **23**(1), 114–125 (2012)
87. H Schwarz, D Marpe, T Wiegand, Overview of the scalable video coding extension of the H.264/AVC standard. Circ Syst Video Technol IEEE Trans. **17**(9), 1103–1120 (2007)
88. L Ferreira, L Cruz, P Assuncao, in *EUROCON 2011 IEEE*. Efficient scalable coding of video summaries using dynamic GOP structures, (Lisbon, 2011), pp. 1–4
89. L Herranz, S Jiang, Scalable storyboards in handheld devices: applications and evaluation metrics. Multimed Tools Appl. **75**, 1–29 (2015)
90. P Over, AF Smeaton, G Awad, in *Proceedings of the 2Nd ACM TRECVid Video Summarization Workshop*. TVS '08. The trecvid 2008 BBC rushes summarization evaluation (ACM, New York, 2008), pp. 1–20
91. P Over, AF Smeaton, P Kelly, in *Proceedings of the International Workshop on TRECVID Video Summarization*. TVS '07. The TRECVID 2007 BBC rushes summarization evaluation pilot (ACM, New York, 2007), pp. 1–15
92. J Bescos, JM Martinez, L Herranz, F Tiburzi, Content-driven adaptation of on-line video. Signal Process Image Commun. **22**(7–8), 651–668 (2007)
93. M Mrak, J Calic, A Kondoz, Fast analysis of scalable video for adaptive browsing interfaces. Comp Vision Image Underst. **113**(3), 425–434 (2009)
94. L Herranz, J Martínez, An integrated approach to summarization and adaptation using H.264/MPEG-4 SVC. Signal Processing: Image Comm. **24**(6), 499–509 (2009)