**Why is shot boundary important**

The recent increase in the number of videos available in cyberspace is due to the availability of multimedia devices, highly developed communication technologies, and low-cost storage devices. These videos are simply stored in databases through text annotation. Content-based video browsing and retrieval are inefficient due to the method used to store videos in databases. Video databases are large in size and contain voluminous information, and these characteristics emphasize the need for automated video structure analyses. Shot boundary detection (SBD) is considered a substantial process of video browsing and retrieval. SBD aims to detect transition and their boundaries between consecutive shots; hence, shots with rich information are used in the content-based video indexing and retrieval.

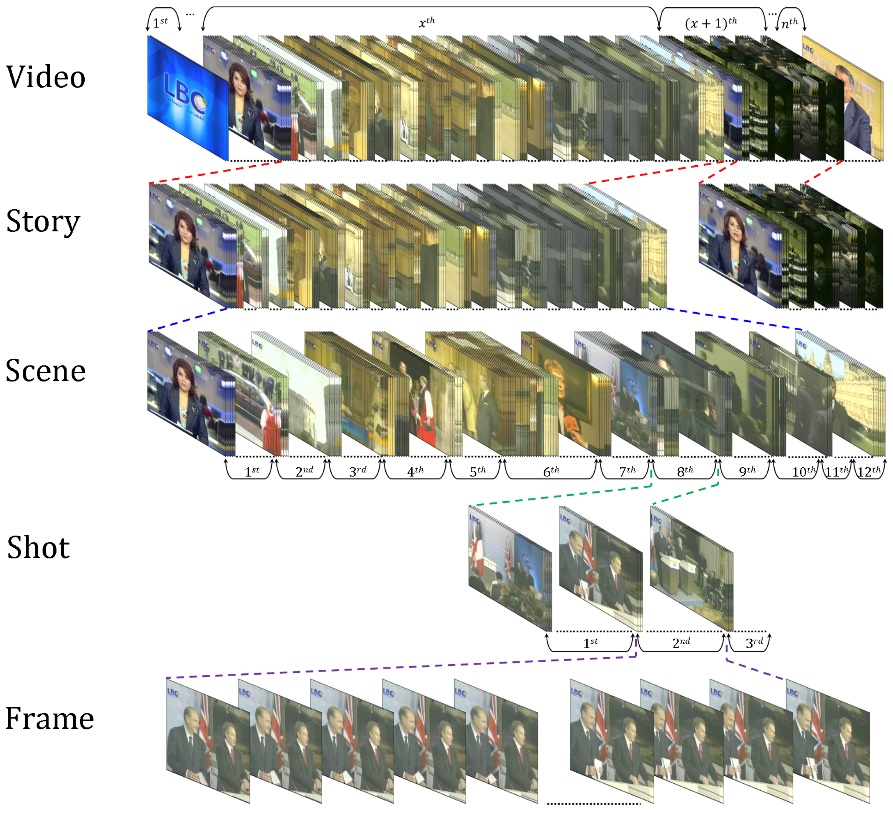
**Why video not text**

Video is the most consumed data type on the Internet. Videos consume a large amount of storage space, and they contain voluminous information [[**3**](https://www.mdpi.com/1099-4300/20/4/214/htm#B3-entropy-20-00214)]. Text, audio, and images are combined to constitute a video [[**4**](https://www.mdpi.com/1099-4300/20/4/214/htm#B4-entropy-20-00214)], so videos are large in size. The human brain gathers most information visually and can process visual media faster than it can process text. Thus, videos facilitate easy communication among individuals [[**5**](https://www.mdpi.com/1099-4300/20/4/214/htm#B5-entropy-20-00214)]. In the past two decades, computer performance, storage media availability, and the number of recording devices have increased considerably, resulting in the active uploading and viewing of videos at inconceivable rates [[**6**](https://www.mdpi.com/1099-4300/20/4/214/htm#B6-entropy-20-00214)]. For example, YouTube is the second most popular video sharing website (VSW). Statistics show that 300 hours of videos were uploaded every minute in 2016, and this figure presents a significant increase from the 72 hours of videos uploaded in 2015; furthermore, five billion hours of videos are being viewed daily. Video consumption increases at a rate of 300% annually. This growth is due to individuals and companies sharing their media through VSWs to broaden their audience. Moreover, individuals can now easily access the Internet as a result of the prevalence of mobile technology [[**2**](https://www.mdpi.com/1099-4300/20/4/214/htm#B2-entropy-20-00214)], which motivates them to upload videos to VSWs or social media. Readily available video editing software on computers and portable devices enable users to manipulate video contents by combining two or more videos, altering videos by adding other video contents, and omitting certain video parts. In addition, uploading videos to hosting sites is no longer restricted to skilled programmers, and this condition has resulted in video duplication. Video repetitions occur in many forms, such as downloading and re-uploading a video as it is, inserting logos, and covering copyrights by replacing video features (e.g., changing illumination or resizing video frames).

**Why is shot boundary hard**

In general, the performance of a SBD algorithm is based on its ability to detect transitions (shot boundaries) in a video sequence. That is, SBD algorithm performance can be measured by its ability in detecting correct transition. Where, a SBD accuracy generally depends on the extracted features and their effectiveness of representing the the visual content of video frames [[**22**](https://www.mdpi.com/1099-4300/20/4/214/htm#B22-entropy-20-00214)]. The second factor that influences a SBD algorithm performance is the computational cost of the algorithm, that need to be reduced where in contrast, algorithm speed is increased. Note that, theoretically, within a shot, frames are very similar in terms of their visual content. Therefore, when shot transition is occurred, a change in similarity/dissimilarity values will be appeared. In hard transition (HT), a very high change in similarity/dissimilarity values, but for soft transition (ST) it is small [[**23**](https://www.mdpi.com/1099-4300/20/4/214/htm#B23-entropy-20-00214)]. Practically, there are some effects that appear in a video shot such as: flash lights or light variations, object/camera motion, camera operation (such as zooming, panning, and tilting), and similar background. These effects are highly provoking the accuracy of transitions detection and thus greatly impact on SBD algorithm performance. To fulfill the maximum efficiency, SBD should be able to detect shot transitions between two consecutive shots by, first, minimizing both false alarm signals (FASs), i.e., false positives, within a shot (intra-shot frames), and second, miss detects (MSD), i.e., false negatives, between two consecutive shots (inter-shot frames) during transition detection process. Currently, there is no complete solution to these problems or most of them in the same algorithm. That is, a favorable and efficient method for detecting transitions between shots remains unavailable despite the increasing attention devoted to SBD in the last two decades. This unavailability is due to the randomness and size of raw video data. Hence, a robust, efficient, automated SBD method is an urgent requirement [[**11**](https://www.mdpi.com/1099-4300/20/4/214/htm#B11-entropy-20-00214),[**19**](https://www.mdpi.com/1099-4300/20/4/214/htm#B19-entropy-20-00214)].

***Video Hierarchy***

To some extent, video hierarchy is comparable to a book. A video consists of a single story (such as a football game) or multiple stories (such as news) [[**11**](https://www.mdpi.com/1099-4300/20/4/214/htm#B11-entropy-20-00214)]. A story is defined as a clip that captures a series of events or a continuous action, and it may be composed of several scenes. A scene is a pool of semantically related and temporally contiguous shots captured at multiple camera angles [[**16**](https://www.mdpi.com/1099-4300/20/4/214/htm#B16-entropy-20-00214),[**28**](https://www.mdpi.com/1099-4300/20/4/214/htm#B28-entropy-20-00214)]. [**Figure 2**](https://www.mdpi.com/1099-4300/20/4/214/htm#fig_body_display_entropy-20-00214-f002) shows the hierarchy of a video. As previously mentioned, the hierarchy of a video closely resembles that of a document, such as a book consisting of chapters, which are similar to stories in a video [[**29**](https://www.mdpi.com/1099-4300/20/4/214/htm#B29-entropy-20-00214)]. Each chapter comprises sections similar to scenes. Sections consist of paragraphs similar to a video comprising shots. A paragraph is a group of interconnected sentences that are similar to the interconnected frames that constitute a shot in a video. Moreover, a sentence is composed of multiple words, similar to a shot being composed of frames. Each frame in a video represents a single image, while a shot represents a continuous sequence of frames captured by a single camera, as explained previously.******

**Shots**

A shot is the building block of a video; it is a set of one or more frames grabbed continually (uninterruptedly) by a single recording device, and these frames symbolize an incessant action in time and space that shows a certain action or event [[**1**](https://www.mdpi.com/1099-4300/20/4/214/htm#B1-entropy-20-00214),[**3**](https://www.mdpi.com/1099-4300/20/4/214/htm#B3-entropy-20-00214),[**6**](https://www.mdpi.com/1099-4300/20/4/214/htm#B6-entropy-20-00214),[**15**](https://www.mdpi.com/1099-4300/20/4/214/htm#B15-entropy-20-00214),[**30**](https://www.mdpi.com/1099-4300/20/4/214/htm#B30-entropy-20-00214),[**31**](https://www.mdpi.com/1099-4300/20/4/214/htm#B31-entropy-20-00214)]. A shot is also considered the smallest unit of temporal visual information [[**3**](https://www.mdpi.com/1099-4300/20/4/214/htm#B3-entropy-20-00214),[**11**](https://www.mdpi.com/1099-4300/20/4/214/htm#B11-entropy-20-00214),[**15**](https://www.mdpi.com/1099-4300/20/4/214/htm#B15-entropy-20-00214)]. The frames within a shot (intra-shot frames) contain similar information and visual features with temporal variations [[**32**](https://www.mdpi.com/1099-4300/20/4/214/htm#B32-entropy-20-00214),[**33**](https://www.mdpi.com/1099-4300/20/4/214/htm#B33-entropy-20-00214)]. These variations in time between shot elements (i.e., frames) may cause small or large changes due to the action between start and stop marks [[**34**](https://www.mdpi.com/1099-4300/20/4/214/htm#B34-entropy-20-00214)]. These changes are due to the fact that a shot captures objects in the real world and the semantics, dynamics, and syntax of these objects are merged to obtain shot frames [[**3**](https://www.mdpi.com/1099-4300/20/4/214/htm#B3-entropy-20-00214)], such as object motion, camera motion, or camera operation. Moreover, a shot is supposed to comprise rigid objects or objects composed of rigid parts connected together [[**3**](https://www.mdpi.com/1099-4300/20/4/214/htm#B3-entropy-20-00214)]. Shots are classified into four types according to the object and/or camera motion; these types are static object with a static camera, static object with a dynamic camera, dynamic object with a static camera, and dynamic object with a dynamic camera [[**35**](https://www.mdpi.com/1099-4300/20/4/214/htm#B35-entropy-20-00214)]. A frame is the smallest unit that constitutes a shot. Hence, the shot and scene hierarchies are analogous to a sentence and paragraph. Shots are essential in depicting a story, and scenes are a necessary unit for a visual narrative [[**16**](https://www.mdpi.com/1099-4300/20/4/214/htm#B16-entropy-20-00214)]. Video frames are temporally ordered, but they are not independent [[**36**](https://www.mdpi.com/1099-4300/20/4/214/htm#B36-entropy-20-00214)].

#### *Video Transition Types*

#### 

**Approaches**

#### 1-*Pixel-Based Approach*

The pixel-based approach (PBA) or pixel-wise comparison is used as a ROVI directly from the pixel intensities of video frames. PBA involves calculating the difference between two corresponding pixels (at location *x* and *y*) of two consecutive video frames (fn and fn+1). In the next stage of PBA, the total sum of pixel differences is determined and compared with a threshold. A transition is declared if the sum exceeds the selected threshold [[**67**](https://www.mdpi.com/1099-4300/20/4/214/htm#B67-entropy-20-00214)].

The earliest researchers who implemented PBA for SBD are [[**68**](https://www.mdpi.com/1099-4300/20/4/214/htm#B68-entropy-20-00214),[**69**](https://www.mdpi.com/1099-4300/20/4/214/htm#B69-entropy-20-00214),[**70**](https://www.mdpi.com/1099-4300/20/4/214/htm#B70-entropy-20-00214)]. The researchers in [[**68**](https://www.mdpi.com/1099-4300/20/4/214/htm#B68-entropy-20-00214)] implemented PBA to locate HT by comparing the sum of the absolute differences of the total pixel (Equations ([**8**](https://www.mdpi.com/1099-4300/20/4/214/htm#FD8-entropy-20-00214)) and ([**9**](https://www.mdpi.com/1099-4300/20/4/214/htm#FD9-entropy-20-00214))) with a threshold value. When the sum was greater than the threshold, HT was declared; otherwise, a frame’s shot was considered.

The researchers in [[**69**](https://www.mdpi.com/1099-4300/20/4/214/htm#B69-entropy-20-00214)] modified the technique proposed in [[**68**](https://www.mdpi.com/1099-4300/20/4/214/htm#B68-entropy-20-00214)] to reduce the disturbance in dissimilarity signal. First, they compared the corresponding pixel differences of two successive frames to threshold T1. When the partial difference exceeded T1 (Equation ([**10**](https://www.mdpi.com/1099-4300/20/4/214/htm#FD10-entropy-20-00214))), they considered that pixel a change. Second, they summed up all the partial differences of the pixels and compared the result to a second threshold T2 (Equation ([**11**](https://www.mdpi.com/1099-4300/20/4/214/htm#FD11-entropy-20-00214))) (the ratio of pixel change). When the value exceeded T2, HT is detected.

PBA are highly sensitive to OCM, and they produce a high rate of false alarms (FAR). As a result of their dependency on spatial location, these techniques are particularly sensitive to motion, even global motion [[**73**](https://www.mdpi.com/1099-4300/20/4/214/htm#B73-entropy-20-00214)]. Although PBA techniques are highly sensitive to motion, missed detections (MSDs) occur [[**74**](https://www.mdpi.com/1099-4300/20/4/214/htm#B74-entropy-20-00214)]. For example, two adjacent frames within intra-shots with different pixel intensity disturbances can result in similar pixel differences. Furthermore, because of the high sensitivity of PBA techniques, intra-shots with camera motion can be incorrectly classified as gradual transitions. These methods rely on the threshold procedure, and they do not consider the temporal relation of dissimilarity/similarity signal. [**Table 1**](https://www.mdpi.com/1099-4300/20/4/214/htm#table_body_display_entropy-20-00214-t001) presents a summary for the previously discussed PBA algorithms, their parameters settings and ability for detecting transitions.

#### *Histogram-Based Approaches*

Color histograms or histograms reflect the distribution of colors in an image. Histograms are considered substitutes for PBAs because they do not consider the spatial information of all color spaces. Hence, histograms, to some extent, are regarded as invariant to local motion or small global motion compared with PBAs [[**75**](https://www.mdpi.com/1099-4300/20/4/214/htm#B75-entropy-20-00214),[**76**](https://www.mdpi.com/1099-4300/20/4/214/htm#B76-entropy-20-00214)].