Video indexing and retrieval based on content: a systematic literature review

Newton Spolaôr[1], Huei Diana Lee[1], Weber Shoity Resende Takaki[1,2], Narco Afonso Ravazzoli Maciejewski[1], Leandro Augusto Ensina[1], Cláudio Saddy Rodrigues Coy[2], and Feng Chung Wu[1,2]

[1] Laboratory of Bioinformatics (LABI), Western Paraná State University (UNIOESTE), Foz do Iguaçu, Brazil
[2] University of Campinas (UNICAMP), Campinas, Brazil
{newtonspolaor,hueidianalee}@gmail.com

Abstract. Content-based video retrieval and indexing have been associated with intelligent methods in many applications such as education, medicine and agriculture. However, an extensive and replicable review of the recent literature is missing. Moreover, relevant topics that can support video retrieval, such as dimensionality reduction, have not been surveyed. This work designs and conducts a systematic review to find papers able to answer the following research question: "what segmentation, feature extraction, dimensionality reduction and machine learning approaches have been applied for content-based video indexing and retrieval?". By applying a research protocol proposed by us, 153 papers published from 2011 to 2018 were selected. As a result, it was found that strategies for cut-based segmentation, color-based indexing, k-means based dimensionality reduction and data clustering have been the most frequent choices in recent papers. All the information extracted from these papers can be found in a publicly available spreadsheet. This work also indicates additional findings and future research directions.

Keywords: Color Features, Unsupervised Learning, Shot Boundary Detection.

1 Introduction

Multimedia documents composed of different media types have increasingly been published and consumed [1, 2]. This fact is due to the larger access to computational resources and the Internet, among other reasons [3]. Video in particular consists in a usual way to capture and share information, as it is able to represent moving objects in space and time accordingly. These benefits come at the price of reasonable storage and processing costs [4].

In general, video content is richer than single image content [5]. A video file typically has much raw data, but little prior structure. Moreover, information available in video occasionally include textual metadata and captions, images (frames) and audio.

Due to the crescent interest in video, automatic indexing and retrieval are usually considered in multimedia research. In particular, the former specifies indexes (fea-

tures) to describe a video, whereas the latter allows one to search for relevant videos. These tasks can be combined, for example, to find video in an indexed database that contains characteristics similar to the ones given by a user's query. In this work, the retrieval and indexing based on the video content is considered. Both tasks have been applied in agriculture [6], cinema [7], discourse analysis [8], education [9], geo-referenced video [10], human action recognition [11], journalism [12], marketing [13], medicine [14], sports [15] and television broadcast [16].

Some surveys identify relevant research on Content-based Video Indexing and Retrieval (CBVIR). One of them describes the background and an extensive review of CBVIR methods and results [4]. In [5] the reader can find a broad survey that organizes procedures inherent to CBVIR, describes their merits and limitations and supplements previous work. Besides describing several approaches from the relevant literature, [17] reports research challenges. Another example, published in [18], focuses on the use of intelligent methods for several tasks, such as multimedia indexing and retrieval. Soft computing arises as an alternative due to the ability to find inexact solutions [3]. Recently, a Systematic literature Review (SR) on content-based multimedia medical retrieval that concentrates efforts on medical images was conducted [19]. Besides image retrieval, the search for medical video is addressed in [20].

This work aims to supplement earlier surveys by designing and conducting an SR on CBVIR. In particular, the SR method allows one to obtain a replicable and wide review of the relevant literature with reduced subjectivity [21].

Besides the SR method application, this work contains three main differences from previous surveys:

- Finding of Dimensionality Reduction (DR) approaches for video retrieval: DR, an important pre-processing procedure to reduce curse of dimensionality effects, is usual in automatic processes to learn from data [22]. This curse involves phenomena regarding the increasing sparsity of the data as the number of dimensions grows [23]. In CBVIR, dimensionality reduction can yield a small set of video indexes more useful for retrieval and cheaper to be extracted from new videos than the original set of features. Thus, this paper pays attention to the DR approaches used in relevant papers. Besides dimensionality reduction, segmentation, feature extraction and machine learning approaches are reviewed due to their frequent use in the literature;
- Proposal of a review protocol on video indexing and retrieval: by sharing the designed protocol, this work supports other researchers to (1) replicate and update the current review and (2) apply the SR method in other research topics associated with video content. Different from [19], our protocol reviews papers that consider video retrieval and indexing regardless of the video domain;
- Publication period: we supplement earlier surveys by focusing on papers published from 2011 to 2018.

This paper is organized as follows. Section 2 describes concepts inherent to content-based video retrieval and indexing. Section 3 presents the review method and protocol applied to identify usual approaches in CBVIR and other findings, which in turn are

reported in Section 4. Sections 5 and 6, respectively, consider future directions and final remarks.

2 Background

In a scenario in which a user retrieves video based on content, a similarity measure is typically used to compare query indexes with indexes describing repository videos [4]. The results can be ranked by relevance to enhance future queries.

In particular, video query is usually based on the following approaches: Query-by-Example (QbE), sketch, image, text and audio [5]. They differ in terms of the input provided by the user. Thus, QbE involves the retrieval of videos similar to the query video (example). Sketch and image queries can feed the search for videos with similar trajectories or frames, respectively. Some CBVIR methods receive query keywords or natural language text from the users. Finally, Automatic Speech Recognition (ASR) methods [24] can be employed to extract text from audio for video retrieval.

Besides traditional measures derived from Minkowski distance, such as Euclidean and Manhattan, the cosine similarity is another alternative employed in CBVIR [25]. In addition, one can note the use of measures designed specifically for video, such as a measure to differentiate trajectories [26].

A CBVIR method typically yields a set of candidate videos that accomplish the query. In general, these videos are ranked by a method, according to a relevance criterion, or by the users. User's feedback on a ranking is useful to refine future queries according to his/her preferences. This feedback can also be simulated, as illustrated in [27].

In what follows, we consider some tasks that support the mentioned procedures and are applied in papers found by the SR method: video segmentation, feature extraction for video indexing, Dimensionality Reduction (DR) and Machine Learning (ML). Other concepts related to CBVIR are presented in detail in the literature [28, 5, 29, 30].

2.1 Video Segmentation

This task divides a video into segments of related frames [31]. A common segment type, named shot, corresponds to a frame sequence that represents an action. This action, continuous in space and time, is recorded by a simple camera operation [32]. Shots are often considered as the fundamental units of video in CBVIR.

In specific domains, a video segment can be associated with other definitions. In lecture-based distance learning, for example, a subsequence of video regarding a topic or subtopic within a lecture is usually regarded as a segment [9]. To identify these subsequences, slides can be used as separators.

The segmentation of generic video into shots consists in an important research topic [33]. In general, the segmentation methods extract information to describe frames and identify segment boundaries. Different Shot Boundary Detection (SBD) approaches useful for video segmentation have been created [34, 4]. This work surveys recent ones in CBVIR context in Section 4.1.

Depending on the approach used for video segmentation, redundant frames may be found. Thus, some frames representing the shot content can be selected as keyframes. These special elements are useful, for example, to perform video summarization [5].

A shot is relatively small and does not necessarily correspond to a meaningful semantic unit. Thus, some CBVIR methods group these segments into larger ones, such as scenes – combinations of adjacent shots associated with the same subject or topic [35]. This combination is obtained, for example, from information extracted from text, image or audio inherent to a video.

2.2 Feature Extraction

This task extracts features that are typically used as indexes for CBVIR. Three abstraction levels are usually considered to categorize video features [4]: raw data, descriptors and concepts. Descriptors and concepts are also known as low-level and high-level features, respectively. If different levels are taken into account, a more complete characterization of video can be obtained, as illustrated in [36]. However, using a high amount of features may cause the curse of dimensionality, demanding dimensionality reduction approaches (Section 2.3).

Descriptors can be applied to characterize several video elements, such as keyframes, movement and objects, i.e., relevant components within a specific domain, such as caption text or human face [5]. Some descriptors are also used for image processing in general [37]. Usual descriptors include bag of visual words [38].

Concepts or semantic indexes are assigned to videos by different approaches, such as manual or automatic annotation [39, 40]. The idea is to associate segments, objects or events – complex activities that can be directly noticed and occur in specific local and time – with pre-defined semantic categories. Automatic annotation in particular is often supported by ML algorithms [22] (Section 2.4). In summary, these intelligent techniques are able to learn patterns (models) from video features (usually descriptors). As a result, each input video is annotated with one or more concepts (classes).

In this work, we focused on approaches to extract descriptors, as they have been the most frequent in the relevant literature. However, it should be emphasized that we also found methods working with concepts for video indexing [39, 1, 41, 42, 43, 44, 36].

2.3 Dimensionality Reduction

Dimensionality Reduction (DR) is an alternative to tackle the curse of dimensionality. Plainly speaking, two close data points in a 2D space are likely distant in a 100D space [23]. As content-based video retrieval typically depends on the similarity calculation based on indexes (dimensions), it can be hindered by the curse if a too large number of irrelevant indexes is used. Problems may also arise, for example, for machine learning algorithms that learn concepts from video data, as it is difficult to predict semantic indexes from a sparse feature space.

Usual DR tasks consist in feature construction and Feature Selection (FS). The former, a.k.a. feature extraction by the data mining community, should not be con-

fused with the indexing task for video retrieval. It aims to build expressive features from the original data attributes by mapping the input dimension space into another one usually smaller. By doing so, it is possible to enlighten video characteristics not directly visible in the input feature space. Although this idea can improve the retrieval performance, domain experts can have more difficulty to understand the new data representation.

Principal Components Analysis (PCA) is a well-known feature construction technique that maps the original dimension into a new one by performing an orthonormal transformation in the data [45]. As a result, components representing data variance are found.

Feature Selection, in turn, aims to remove irrelevant and or redundant features from video data, selecting the remaining ones [23]. Irrelevant features can be removed without affecting the learning performance. Their removal can be especially useful to the popular Nearest Neighbors (NN) algorithm, as it uses a similarity measure during its training (Section 2.4). A redundant feature implies the co-presence of another feature with similar representation power. The withdrawing of irrelevant and or redundant features may bring benefits, such as learning performance improvement and or model comprehensibility by reducing the complexity of the patterns. It should be emphasized that some FS algorithms rank features according to their importance. To specify a subset in this scenario, one can choose, for example, the features with importance score better than a threshold.

Using a DR task in CBVIR can promote several benefits, such as the improvement of video retrieval based on machine learning and the saving of computational resources by avoiding the extraction of unimportant and costly video indexes.

2.4 Machine Learning

After applying DR, automatic processes to learn from the data, such as the Knowledge Discovery in Databases, usually apply Machine Learning (ML) algorithms. These algorithms, in particular, build models with complex data patterns to make intelligent decisions [22].

Let D be a dataset typically submitted for ML, composed of N instances E_i, i = 1 … N. A vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ composed of M features x_j, j = 1 … M describes each instance E_i. On one hand, some ML strategies, such as unsupervised algorithms for clustering, discover a model with groups (clusters) of instances close in the feature space. On the other hand, supervised algorithms also take into account a class or target feature Y_i in the data, i.e., $E_i = (x_i , Y_i)$. These methods build a model H to predict one or more labels (class values) for the class Y of a new instance E = (x, ?). To predict discrete or numeric labels, classification or regression models can be respectively used.

Support Vector Machines (SVM) are well-known supervised learning algorithms [3]. This method, grounded in statistical learning theory, is able to transform the initial feature space into a higher dimensional space. By doing so, SVM is able to build a model based on a hyperplane, i.e., a decision boundary that separates instances from different classes.

A simpler alternative for supervised learning, Nearest Neighbors (NN), is based on lazy learning. Instead of building a model, this scheme simply stores labeled instances

from the input dataset D. Only after receiving a new instance E, NN performs generalization to predict the corresponding class based on the similarity between E and the stored instances. In particular, this class is typically given by the most common class within the k nearest instances (neighbors) to E. As the similitude calculation uses all data features, NN is sensitive to the curse of dimensionality (Section 2.3).

The most applied algorithm for unsupervised clustering consists in k-means [22], which organizes the instances into k exclusive groups (clusters). The clusters optimize an objective criterion, such as a dissimilarity measure, such that instances clustered together are similar and instances in different clusters are not similar. It should be emphasized that unsupervised learning algorithms usually disregard the class, such that only data features are taken into account.

ML algorithms have been applied, for example, to predict concepts (semantic indexes) or categories from low-level features, as well as to assist video segmentation and retrieval [5]. In what follows, we describe the review method applied by us to find recent CBVIR methods associated with approaches for ML, DR, feature extraction and video segmentation.

3 Systematic Review Process

To capture a replicable and wide panorama on Content-based Video Indexing and Retrieval (CBVIR), we instantiated the Systematic literature Review process (SR) [21]. Fig. 1 summarizes the workflow regarding the three SR steps and their relevant inputs and outputs [46].

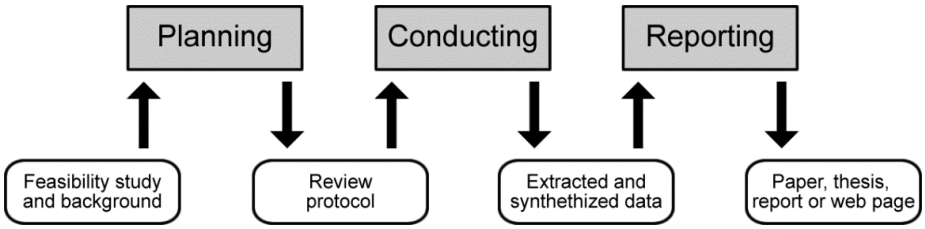

Fig. 1. Systematic literature review workflow (this figure is an adaptation of material published in [46]. Any citation to the material should consider that paper).

The first step receives two inputs: a feasibility study and the background on the research questions to be answered [46]. In particular, the study allows one to verify the need for a review, identifying and analyzing any existing systematic review on the subject of interest [21]. By processing these inputs, planning generates a protocol that ease new applications of the SR process.

In the next step, a researcher can follow the protocol to yield data able to answer research questions, which are the core of the SR process. Finally, the data obtained after conducting the review is published, for example, as a piece of a paper.

In what follows, we present detail on the planning step proposed for this work, including the protocol components considered. The main results of the conduction step are reported in Section 4.

3.1 Systematic Review Planning

During the feasibility study, six recent surveys associated with CBVIR were found [19, 3, 4, 5, 17, 18]. However, differently from these surveys, this work: (1) pays attention on dimensionality reduction approaches, (2) proposes a research protocol on indexing and retrieval of videos from any domain and (3) reviews papers published from 2011 to 2018.

As this work is innovative, the need for a review is accomplished. Moreover, we take advantage of the background considered in related work, briefly described in Section 2, to establish pieces of the current review protocol.

Although a systematic review protocol can include many components [21], the following ones are usually considered: (1) research question, (2) study search strategy, (3) selection criteria and strategy, (4) quality criteria and strategy, (5) information to be extracted and (6) synthesis strategy.

This work surveys the literature to answer the following research question: what segmentation, feature extraction, dimensionality reduction and machine learning approaches have been applied for content-based video indexing and retrieval?

The search strategy used involves applying a string to seven bibliographic databases: ACM Digital Library, CiteSeerX, IEEE Xplore, Science Direct, Scopus, Web of Science and Wiley. In particular, the string employed is: *((indexing OR retrieval OR retrieving OR retrieve OR summarization OR summary OR skimming OR skim OR skims OR abstraction OR abstract OR synopsis OR recover OR recovering) AND ("content-based" OR "semantic-based" OR "context-aware" OR "context-preserving" OR "concept-oriented" OR "keyword-focused") AND (video OR videos OR "video-based" OR multimedia OR "multi-media" OR "multi media" OR audio-visual OR shot OR shots OR keyframe OR "key-frame" OR keyframes OR "key-frames" OR headshot OR headshots))*. Whenever a database supports searches restricted to paper title, abstract and keywords, the feature is used.

Regarding the selection strategy, we specified 13 exclusion criteria. Thus, if a study fulfills an Exclusion Criterion (EC), it is removed from the next SR components. It should be emphasized that the 13 criteria described in what follows are verified in the study title, abstract and full text, if necessary.

- EC1. The piece of work deals with 3D elements, such as objects;
- EC2. The piece of work composed of only one page (abstract paper), poster, presentation, proceeding, program of scientific events and tutorial slides;
- EC3. The piece of work published before 2011;
- EC4. The piece of work does not suit the research question;
- EC5. Duplicated pieces of work written by the same authors (Similar title, abstract, results or text). In this case, only one is kept;
- EC6. The piece of work written in a language different than English;
- EC7. The piece of work hosted in web pages that can not be accessed by using the UNIOESTE and UNICAMP login credentials;
- EC8. The piece of work does not focus mainly on video retrieval;
- EC9. The piece of work does not conduct experimental evaluation (quantitative study) on video retrieval;
- EC10. A patent;

- EC11. The piece of work deals with video copy or near-duplicate retrieval;
- EC12. The piece of work related to academic challenges;
- EC13. The piece of work deals with object recognition or identification.

The strategy used to estimate the methodological quality of each selected paper involves applying four criteria, such that each Quality Criterion (QC) consists in a yes/no question.

In this work, 18 information items are extracted from each selected paper to verify the quality criteria and to conduct the synthesis. The complete description of the four quality criteria and 18 information items taken into account by us, are available at http://tiny.cc/58nohy.

We conducted a qualitative synthesis to answer the research question from quality criteria and other extracted information, as this strategy is usual in Computer Science [21]. This synthesis yielded the results summarized in Section 4.

First Results from the Systematic Review Conduction. We applied the search strategy in 2017 and updated it in February 2019. Altogether, we found a set of 3477 pieces of work. After applying the 13 exclusion criteria, 153 papers (nearly 4% of the initial set) were chosen. An electronic spreadsheet with all the information extracted from the 153 references is available at http://tiny.cc/4inohy.

4 Approaches for Video Indexing and Retrieval

This section organizes approaches used by the 153 papers found by systematic review into four topics: (1) video segmentation, (2) feature extraction, (3) dimensionality reduction and (4) machine learning. Fig. 2 indicates the most frequent approaches per topic.

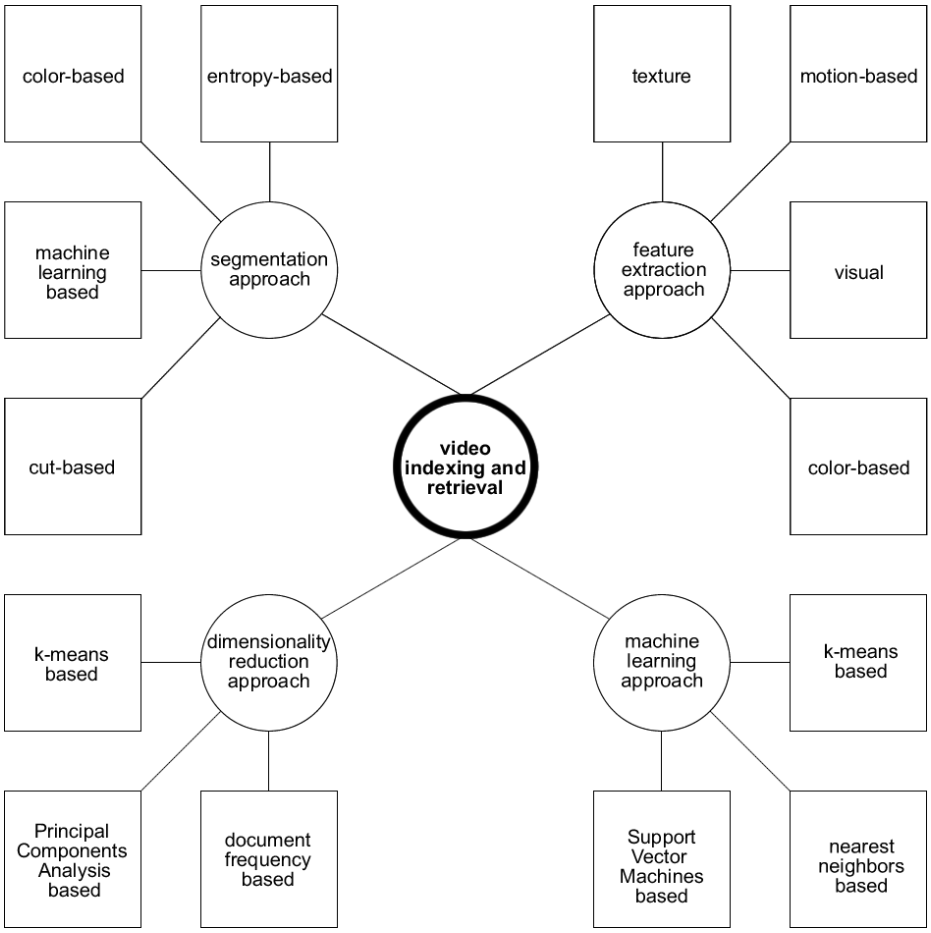

Fig. 2. Usual approaches found in content-based video indexing and retrieval.

4.1 Video Segmentation

We found that 44 selected papers reported the segmentation approach used. Nearly 89% of these publications are associated with Shot Boundary Detection (SBD) [47, 34, 4], while the remaining ones consider alternatives related with frames, scenes or other structures [48, 49]. Table 1 indicates the approaches most commonly applied in the reviewed literature. Some properties of these approaches are considered in what follows. For further detail and discussion regarding the merits of the main approaches, papers cited in this section and literature surveys are appropriate [50, 51, 5].

Table 1. Segmentation approaches found in recent papers.

Approach	Number of occurrences	References
Cut-based	7	[52], [8], [53], [54], [55], [25], [56]
Machine learning	5	[16], [57], [58], [59], [60]

Color-based	4	[61], [62], [63], [64]
Entropy-based	2	[65], [66]

The cut-based category stands out as the most usual one, as indicated by seven references. Black frames, silence segments and other clues are associated with electronic program guide information to break television content into smaller pieces in [8]. Another idea, described in [53], applies the thresholds t_b and t_s on the histogram difference between frames. In particular, t_b is useful to identify hard cuts, while the latter is applied to identify the beginning of a sequence of frames with gradual transition. This sequence ends when the accumulated frame difference reaches t_b.

Segmentation based on a machine learning algorithm has been another usual choice for SBD. The Boosting method used in [59] combines weak classifiers, i.e., classification models slightly better than random guessing, to define a stronger learner able to identify the boundaries. During the learning, this method assigns a weight for each weak classifier before associating them [67]. To feed Boosting, features focusing on the hue channel and optical flow – the apparent motion of something on pairs of consecutive frames – are extracted from the video frames. Another ML alternative, based on Nearest Neighbors (NN), is applied in [57]. Besides supervised Boosting and NN approaches, unsupervised algorithms have also been used to support segmentation [16].

Color-based segmentation can also take into account thresholds. The RGB histogram difference measure reported in [64] considers an iterative frame skip strategy to save computation time in a video with N frames. In particular, the measure is applied between the frames f_i and f_j , i ∈ 1 ... N, j ∈ 1 ... N, which are separated by k frames. If the difference reaches a specific threshold, then a shot boundary is defined.

An alternative category considers Entropy and related measures, such as mutual information [65, 66].

4.2 Feature Extraction

A typical procedure conducted in content-based video retrieval consists in feature or index extraction [5]. As a result, 148 out of the 153 selected papers indicated the features (indexes) extracted from video. Table 2 describes the approaches most frequently used in the 148 mentioned papers, also indicated in the electronical spreadsheet available at http://tiny.cc/4inohy. It should be emphasized that, although 83 reviewed publications consider more than one indexing type, the table counts separately the frequency for each approach. Some properties of these approaches are considered in what follows. For further detail and discussion regarding the merits of the main approaches, pieces of work cited in this section and literature surveys are appropriate [51, 39, 4, 5, 37].

Table 2. Most commonly used feature extraction approaches.

Approach	Number of occurrences
Color-based	56
Visual	38
Motion-based	36

Texture	22

Besides segmentation, color is useful to describe image content. Huang and Chen [64] select two MPEG-7 color descriptors: layout and structure. The former type considers the YCbCr color space and applies a discrete cosine transform to the data. The latter type takes into account the hue-min-max difference space to represent color contents and structural information of image regions.

Another example applying color features can be found in [61], in which the authors employed both global and regional (local) descriptors. The former group extracts, for example, the color value averaged across all frame pixels in a shot. The latter group involves color moments, histograms and other descriptors extracted from pieces of a shot, such as a keyframe.

Visual features in turn are associated, for example, with visual words analogous to textual words in documents [68]. In this context, a bag of visual words approach can be combined with Scale Invariant Feature Transform [69] (SIFT) to characterize videos. An alternative to do so initially detects key points regarding salient regions from (key) frames. Then SIFT and other descriptors are calculated by taking these points as references and grouped to yield visual words ("labels"). Afterwards, statistics are obtained to identify the relevance of the extracted words [70].

A strategy to build a dictionary of visual words by detecting Space-Time Interest Points in frames is reported in [14]. In particular, each point is on the center of a cubic region that is considered to calculate optical flows and histograms of oriented gradient. As a result, visual features are obtained. This technique is also used, for example, in [71, 72].

Different from color, motion is a dynamic property of videos associated with the temporal variation of content [5]. In this context, it is suitable to describe the content of sequences of frames. To extract motion-based features to characterize surgical gestures, [73] propose a motion model based on spatiotemporal polynomials. In particular, the model considers polynomials that approximate the optical flow mapping spatiotemporal coordinates to specific displacements.

The approach used in [74] to describe motion partitions object trajectories into segments. Then, the segments are clustered to yield a codebook with k cluster centers. Finally, a bag of motions is defined from a histogram composed of bins associated with the codebook centers. It should be emphasized that the approach is also used to provide sketch-based queries.

Texture, an important property to characterize images and frame videos, can be seen as an approach to describe local variations that follow specific patterns [37]. These variations typically focus on the neighborhood of pixels delimited by parameters. A literature example applying textons and widely used Haralick's texture features for video retrieval is found in [43]. The main idea of textons is to describe geometric and photometric properties regarding, for example, spots and stripes. On the other hand, Haralick's features are extracted from a matrix representing transitions between pixels.

Another approach to describe the texture from video frames is applied in [57]. Specifically, the technique analyses moments of the distribution of wavelets coefficients at different scales and directions. By employing the idea for each color channel, it also takes into account the frame color content. It should be emphasized that, in this

reference, feature extraction includes descriptors based on motion (optical flow) as well.

4.3 Dimensionality Reduction

We found that 55 out of the 153 selected papers described the approach used for Dimensionality Reduction (DR) in the context of CBVIR. Table 3 shows the frequency of the approaches most employed in the literature. As is the case in the previous section, we count separately the frequency for each table row, such that if a piece of work uses two DR approaches, it is counted for each approach. Some properties of the DR approaches found are considered in what follows. For further detail and discussion regarding the merits of the main approaches, pieces of work cited in this section and references from the DR area are appropriate [75, 76, 77, 23].

Table 3. Most commonly used dimensionality reduction approaches.

Approach	Number of occurrences
k-means based	21
Principal components analysis based	15
Document frequency based	3

We found that k-means based clustering [78] has been the most common approach considered to reduce the video dimensionality. Huang and Chen [64] illustrate this application by clustering the features of each index type extracted from a keyframe into four groups. Each group is numbered from 0 to 3. Afterwards, a 4-digit MPEG-7 signature can be defined, such that each digit consists in the number of the group closest to the corresponding index type extracted from a video. By using the signature as part of the indexing process, the authors achieved reasonable retrieval performance.

One can note that k-means is also useful to group trajectory segments for sketch-based retrieval [74]. Finally, this algorithm and variations are often associated with usual visual words techniques for video indexing [38, 59, 70]. As k-means is an unsupervised learning algorithm, it is applicable in DR problems without class or target feature (Section 2.4).

Principal Components Analysis (PCA) has been frequently used in many domains to transform an original space into another space with less dimensions [45]. An application specific for video retrieval is identified in [11]. In particular, the authors combine PCA with the k-means clustering technique during video preprocessing to group local features into 1000 codewords. Afterwards, each database or query video is described by a set of dimensions. In this scenario, each dimension consists in a tuple with the feature spatio-temporal location and corresponding codeword. It should be emphasized that, although the dimensionality reduction procedure can be costly, it needs to be applied only once to the video database. Another example employing PCA for a diverse set of features is found in [57]. As a result, the authors were able to obtain a more compact and less redundant medical video representation.

A proposal for human action recognition is described in [26]. The method replaces each original video frame with a simplified image, which is projected into a lower dimension space by using a non-linear PCA-based transformation. By doing so,

frames of the entire video sequence trace out a trajectory curve to provide efficient differentiation among actions. PCA was also used in [79] to reduce the number of video indexes.

The mentioned approaches transform an input feature space into another one with new dimensions (Section 2.3). However, there are also methods based on the feature selection scheme, which allows one to weigh and or select original video indexes according to their importance within the CBVIR context. The study reported in [80] adapts document frequency based measures to weigh video indexes. In particular, after representing videos according to visual features, the authors are interested in the occurrence of these features as an estimate of their importance. One can also note that the feature relevance estimation technique [27] calculates the importance of indexes according to the relevance feedback provided by users' subjective judgment on queries. The selection of features in turn is illustrated in [81], which considers the performance achieved by a supervised SVM model built from videos described by specific feature extraction approaches. The best performing models motivate the choice of the corresponding video indexes for further concept-based video retrieval.

4.4 Machine Learning

As mentioned, machine learning has been applied by some video retrieval methods to support video segmentation, indexing and retrieval [5]. As a result of the current systematic review, we found 89 out of the 153 selected papers that described the ML algorithm used. The number relatively high of publications employing machine learning strengthens the relevance of this topic in CBVIR. Table 4 indicates the occurrence of the approaches most used in the literature, counting separately the frequency for each table row. Some properties of these approaches found are considered in what follows. For further detail and discussion regarding the merits of the main approaches, pieces of work cited in this section and references from the ML area are appropriate [20, 82, 22, 5].

Table 4. Most commonly used machine learning algorithms.

Algorithm	Number of occurrences
k-means based	24
Nearest neighbors based	15
Support vector machines based	3

As mentioned in Section 4.3, k-means and variants are popular in CBVIR. Besides the 21 occurrences regarding DR, three papers apply the algorithm to group video images or shots [83, 84, 85]. It should be emphasized that other unsupervised algorithms, such as hierarchical clustering, have also been applied to reduce the number of video indexes, assist relevance feedback, group video shots or support Optical Character Recognition (OCR) [86, 87, 88, 27].

Nearest Neighbors (NN) is another supervised learning algorithm common in the reviewed literature. A typical NN application involves the search for the k videos closest to a user query, as exemplified in [49]. In this piece of work, the similarity measure is highlighted as one of the main method parameters. By speeding up the

dynamic time warping [89] similarity calculation, the authors achieved a fast and accurate retrieval method. The same basic measure is applied in [74] to query similar videos based on motion trajectories. Besides the typical use, one can use NN classification accuracy as an indirect objective indicator for retrieval relevance according to diagnoses based on histopathology [70]. Moreover, NN principles can be associated with similarity approaches to compare image feature signatures in Medicine [90].

Support Vector Machines (SVM) and variants have been one of the most usual ML techniques to assist video retrieval. As illustrated in [91, 11, 27], SVM can be used in relevance feedback as an attempt to enhance retrieval results in further queries and to deal with the gap between descriptors and users' feature perception. Other SVM-based applications include video re-ranking to improve the initial retrieved results [92], segmentation [9], semantic indexing with concepts [43] and video classification or annotation [5].

Limitations. The SR method conducted in this work contains some limitations in its wideness due to the selection criteria described in Section 3.1. The main limitations arose because in the current review we did not consider:

- Relevant research topics associated with 3D elements, object identification/recognition, event detection, near-duplicate and video copy detection;
- Full papers that could not be accessed by our institutions;
- Publications that do not simultaneously focus on content-based video indexing and retrieval, which includes pieces of work focusing on the indexing and retrieval of other media types;
- Patents and academic challenges;
- Papers written in language different than English.

The protocol designed in Section 3.1 should be modified to tackle these issues and cover more potential references, making possible a new systematic review application. We believe that SR dealing with specific research questions to address the mentioned topics inspire future work.

Also, the search string used restricted the extent of this review, as occasionally a relevant publication do not accomplish the following piece of string: *("content-based" OR "semantic-based" OR "context-aware" OR "context-preserving" OR "concept-oriented" OR "keyword-focused")*. However, we decided to keep the string as it is, as removing it would bring a large number of pieces of work that index and retrieve videos regardless of its content.

Another limitation is due to changes in the search tools inherent to bibliographic databases across the years. The first application of the search string in 2017 worked in the seven databases indicated in Section 3.1, yielding most of the publications analyzed by us. However, only Scopus, Web of Science and Wiley databases worked as before during the review update conducted in February 2019. The other databases were not able to process our search string anymore. Future updates should summarize the search string to increase the number of sources to find publications.

5 Future Directions

The use of filter feature selection algorithms, i.e., DR techniques that choose features regardless of machine learning approaches, is incipient in content-based video retrieval. Although 14 reviewed papers used filter methods, only one selects and assesses subsets of features by evaluating the inter-class distance [36]. As a result, most of the methods analyzed are unable to identify redundant/correlated features (indexes) in the original feature space, as they focus on the individual relevance of each index to the class or perform space transformations that create new dimensions. To bridge this gap, one could take into account, for example, the methods Correlation-based Feature Selection or Consistency-based Filter [23], which are publicly available in the Weka framework (www.cs.waikato.ac.nz/ml/weka). Besides traditional linear and nonlinear DR approaches, new researchers in CBVIR can also take into account recent ideas to preserve the geometric structure of the data inherent to video features [93] or deal with streaming data associated with real-time video transmission [76].

As indicated in [5], machine learning algorithms are useful in video retrieval. In fact, most of the selected papers employ a variety of algorithms for different purposes. However, there are still research points regarding machine learning to be more studied in video retrieval. In this context, an issue to be better examined consists in the use of multi-label learning algorithms [94], i.e., algorithms that predict multiple class values for each instance (Section 2.4). This idea stands out as, except for a few cases in conventional or deep learning approaches [16], most of the supervised algorithms applied in the relevant literature predict only one label per instance. By taking into account the label dependence for multi-label semantic video indexing, one could explore, for example, the correlation between concepts (labels) as an additional information to improve the annotation performance [95, 96, 97].

Another issue involves the study of the values assigned to the parameters of machine learning algorithms. Despite of initial efforts into this direction [70], few pieces of work concern on the influence that parameters may have on the learning performance. Indeed, a simple learning method, such as Nearest Neighbors, is already sensitive to its few parameters – especially the dissimilarity measure and the number of neighbors. This issue is even harder to deal with on algorithms associated with more parameters, such as SVM [98]. The use of computationally demanding deep learning approaches also depends on the choice of an appropriate architecture [99].

As noted in [4], few pieces of work used audio to support video retrieval. This finding remains valid, as our SR found only three papers considering audio-based query [1, 100, 101], five references extracting audio-based features and 11 publications using indexes transcribed from speech. Despite of the current limitations on ASR [102], such as relatively low performance on under-resourced languages [24], it is necessary to better investigate audio as an additional information source to retrieve video files. Additional benefits from audio in CBVIR context include its use to support segmentation approaches [103].

6 Final Remarks

This work surveyed the literature on video retrieval and indexing based on content. Altogether, 153 recent papers were summarized and organized into categories regarding video segmentation, indexing, dimensionality reduction and machine learning approaches. This paper updates and extends previous surveys [19, 3, 4, 5, 17, 18] by highlighting dimensionality reduction approaches considered by the selected references, as well as by exploring relevant and recent publications with the replicable systematic review method.

The review protocol can be updated in future work by enabling the selection and summary of recent patents, as they contain innovative ideas closer to actual products. International challenge papers, such as the ones from TRECVID workshops, could also be included to find research insights. Extending the systematic review method to answer specific research questions on trendy topics, such as deep learning, big data or other video technology issues, can also be a future direction.

Acknowledgments. This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and the Graduate Program in Electrical Engineering and Computer Science at Western Paraná State University (UNIOESTE).

References

1. Guo, K., Pan, W., Lu, M., Zhou, X., Ma, J.: An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval. J Syst Softw 102, 207–216 (2015). https://doi.org/http://dx.doi.org/10.1016/j.jss.2014.09.016
2. Benois-Pineau, J., Precioso, F., Cord, M.: Visual Indexing and Retrieval. Springer-Verlag New York, New York, United States (2012)
3. Bhaumik, H., Bhattacharyya, S., Nath, M.D., Chakraborty, S.S.: Hybrid soft computing approaches to content based video retrieval: A brief review. Appl Soft Comput 46, 1008–1029 (2016). https://doi.org/http://dx.doi.org/10.1016/j.asoc.2016.03.022
4. Priya, R., Shanmugam, T.N.: A comprehensive review of significant researches on content based indexing and retrieval of visual information. Front Computer Science 7(5), 782–799 (2013). https://doi.org/10.1007/s11704-013-1276-6
5. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41(6), 797–819 (2011). https://doi.org/10.1109/TSMCC.2011.2109710
6. Luong, T.H., Pham, N.M., Vu, Q.H.: Vietnamese Multimedia Agricultural Information Retrieval System as an Info Service, pp. 147–160. Springer International Publishing, Cham, Switzerland (2016). https://doi.org/10.1007/978-3-319-31468-6_11
7. Mitrović, D., Zeppelzauer, M., Zaharieva, M., Breiteneder, C.: Retrieval of visual composition in film. In: International Workshop on Image Analysis for Multimedia Interactive Services. pp. 1–4. TU Delft, Delft, The Netherlands (2011)
8. Pereira, M.H.R., de Souza, C.L., Pádua, F.L.C., Silva, G.D., de Assis, G.T., Pereira, A.C.M.: Sapte: A multimedia information system to support the discourse analysis and in-

formation retrieval of television programs. Multimedia Tools and Applications 74(23), 10923–10963 (2015). https://doi.org/10.1007/s11042-014-2311-9
9. Yang, H., Meinel, C.: Content based lecture video retrieval using speech and video text information. IEEE Trans Learn Technol 7(2), 142–154 (2014)
10. Yin, Y., Seo, B., Zimmermann, R.: Content vs. context: Visual and geographic information use in video landmark retrieval. ACM Transactions on Multimedia Computing, Communications and Applications 11(3), 39:1–39:21 (2015). https://doi.org/10.1145/2700287
11. Shao, L., Jones, S., Li, X.: Efficient search and localization of human actions in video databases. IEEE Trans Circuits Syst Video Technol 24(3), 504–512 (2014). https://doi.org/10.1109/TCSVT.2013.2276700
12. Younessian, E., Rajan, D.: Multi-modal solution for unconstrained news story retrieval. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.W., Andreopoulos, Y., Breiteneder, C. (eds.) Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. 7131, pp. 186–195. Springer Berlin Heidelberg, Berlin, Germany (2012)
13. Sharma, R., Mummareddy, S., Hershey, J., Jung, N.: Method and system for analyzing shopping behavior in a store by associating RFID data with video-based behavior and segmentation data. Patent (2013), US 8380558
14. Charrière, K., Quellec, G., Lamard, M., Coatrieux, G., Cochener, B., Cazuguel, G.: Automated surgical step recognition in normalized cataract surgery videos. In: International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4647–4650. IEEE, Chicago, IL, United States (2014). https://doi.org/10.1109/EMBC.2014.6944660
15. Al Kabary, I., Schuldt, H.: Enhancing sketch-based sport video retrieval by suggesting relevant motion paths. In: International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1227–1230. ACM, New York, NY, United States (2014). https://doi.org/10.1145/2600428.2609551
16. Mühling, M., Meister, M., Korfhage, N., Wehling, J., H¨orth, A., Ewerth, R., Freisleben, B.: Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive, pp. 67–78. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-43997-6_6
17. Smeaton, A.F.: Techniques used and open challenges to the analysis, indexing and retrieval of digital video. Inf Syst 32(4), 545–559 (2007). https://doi.org/http://dx.doi.org/10.1016/j.is.2006.09.001
18. Antani, S., Kasturi, R., Jain, R.: A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognit 35(4), 945–965 (2002). https://doi.org/10.1016/S0031-3203(01)00086-3
19. Müller, H., Unay, D.: Retrieval from and understanding of large-scale multi-modal medical datasets: A review. IEEE Trans Multimedia 19(9), 2093–2104 (2017). https://doi.org/10.1109/TMM.2017.2729400
20. Münzer, B., Schoeffmann, K., B¨osz¨ormenyi, L.: Content-based processing and analysis of endoscopic images and videos: A survey. Multimedia Tools and Applications 77(1), 1323–1362 (Jan 2018). https://doi.org/10.1007/s11042-016-4219-z
21. Kitchenham, B.A., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Evidence-based Software Engineering Technical Report (2007)
22. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann, Burlington, MA, United States (2011)
23. Liu, H., Motoda, H.: Computational methods of feature selection. Chapman and Hall/CRC, Boca Raton, United States (2007)

24. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Commun 56, 85–100 (2014). https://doi.org/http://dx.doi.org/10.1016/j.specom.2013.07.008
25. Kant, S.: Activity-based exploitation of full motion video (fmv). Proc SPIE 8386, 83860D–83860D–11 (2012). https://doi.org/10.1117/12.920280
26. Goméz-Conde, I., Olivieri, D.N.: A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system. Expert Systems with Applications 42(13), 5472–5490 (2015). https://doi.org/http://dx.doi.org/10.1016/j.eswa.2015.03.010
27. Mironica, I., Vertan, C., Ionescu, B.: A relevance feedback approach to video genre retrieval. In: IEEE International Conference on Intelligent Computer Communication and Processing. pp. 327–330. IEEE, Cluj-Napoca, Romania (2011). https://doi.org/10.1109/ICCP.2011.6047890
28. Raieli, R.: Multimedia Information Retrieval: Theory and Techniques. Chandos Publishing, Oxford, United Kingdom, 1st edn. (2013)
29. Fan, J., Zhu, X., Xiao, J.: Content-based video indexing and retrieval. In: DiMarco, J. (ed.) Computer Graphics and Multimedia: Applications, Problems and Solutions, pp. 110–144. IGI Global, Hershey, PA, United States (2004). https://doi.org/10.4018/978-1-59140-196-4.ch007
30. Petković, M., Jonker, W.: Content-Based Video Retrieval - a database perspective. Springer US, New York, NY, United States, 1 edn. (2004)
31. Lelescu, D., Schonfeld, D.: Video skimming and summarization based on principal component analysis. In: Al-Shaer, E.S., Pacifici, G. (eds.) Management of Multimedia on the Internet, Lecture Notes in Computer Science, vol. 2216, pp. 128–141. Springer Berlin Heidelberg, Berlin, Germany (2001)
32. Gargi, U., Kasturi, R., Strayer, S.H.: Performance characterization of video-shot-change detection methods. IEEE Trans Circuits Syst Video Technol 10(1), 1–13 (2000)
33. Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B.: A formal study of shot boundary detection. IEEE Trans Circuits Syst Video Technol 17(2), 168–186 (2007). https://doi.org/10.1109/TCSVT.2006.888023
34. SenGupta, A., Thounaojam, D.M., Singh, K.M., Roy, S.: Video shot boundary detection: A review. In: IEEE International Conference on Electrical, Computer and Communication Technologies. pp. 1–6. IEEE, Coimbatore, India (2015). https://doi.org/10.1109/ICECCT.2015.7226084
35. Safadi, B., Sahuguet, M., Huet, B.: When textual and visual information join forces for multimedia retrieval. In: International Conference on Multimedia Retrieval. pp. 265–265. ACM, New York, NY, USA (2014)
36. Ji, X., Han, J., Hu, X., Li, K., Deng, F., Fang, J., Guo, L., Liu, T.: Retrieving video shots in semantic brain imaging space using manifold-ranking. In: IEEE International Conference on Image Processing. pp. 3633–3636. IEEE, Brussels, Belgium (2011). https://doi.org/10.1109/ICIP.2011.6116505
37. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, Upper Saddle River, NJ, United States, 3 edn. (2007)
38. Shen, X., Zhang, L., Wang, Z., Feng, D.: Spatial-temporal correlation for trajectory based action video retrieval. In: IEEE International Workshop on Multimedia Signal Processing. pp. 1–6. IEEE, Xiamen, China (2015). https://doi.org/10.1109/MMSP.2015.7340811
39. Inoue, N., Shinoda, K.: Semantic indexing for large-scale video retrieval. ITE Transactions on Media Technology and Applications 4(3), 209–217 (2016). https://doi.org/10.3169/mta.4.209

40. Zha, Z.J., Wang, M., Zheng, Y.T., Yang, Y., Hong, R., Chua, T.S.: Interactive video indexing with statistical active learning. IEEE Trans Multimedia 14(1), 17–27 (2012). https://doi.org/10.1109/TMM.2011.2174782
41. Memar, S., Affendey, L.S., Mustapha, N., Doraisamy, S.C., Ektefa, M.: An integrated semantic-based approach in concept based video retrieval. Multimedia Tools and Applications 64(1), 77–95 (2013). https://doi.org/10.1007/s11042-011-0848-4
42. Wei, X.Y., Yang, Z.Q.: Coaching the exploration and exploitation in active learning for interactive video retrieval. IEEE Trans Image Process 22(3), 955–968 (2013). https://doi.org/10.1109/TIP.2012.2222902
43. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: A smart atlas for endomicroscopy using automated video retrieval. Méd Image Anal 15(4), 460–476 (2011). https://doi.org/10.1016/j.media.2011.02.003
44. Huurnink, B., Snoek, C.G.M., de Rijke, M., Smeulders, A.W.M.: Content-based analysis improves audiovisual archive retrieval. IEEE Trans Multimedia 14(4), 1166–1178 (2012). https://doi.org/10.1109/TMM.2012.2193561
45. Jackson, J.E.: A user's guide to principal components. Wiley-Interscience, New York, United States (2003)
46. Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D.: A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing 180, 3–15 (2016). https://doi.org/10.1016/j.neucom.2015.07.118
47. Münzer, B., Primus, M.J., Hudelist, M., Beecks, C., H¨urst, W., Schoeffmann, K.: When content-based video retrieval and human computation unite: Towards effective collaborative video search. In: IEEE International Conference on Multimedia Expo Workshops (ICMEW). pp. 214–219 (2017). https://doi.org/10.1109/ICMEW.2017.8026262
48. Ngo, T.T., Vo, D.: A novel content based scene retrieval using multi-frame features. In: International Conference on Advanced Technologies for Communications. pp. 105–108. IEEE, Hanoi, Vietnam (2014). https://doi.org/10.1109/ATC.2014.7043365
49. Cui, H., Zhu, M.: A novel multi-metric scheme using dynamic time warping for similarity video clip search. In: IEEE International Conference on Signal Processing, Communication and Computing. pp. 1–5. IEEE, KunMing, China (2013). https://doi.org/10.1109/ICSPCC.2013.6663926
50. Abdulhussain, S.H., Ramli, A.R., Saripan, M.I., Mahmmod, B.M., Al-Haddad, S.A.R., Jassim, W.A.: Methods and challenges in shot boundary detection: A review. Entropy 20(4), 1–42 (2018). https://doi.org/10.3390/e20040214
51. Ushapreethi, P., Lakshmipriya, G.G.: Survey on video big data: Analysis methods and applications. Int J Appl Eng Res 12, 2221–2231 (01 2017)
52. Li, K., Li, S., Oh, S., Fu, Y.: Videography-based unconstrained video analysis. IEEE Transactions on Image Processing 26(5), 2261–2273 (2017). https://doi.org/10.1109/TIP.2017.2678800
53. Asha, S., Sreeraj, M.: Content based video retrieval using surf descriptor. In: International Conference on Advances in Computing and Communications. pp. 212–215. IEEE, Cochin, India (2013). https://doi.org/10.1109/ICACC.2013.49
54. Wu, G.L., Kuo, Y.H., Chiu, T.H., Hsu, W.H., Xie, L.: Scalable mobile video retrieval with sparse projection learning and pseudo label mining. IEEE Multimedia 20(3), 47–57 (2013). https://doi.org/10.1109/MMUL.2013.13
55. Loganathan, D., Jamal, J., Nijanthan, P., Balamurugan, V.K.: Advances in Communication, Network, and Computing International Conference, chap. Enhanced Video Indexing and Retrieval Based on Face Recognition through Combined Detection and Fast LDA, pp.

351–357. Springer Berlin Heidelberg, Berlin (2012). https://doi.org/10.1007/978-3-642-35615-5_56
56. Lakshmi Rupa, G., Gitanjali, J.: A video mining application for image retrieval. International Journal of Computer Applications 20(3), 46–51 (2011). https://doi.org/10.5120/2410-3214
57. Quellec, G., Lamard, M., Droueche, Z., Cochener, B., Roux, C., Cazuguel, G.: A polynomial model of surgical gestures for real-time retrieval of surgery videos. In: International Conference on Medical Content-Based Retrieval for Clinical Decision Support. pp. 10–20. Springer Berlin Heidelberg, Berlin, Germany (2013). https://doi.org/10.1007/978-3-642-36678-9_2
58. Quellec, G., Charrière, K., Lamard, M., Droueche, Z., Roux, C., Cochener, B., Cazuguel, G.: Real-time recognition of surgical tasks in eye surgery videos. Méd Image Analysis 18(3), 579–590 (2014). https://doi.org/10.1016/j.media.2014.02.007
59. Choi, J., Wang, Z., Lee, S.C., Jeon, W.J.: A spatio-temporal pyramid matching for video retrieval. Computer Vis Image Underst 117(6), 660–669 (2013). https://doi.org/10.1016/j.cviu.2013.02.003
60. Quellec, G., Lamard, M., Cazuguel, G., Droueche, Z., Roux, C., Cochener, B.: Real-time retrieval of similar videos with application to computer-aided retinal surgery. In: International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4465–4468. IEEE, Boston, MA, United States (2011). https://doi.org/10.1109/IEMBS.2011.6091107
61. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A multi-feature sketch-based video retrieval engine. In: IEEE International Symposium on Multimedia. pp. 18–23. IEEE, Taichung, Taiwan (2014). https://doi.org/10.1109/ISM.2014.38
62. Daga, B.: Advances in Computing, Communication, and Control International Conference, chap. Content Based Video Retrieval Using Color Feature: An Integration Approach, pp. 609–625. Springer Berlin Heidelberg, Berlin (2013). https://doi.org/10.1007/978-3-642-36321-4_57
63. Liang, B., Xiao, W., Liu, X.: Design of video retrieval system using mpeg-7 descriptors. Procedia Engineering 29, 2578–2582 (2012). https://doi.org/10.1016/j.proeng.2012.01.354
64. Huang, Y.F., Chen, H.W.: Active Media Technology International Conference, chap. A Multi-type Indexing CBVR System Constructed with MPEG-7 Visual Features, pp. 71–82. Springer Berlin Heidelberg, Berlin (2011). https://doi.org/10.1007/978-3-642-23620-4_11
65. Kamde, P.M., Shiravale, S., Algur, S.P.: Entropy supported video indexing for content based video retrieval. International Journal of Computer Applications 62(17), 1–6 (2013)
66. Yarmohammadi, H., Rahmati, M., Khadivi, S.: Content based video retrieval using information theory. In: Iranian Conference on Machine Vision and Image Processing. pp. 214–218. IEEE, Zanjan, Iran (2013). https://doi.org/10.1109/IranianMVIP.2013.6779981
67. Zhi-Hua, Z.: Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, Boca Raton, United States (2012)
68. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2036–2043. IEEE, Miami, FL, United States (2009). https://doi.org/10.1109/CVPR.2009.5206718
69. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. pp. 1150–1157. IEEE, Kerkyra, Greece (1999). https://doi.org/10.1109/ICCV.1999.790410

70. André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N.: Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Trans Méd Imaging 31(6), 1276–1288 (2012). https://doi.org/10.1109/TMI.2012.2188301
71. Ramezani, M., Yaghmaee, F.: Motion pattern based representation for improving human action retrieval. Multimedia Tools and Applications. 77(19), 26009–26032 (2018). https://doi.org/10.1007/s11042-018-5835-6
72. Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B.: Real-time analysis of cataract surgery videos using statistical models. Multimedia Tools and Applications 76(21), 22473–22491 (2017). https://doi.org/10.1007/s11042-017-4793-8
73. Quellec, G., Lamard, M., Cochener, B., Cazuguel, G.: Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans Méd Imaging 33(12), 2352–2360 (2014). https://doi.org/10.1109/TMI.2014.2340473
74. Ghosal, K., Namboodiri, A.: A sketch-based approach to video retrieval using qualitative features. In: Indian Conference on Computer Vision Graphics and Image Processing. pp. 1–8. ACM, New York, NY, United States (2014). https://doi.org/10.1145/2683483.2683537
75. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79 (2018). https://doi.org/https://doi.org/10.1016/j.neucom.2017.11.077
76. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Comput Surv 50(6), 94:1–94:45 (2017). https://doi.org/10.1145/3136625
77. Praveena, M.D.A., Bharathi, B.: A survey paper on big data analytics. In: International Conference on Information Communication and Embedded Systems. pp. 1–9 (2017)
78. Hennig, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of cluster analysis. Chapman and Hall/CRC, Boca Raton, United States, 1 edn. (2015)
79. Liu, Y., Sui, A.: Research on feature dimensionality reduction in content based public cultural video retrieval. In: IEEE/ACIS International Conference on Computer and Information Science. pp. 718–722 (2018). https://doi.org/10.1109/ICIS.2018.8466379
80. Murata, M., Nagano, H., Mukai, R., Kashino, K., Satoh, S.: Bm25 with exponential idf for instance search. IEEE Trans Multimedia 16(6), 1690–1699 (2014). https://doi.org/10.1109/TMM.2014.2323945
81. Guo, X., Zhao, Z., Chen, Y., Cai, A.: An improved system for concept-based video retrieval. In: IEEE International Conference on Network Infrastructure and Digital Content. pp. 391–395. IEEE, Beijing, China (2012). https://doi.org/10.1109/ICNIDC.2012.6418781
82. Puthenputhussery, A., Chen, S., Lee, J., Spasovic, L., Liu, C.: Learning and recognition methods for image search and video retrieval. In: Liu, C. (ed.) Recent Advances in Intelligent Image Search and Video Retrieval. pp. 21–43. Springer International Publishing, Cham (2017)
83. Chamasemani, F.F., Affendey, L.S., Mustapha, N., Khalid, F.: Surveillance video retrieval using effective matching techniques. In: International Conference on Information Retrieval and Knowledge Management. pp. 137–141 (2018). https://doi.org/10.1109/INFRKM.2018.8464772
84. Kulkarni, P., Patil, B., Joglekar, B.: An effective content based video analysis and retrieval using pattern indexing techniques. In: International Conference on Industrial Instrumentation and Control. pp. 87–92. IEEE, Pune, India (2015). https://doi.org/10.1109/IIC.2015.7150717

85. Kumar, C.R., Sujatha, S.N.N.: Star: Semi-supervised-clustering technique with application for retrieval of video. In: International Conference on Intelligent Computing Applications. pp. 223–227. IEEE, Coimbatore, India (2014). https://doi.org/10.1109/ICICA.2014.55
86. Wattanarachothai, W., Patanukhom, K.: Key frame extraction for text based video retrieval using maximally stable extremal regions. In: International Conference on Industrial Networks and Intelligent Systems. pp. 29–37. European Alliance for Innovation, Begijnhoflaan, Belgium (2015). https://doi.org/10.4108/icst.iniscom.2015.258410
87. Anh, T.Q., Bao, P., Khanh, T.T., Thao, B.N.D., Tuan, T.A., Nhut, N.T.: A content based video retrieval analysis system with extensive features by using kullback-leibler. International Journal of Computational Intelligence Systems 8(6), 853–858 (2012)
88. Amiri, A., Abdollahi, N., Jafari, M., Fathy, M.: Hierarchical Key-Frame Based Video Shot Clustering Using Generalized Trace Kernel, pp. 251–257. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-27337-7_23
89. Muller, M.: Information Retrieval for Music and Motion. Springer, New York, United States (2007)
90. Schoeffmann, K., Beecks, C., Lux, M., Uysal, M.S., Seidl, T.: Content-based retrieval in videos from laparoscopic surgery. Proc SPIE 9786, 9786–1–9786–10 (2016). https://doi.org/10.1117/12.2216864
91. Ramezani, M., Yaghmaee, F.: Retrieving human action by fusing the motion information of interest points. International Journal on Artificial Intelligence Tools 27(3), 1850008–1–1850008–18 (2018). https://doi.org/10.1142/S0218213018500082
92. Yu, S.I., Jiang, L., Xu, Z., Yang, Y., Hauptmann, A.G.: Content-based video search over 1 million videos with 1 core in 1 second. In: ACM on International Conference on Multimedia Retrieval. pp. 419–426. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2671188.2749398
93. Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A.G., Zheng, Q.: Adaptive unsupervised feature selection with structure regularization. IEEE Trans Neural Netw Learn Syst 29(4), 944–956 (2018). https://doi.org/10.1109/TNNLS.2017.2650978
94. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. Data Min Knowl Discov Handb pp. 1–19 (2009)
95. Halder, K., Poddar, L., Kan, M.Y.: Cold start thread recommendation as extreme multi-label classification. In: Companion Proceedings of The Web Conference. pp. 1911–1918 (2018). https://doi.org/10.1145/3184558.3191659
96. Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.C.: Categorizing feature selection methods for multi-label classification. Artificial Intelligence Review 49(1), 57–78 (2018). https://doi.org/10.1007/s10462-016-9516-4
97. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8), 1819–1837 (2014). https://doi.org/10.1109/TKDE.2013.39
98. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine learning 46(1-3), 131–159 (2002)
99. Zheng, L., Yang, Y., Tian, Q.: Sift meets cnn: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(5), 1224–1244 (2018). https://doi.org/10.1109/TPAMI.2017.2709749
100. Pranali, B., Anil, W., Kokhale, S.: Inhalt based video recuperation system using OCR and ASR technologies. In: International Conference on Computational Intelligence and Communication Networks. pp. 382–386. IEEE, Jabalpur, India (2015). https://doi.org/10.1109/CICN.2015.315

101. Vigneshwari, G., Juliet, A.N.M.: Optimized searching of video based on speech and video text content. In: International Conference on Soft-Computing and Networks Security. pp. 1–4. IEEE, Coimbatore, India (2015). https://doi.org/10.1109/ICSNS.2015.7292369
102. Spille, C., Kollmeier, B., Meyer, B.T.: Comparing human and automatic speech recognition in simple and complex acoustic scenes (in press). Computer Speech & Language (2018). https://doi.org/https://doi.org/10.1016/j.csl.2018.04.003
103. Cao, Y., Tavanapong, W., Li, D., Oh, J., de Groen, P.C., Wong, J.: A Visual Model Approach for Parsing Colonoscopy Videos, pp. 160–169. Springer Berlin Heidelberg, Berlin, Germany (2004). https://doi.org/10.1007/978-3-540-27814-6_22