Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A systematic review on content-based video retrieval☆

Newton Spolaôr[a], Huei Diana Lee[a,*], Weber Shoity Resende Takaki[a,b], Leandro Augusto Ensina[a], Claudio Saddy Rodrigues Coy[b], Feng Chung Wu[a,b]

[a] *Laboratory of Bioinformatics, Western Paraná State University, Foz do Iguaçu, Paraná, Presidente Tancredo Neves Avenue, 6731, ZIP code: 85867-900, Brazil*
[b] *Service of Coloproctology, Faculty of Medical Sciences, University of Campinas, Campinas, São Paulo, Brazil*

ARTICLE INFO

Keywords:
Color features
Unsupervised learning
Shot boundary detection

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.

1. Introduction

Multimedia documents composed of different media types have increasingly been published and consumed (Guo et al., 2015; Benois-Pineau et al., 2012). This fact is due to the larger access to computational resources and the Internet, among other reasons (Bhaumik et al., 2016). 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 (Priya and Shanmugam, 2013).

In general, video content is richer than single image content (Hu et al., 2011). 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 (features) 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 (Luong et al., 2016), cinema (Mitrović et al., 2011), discourse analysis (Pereira et al., 2015), education (Yang and Meinel, 2014), geo-referenced video (Yin et al., 2015), human action recognition (Shao et al., 2014), journalism (Younessian and Rajan, 2012), marketing (Sharma et al., 2013), medicine (Charrière et al., 2014), sports (Al Kabary and Schuldt, 2014) and television broadcast (Mühling et al., 2016).

To support video indexing and retrieval, some additional topics can be useful (Puthenputhussery et al., 2017; Priya and Shanmugam, 2013; Hu et al., 2011). Video segmentation is a classical preliminary step typically implemented to separate a video into several units that potentially improve indexing (Pereira et al., 2015; Yuan et al., 2007; Lelescu and Schonfeld, 2001). Dimensionality Reduction (DR) techniques, usual in data mining research (Han and Kamber, 2011), represent an alternative to keep only relevant and non-redundant video indexes or combine the original indexes to create new features (Shao et al., 2014; Huang and Chen, 2011). Machine Learning (ML), in turn, can be associated with the remaining topics, for example, to discover segment boundaries or learn meaningful video indexes (Yang and Meinel, 2014; Choi et al., 2013; André et al., 2012). ML has also been useful to support video retrieval by searching for the nearest videos of a user query or reranking an initial list of retrieved videos (Yu et al., 2015; Cui and Zhu, 2013). Fig. 1 summarizes a possible workflow combining the mentioned topics for Content-based Video Indexing and Retrieval (CBVIR). The workflow is expanded later by including the results analyzed by this work.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2020.103557.

* Corresponding author.
E-mail addresses: huei.lee@unioeste.br (H.D. Lee), newtonspolaor@gmail.com (N. Spolaôr).

https://doi.org/10.1016/j.engappai.2020.103557
Received 3 October 2019; Received in revised form 7 December 2019; Accepted 10 February 2020
Available online xxxx

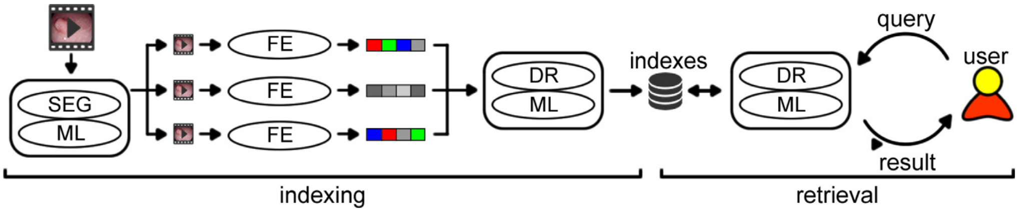

Fig. 1. A possible workflow for content-based video indexing and retrieval based on the Segmentation (SEG), Machine Learning (ML), Feature Extraction for video indexing (FE) and Dimensionality Reduction (DR) topics.

Some surveys identify relevant research on CBVIR. One of them describes the background and an extensive review of CBVIR methods and results (Priya and Shanmugam, 2013). In Hu et al. (2011) 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, Smeaton (2007) reports research challenges. Another example, published in Antani et al. (2002), focuses on the use of intelligent methods for several tasks, such as multimedia indexing and retrieval. ML methods for video retrieval, including deep learning approaches, are discussed in Puthenputhussery et al. (2017). Soft computing arises as an alternative due to the ability to find inexact solutions (Bhaumik et al., 2016). Recently, a Systematic literature Review (SR) on content-based multimedia medical retrieval that concentrates efforts on medical images was conducted (Müller and Unay, 2017). Besides image retrieval, the search for medical video is addressed in Münzer et al. (2018).

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 (Kitchenham and Charters, 2007).

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 (Han and Kamber, 2011). This curse involves phenomena regarding the increasing sparsity of the data as the number of dimensions grows (Liu and Motoda, 2007). 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, this work reviews segmentation, feature extraction and machine learning approaches due to their frequent use in the literature and their discussion in previous surveys (Puthenputhussery et al., 2017; Priya and Shanmugam, 2013; Hu et al., 2011);
- 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 Cedillo-Hernandez et al. (2014), 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 Sections 4 and 5. Sections 6 and 7, 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 (Priya and Shanmugam, 2013). 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 (Hu et al., 2011). 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 (Besacier et al., 2014) 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 (Kant, 2012). In addition, one can note the use of measures designed specifically for video, such as a measure to differentiate trajectories (Goméz-Conde and Olivieri, 2015).

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 Mironica et al. (2011).

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 (Raieli, 2013; Fan et al., 2004; Hu et al., 2011; Petković and Jonker, 2004).

2.1. Video segmentation

This task divides a video into segments of related frames (Lelescu and Schonfeld, 2001). A common segment type, named shot, corresponds to a frame sequence potentially cohesive representing an action. This action, continuous in space and time, is recorded by a simple camera operation (Gargi et al., 2000). 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 (Yang and Meinel, 2014). To identify these subsequences, slides can be used as separators.

The segmentation of generic video into shots consists in an important research topic (Yuan et al., 2007). 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 (SenGupta et al., 2015; Priya and Shanmugam, 2013). This work surveys recent ones in CBVIR context in Section 4.1.

Many SBD approaches can be organized into two main categories according to the shot transition type: abrupt and gradual. The former group includes methods that capture a boundary in a single frame. To illustrate abrupt transitions, consider a documentary video showing natural landscapes. A cut would be detected between shots, each one associated with a different landscape, if the last frame of the first shot is followed by the first frame of the second one. In turn, the gradual group includes approaches that identify transitions regarding, for example, a fade-out. A fade-out can be understood as a series of frames representing a slight decrease of brightness that frequently ends in a solid black frame. Fade-in, dissolve and wipe are other frequent gradual transitions (SenGupta et al., 2015).

Beyond brightness, other frame and video properties are useful to support boundary detection, such as color, texture and shape (Dergachyova et al., 2016; Kamde et al., 2013; Huang and Chen, 2011), audio (Pereira et al., 2015; Yu Cao et al., 2004), features learned by advanced machine learning algorithms (Twinanda et al., 2017), motion and objectness (Varytimidis et al., 2016; Primus et al., 2013). In particular, these properties can feed Euclidean distance or other similarity measures used by SBD approaches. In turn, the SBD methods apply thresholds or machine learning algorithms, for example, to identify relevant dissimilarities as boundaries (Hu et al., 2011).

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. Identifying keyframes with low redundancy for summarization purpose in medical videos, without clear shot boundaries, is illustrated in Schoeffmann et al. (2015). The proposed method is able to detect as representative unblurred frames that show anatomical structures, such as blood vessels. Video browsing approaches could then focus on these images to give users faster access to valuable content.

Although irrelevant frames may also be present in videos, some literature approaches can deal with this issue. For example, the proposal described in Münzer et al. (2013) segments video through four steps: pre-processing, indicator (feature) extraction, fuzzy frame classification and segment composition. The first step omits irrelevant pixels to ease the calculation of varied features from frames, such as the Difference of Gaussians – also used in Schoeffmann et al. (2015) to estimate the blurring level. Afterward, the proposal trains a fuzzy machine learning algorithm to assign for each frame a membership score in predefined classes. Finally, the method builds segments with a specific minimum size that consist, mainly, of frames labeled as irrelevant ones. By implementing the method, a video player could, for example, hide irrelevant segments to give the user better navigation.

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 (Safadi et al., 2014). This combination is obtained, for example, from information extracted from text, image or audio inherent to a video. Abrupt and gradual scene transition detection is illustrated in Ngo and Vo (2014) with a technique grounded on pixel and histogram-based differences between frames, edge detection and dilation operations (Huang and Liao, 2001).

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 (Priya and Shanmugam, 2013): 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 Ji et al. (2011). 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 (Hu et al., 2011). Some descriptors are also used for image processing in general (Gonzalez and Woods, 2007). Usual descriptors include bag of visual words (Shen et al., 2015).

Concepts or semantic indexes are assigned to videos by different approaches, such as manual or automatic annotation (Inoue and Shinoda, 2016; Zha et al., 2012). 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 (Han and Kamber, 2011) (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 (Inoue and Shinoda, 2016; Guo et al., 2015; Memar et al., 2013; Wei and Yang, 2013; André et al., 2011; Huurnink et al., 2012; Ji et al., 2011).

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 (Liu and Motoda, 2007). 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 confused 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 (Jackson, 2003). 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 (Liu and Motoda, 2007). 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.

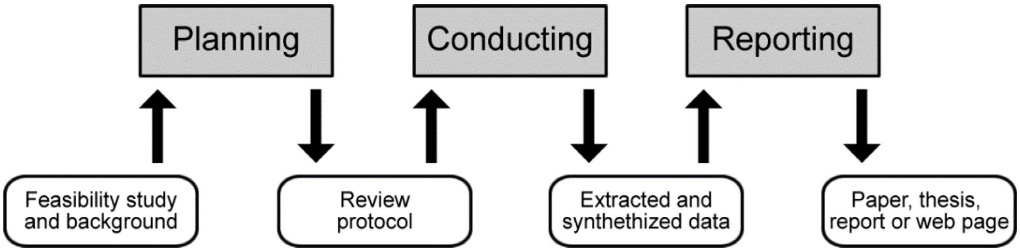
Planning
Conducting
Reporting
Feasibility study and background
Review protocol
Extracted and synthethized data
Paper, thesis, report or web page

Fig. 2. Systematic literature review workflow. *Source*: This figure is an adaptation of material published in Spolaôr et al., 2016. Any citation to the material should consider that paper.

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 (Han and Kamber, 2011).

Let D be a dataset typically submitted for ML, composed of N instances E_i, $i = 1, \dots, N$. A vector $x_i = (x_{i1}, x_{i2}, \dots, x_{iM})$ composed of M features x_j, $j = 1, \dots, M$ 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 (Bhaumik et al., 2016). 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 (Han and Kamber, 2011), 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 (Hu et al., 2011). 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) (Kitchenham and Charters, 2007). Fig. 2 summarizes the workflow regarding the three SR steps and their relevant inputs and outputs (Spolaôr et al., 2016).

The first step receives two inputs: a feasibility study and the background on the research questions to be answered (Spolaôr et al., 2016). 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 (Kitchenham and Charters, 2007). 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 (Müller and Unay, 2017; Bhaumik et al., 2016; Priya and Shanmugam, 2013; Hu et al., 2011; Smeaton, 2007; Antani et al., 2002). 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 (Kitchenham and Charters, 2007), 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 audiovisual 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.

- 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 (Han and Kamber, 2011). This curse involves phenomena regarding the increasing sparsity of the data as the number of dimensions grows (Liu and Motoda, 2007). 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;
- 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 cannot 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 (Kitchenham and Charters, 2007). 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.

3.2. *Justification for the selection criteria chosen*

The literature in CBVIR is vast and diverse. Applying selection criteria in thousands of candidate publications leads to the removal of research papers that fall within a criterion, even if these papers are important for several CBVIR researchers. In what follows, each criterion is justified. It should be emphasized that future SR could answer a more specific research question, focusing only on topics associated with some current criteria.

- EC1. Searching for objects and other elements with three dimensions is useful in Medicine and other applications. Research in video features (Kumar and Suguna, 2016) and dissimilarity measures (Gregor et al., 2015) illustrate the importance of this topic. Furthermore, tasks involving objects, such as identification or recognition, can still be challenging due to the difficulty to efficiently find several objects in different scenes (Hu et al., 2011). Although these issues are relevant, CBVIR associated with 3D elements was not considered in this work. The reader interested in the topic can refer to recent papers that focus on objects, such as Ranjith Kumar and Suguna (2018);
- EC2. Publications composed of a single page makes harder to find all the information required by an SR. Other publication formats, such as posters, slides and event programs, usually lead to the same issue;
- EC3. A relevant survey was published in 2011 (Hu et al., 2011), citing papers up to 2010. Later, other good reviews were published, as described in Section 1. Thus, we decided to focus on papers published from 2011 onwards;
- EC4. The search string contains keywords that may be present in relevant papers. However, it can return manuscripts that have the keywords but are unable to answer the question, as they deal with unrelated issues. An example consists in papers dealing exclusively with content-based image retrieval, such as Ayadi et al. (2016). Also, some publications not covered by the search can also be returned, as illustrated by Fei et al. (2016), due to the complexity of the string. EC4 was used to remove these publications;
- EC5. Only one copy of a set of duplicated papers was kept to reduce the number of candidate publications to analyze;
- EC6. Although ignoring languages limits the SR wideness, the language kept (English) is used in a large number of relevant references;
- EC7. If the full paper of a copyrighted publication cannot be accessed by our institutions, it could be purchased. However, if this fact happens with a large number of publications, financial support could be needed. As this SR was not commissioned, we decided to include this criterion;
- EC8. This criterion was designed to remove papers that contribute to different topics but do not simultaneously focus on the content-based indexing and retrieval of video. Examples include papers developing approaches to index and retrieve image, video, text and audio, which will be better addressed by surveys dealing with multimedia retrieval (Pouyanfar et al., 2018). EC8 was also used to remove papers focusing on event detection (Hu et al., 2011). We believe that a specific SR could be proposed to explore and summarize recent research on this valuable topic (Fan et al., 2017; Liu et al., 2017; Qin and Shelton, 2017; Xu et al., 2015) and developments in corresponding TRECVID tasks (https://www-nlpir.nist.gov/projects/tv2017/);
- EC9. To find if a method performs well, we use information from quantitative experimental evaluations described in the corresponding publication;
- EC10. Although patents bring innovative ideas, searching for them typically involves strategies and databases different from the ones used for research papers. Thus, we did not consider them in this work, but an extension of the current SR protocol could be designed to deal specifically with patents in future work;
- EC11. Near-duplicate video copy detection has been a relevant CBVIR task and inspired TRECVID evaluations (Wang et al., 2017; Rouhi and Thom, 2014; Tao Shen et al., 2013; Lian et al., 2010). Furthermore, copy detection is useful, for example, to find copies of copyrighted videos, allowing owners to make further actions. Research on this topic faces challenges such as efficiency in video indexing and retrieval, even if the amount of data is large (Boukhari and Serir, 2016; Wang et al., 2016; Zhu et al., 2015; Song et al., 2013). We think that designing a specific SR to extend and update previous reviews in this broad topic is appropriate in future work;

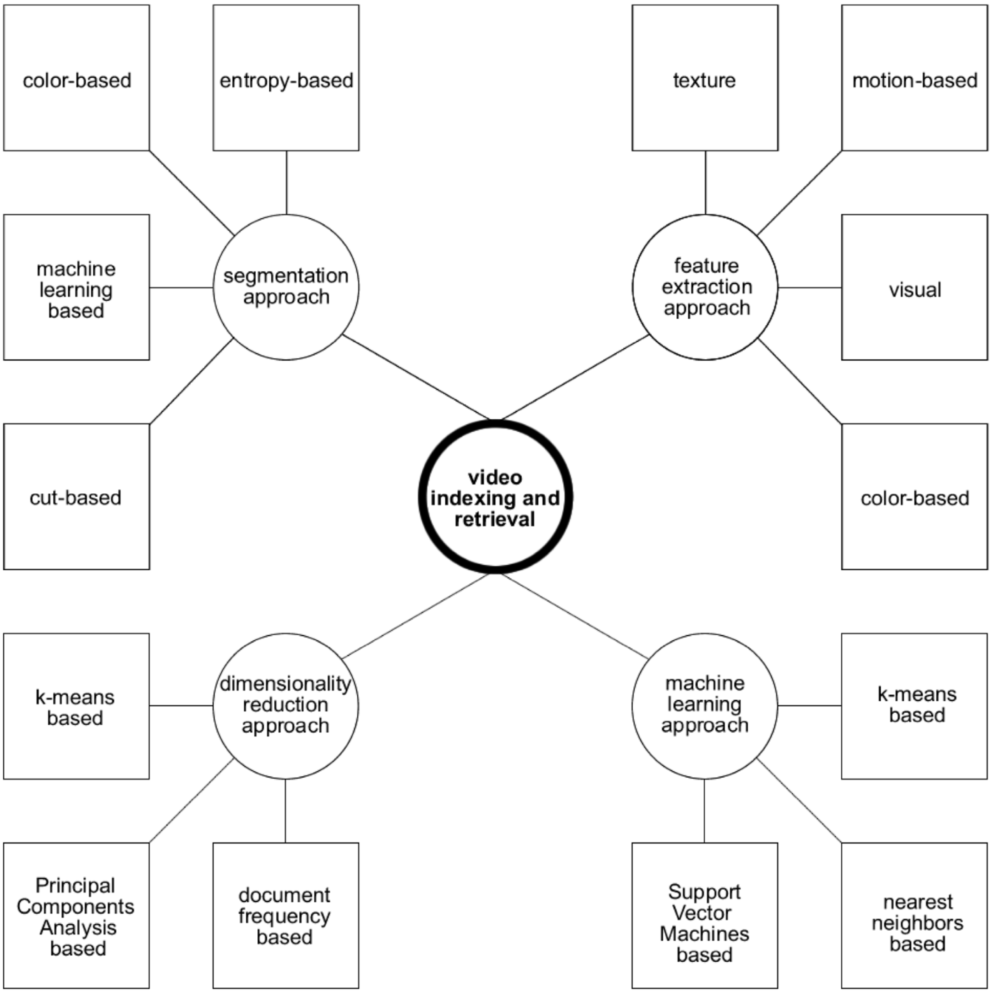

Fig. 3. Usual approaches found in content-based video indexing and retrieval.

- EC12. Academic challenges motivate innovations in an area of study. However, as was the case in the EC10 justification, a specific search strategy would be needed to capture all relevant contributions in these events;
- EC13. The justification is similar to the EC1 one.

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. 3 indicates the most frequent approaches per topic.

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) (Münzer et al., 2017; SenGupta et al., 2015; Priya and Shanmugam, 2013), while the remaining ones consider alternatives related with frames, scenes or other structures (Ngo and Vo, 2014; Cui and Zhu, 2013). 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 (Abdulhussain et al., 2018; Ushapreethi and Lakshmipriya, 2017).

The cut-based or abrupt 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 Pereira et al. (2015). In particular, the authors followed a procedure similar to the one reported in Sadlier et al. (2002) to automatically detect television advertisements in MPEG videos. A frame is considered as a black one for abrupt transition detection if the average luminance intensity based on its discrete cosine transform coefficients was smaller than an adaptive threshold. A silent frame also represents a sudden change, but its identification considers an audio-based threshold.

Another idea, described in Asha and Sreeraj (2013), applies the thresholds t_b and t_s on the histogram difference between frames. In particular, t_b is useful to identify hard or abrupt 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. One can note that this work hybridizes the two categories described in Section 2.1. Also, Asha and Sreeraj (2013) takes into account one of the most common SBD techniques: thresholds (Hu et al., 2011).

Segmentation based on a machine learning algorithm has been another usual choice for SBD. The Boosting method used in Choi et al. (2013) combines weak classifiers, *i.e.*, classification models slightly better than random guessing, to define a stronger learner able to identify sharp boundaries. During the learning, this method assigns a weight for each weak classifier before associating them (Zhi-Hua, 2012). To feed Boosting, frame-by-frame 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.

Table 1
Segmentation approaches found in recent papers.

Approach	Number of occurrences	References
Cut-based	7	Li et al. (2017), Lakshmi Rupa and Gitanjali (2011), Kant (2012), Asha and Sreeraj (2013), Loganathan et al. (2012), Pereira et al. (2015) and Wu et al. (2013)
Machine learning	5	Choi et al. (2013), Quellec et al. (2014a, 2013, 2014b) and Mühling et al. (2016)
Color-based	4	Huang and Chen (2011), Rossetto et al. (2014), Daga (2013) and Liang et al. (2012)
Entropy-based	2	Yarmohammadi et al. (2013), Kamde et al. (2013)

Another ML alternative, based on Nearest Neighbors (NN), is applied in Quellec et al. (2014b). One of the preliminary steps in this alternative involves defining frame subsequences of fixed size from video segments representing action phases in cataract surgeries or idle phases. Then, each subsequence is submitted to ML to find its nearest neighbors in a reference dataset. If most neighbors of a subsequence came from idle periods, it could be in an idle phase as well. Action phases are delimited by two consecutive idle phases. Thus, Quellec et al. (2014b) uses ML to support medical video segmentation into action and idle phases. NN is also useful in Quellec et al. (2014a) to label video subsequences as important ones, for an eye surgery task (segment) to be recognized, by searching their neighbors in an annotated dataset.

Diverse Density (DD) (Maron and Lozano-Pérez, 1998), a multiple instance learning method, is considered in Quellec et al. (2013). The idea is to identify semantically relevant video subsequences in eye surgery tasks (segments). In this scenario, each video subsequence, represented as a signature, is assumed as an instance. Each video, in turn, is regarded as a bag of instances. A video is relevant (positive) if it contains at least one case relevant to a specific surgical task (class). Given a set of labeled videos, diverse density trains a classifier to detect key subsequences. During this process, the classifier measures the intersection of the positive bags minus the union of the negative ones. Estimating an intersection takes into account the number of positive bags with instances close to this point and how far the negative cases are from it. Thus, DD depends on the closeness between instances, as NN also does. It should be emphasized that this algorithm was also useful in Quellec et al. (2013) to split segments into subsequences, as the number of frames in each subsequence was found after evaluating DD in a cross-validation setting.

Besides supervised Boosting and NN approaches, unsupervised algorithms have also been used to support segmentation, as illustrated in Mühling et al. (2016). In particular, the authors use a boundary detection method based on histogram difference and the c-means algorithm to cluster cut candidates identified without thresholds (Ewerth and Freisleben, 2004). One could note that all papers classified into the ML category can be considered as abrupt segmentation representatives.

Color-based segmentation can take into account thresholds. The RGB histogram difference measure reported in Huang and Chen (2011) 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, $i \in 1, \ldots, N$, $j \in 1, \ldots, N$, which are separated by k frames. If the difference reaches an empirically found threshold, then a shot boundary is defined. Although a threshold is also used for cut detection in Rossetto et al. (2014), it is applied as a reference for the difference between fuzzy color histograms based on small frame sub-images. Cut identification in Liang et al. (2012) is grounded on the RGB histogram difference, as is the case in Huang and Chen (2011). Moreover, the authors also detect dissolve transition by skipping some frames and calculating the difference between specific frames within an interval.

An alternative category considers Entropy and related measures. Entropy calculated from gray level co-occurrence matrices for texture analysis of consecutive frames identifies a new shot boundary in Kamde et al. (2013) if the measure is higher than a threshold. Estimating mutual information between frames led to competitive abrupt and gradual transition performance in Yarmohammadi et al. (2013).

Table 2
Most commonly used feature extraction approaches.

Approach	Number of occurrences
Color-based	56
Visual	38
Motion-based	36
Texture	22

4.2. Feature extraction

A typical procedure conducted in content-based video retrieval consists in feature or index extraction (Hu et al., 2011). 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 electronic 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 (Ushapreethi and Lakshmipriya, 2017; Inoue and Shinoda, 2016; Priya and Shanmugam, 2013; Hu et al., 2011; Gonzalez and Woods, 2007).

Besides segmentation, color is useful to describe image content. Huang and Chen (Huang and Chen, 2011) 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 Rossetto et al. (2014), 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 (Li et al., 2009). In this context, a bag of visual words approach can be combined with Scale Invariant Feature Transform (Lowe, 1999) (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”). Afterward, statistics are obtained to identify the relevance of the extracted words (André et al., 2012).

A strategy to build a dictionary of visual words by detecting Space-Time Interest Points in frames is reported in Charrière et al. (2014). 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 Ramezani and Yaghmaee (2018a) and Charrière et al. (2017).

Different from color, motion is a dynamic property of videos associated with the temporal variation of content (Hu et al., 2011). In this context, it is suitable to describe the content of sequences of frames. To extract motion-based features to characterize surgical gestures, Quellec et al. (2013) propose a motion model based on

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

spatiotemporal polynomials. In particular, the model considers polynomials that approximate the optical flow mapping spatiotemporal coordinates to specific displacements.

The approach used in Ghosal and Namboodiri (2014) 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 (Gonzalez and Woods, 2007). 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 André et al. (2011). 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 Quellec et al. (2014a). Specifically, the technique analyzes 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 (Cai et al., 2018; Li et al., 2017; Praveena and Bharathi, 2017; Liu and Motoda, 2007).

We found that k-means based clustering (Hennig et al., 2015) has been the most common approach considered to reduce the video dimensionality. Huang and Chen (2011) 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. Afterward, 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 (Ghosal and Namboodiri, 2014). Finally, this algorithm and variations are often associated with usual visual words techniques for video indexing (Shen et al., 2015; Choi et al., 2013; André et al., 2012). 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 (Jackson, 2003). An application specific for video retrieval is identified in Shao et al. (2014). In particular, the authors combine PCA with the k-means clustering technique during video preprocessing to group local features into 1000 codewords. Afterward, 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. Other examples employing PCA for a diverse set of features are found in Quellec et al. (2014a, 2011). 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 Goméz-Conde and Olivieri (2015). 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 Liu and Sui (2018) 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 Murata et al. (2014) 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 (Mironica et al., 2011) calculates the importance of indexes according to the relevance feedback provided by users' subjective judgment on queries. The selection of feature extractors, in turn, is illustrated in Guo et al. (2012), 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 extractors for further concept-based video retrieval.

Only three CBVIR papers from the 153 selected publications performed feature subset selection, *i.e.*, the search for the best subset of video indexes, which can consist of representatives from different feature extraction approaches. Inter-class distance evaluates various feature subsets according to the filter approach, which disregards machine learning algorithms, in Ji et al. (2011). The authors used a statistical test to find discriminative features for video shots associated with different classes. As a result, a subset of features relevant in a classification setting is yielded. The remaining papers randomly sample a subset of features and instances as part of an ensemble method based on support vector machines (Jones et al., 2012; Jones and Shao, 2011). Thus, they represent the embedded approach, in which a learning algorithm performs feature selection during model training (Liu and Motoda, 2007).

The incipiency of filter feature subset selection in CBVIR is considered a future research direction, as described in Section 6.

One can also note that regularization, a mathematical approach applicable to different types of optimization problems (Bühlmann and van de Geer, 2011), has been applied to support dimensionality reduction in the CBVIR literature, as illustrated in what follows:

- The proposal reported in Jiang et al. (2015) employs a regularizer to support the search for salient and consistent concepts for efficient video indexing. In particular, the proposal aims to adjust the high-level features found by an off-the-shelf concept detection method to produce a relatively compact video description. This description is expected to be consistent with the underlying concept representation. In turn, the regularizer is designed to yield sparse representations that can deal with scenarios in which specific concepts occur in the same video. A regularization term parameter with range [0,1] makes it possible to control the

Table 4
Most commonly used machine learning algorithms.

Algorithm	Number of occurrences
k-means based	24
Nearest neighbors based	24
Support vector machines based	22

sparsity. If it is one, the term becomes lasso. But if it is zero, the term becomes group lasso, which can deal with groups including, for example, co-occurring concepts.

- A regularizer to enforce the sparsity in a projection matrix representation learned from a video is adopted in Wu et al. (2013). In particular, the regularizer promotes the learning of a compact matrix based on a few representative dimensions. The dimensionality reduction achieved is important in the context of video retrieval from mobile platforms with relatively limited memory.
- The unsupervised replicated softmax model described in Zhao et al. (2013) uses a regularizer to support parameter estimation. In particular, the softmax model extracts latent semantic topics from a bag of words representation and, consequently, reduces its dimensionality. The regularizer, in turn, controls the topic sparsity to improve descriptor interpretability and discriminability.

4.4. Machine learning

As mentioned, machine learning has been applied by some video retrieval methods to support video segmentation, indexing and retrieval (Hu et al., 2011). Section 4.1 also reports recent CBVIR papers using ML to support segmentation. 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 (Münzer et al., 2018; Puthenputhussery et al., 2017; Han and Kamber, 2011; Hu et al., 2011).

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 (Chamasemani et al., 2018; Kulkarni et al., 2015; Kumar and Sujatha, 2014). 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) (Wattanarachothai and Patanukhom, 2015; Anh et al., 2012; Amiri et al., 2011; Mironica et al., 2011).

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 Cui and Zhu (2013). In this piece of work, the similarity measure is highlighted as one of the main method parameters. By speeding up the dynamic time warping (Müller, 2007) similarity calculation, the authors achieved a fast and accurate retrieval method. The same basic measure is applied in Ghosal and Namboodiri (2014) 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 (André et al., 2012). Moreover, NN principles can be associated with similarity approaches to compare image feature signatures in Medicine (Schoeffmann et al., 2016).

Support Vector Machines (SVM) and variants have been one of the most usual ML techniques to assist video retrieval. As illustrated in Ramezani and Yaghmaee (2018b), Shao et al. (2014) and Mironica et al. (2011), 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 (Yu et al., 2015), segmentation (Yang and Meinel, 2014), semantic indexing with concepts (André et al., 2011) and video classification or annotation (Hu et al., 2011).

Table 5
Machine learning algorithms used in CBVIR papers.

Algorithm	Number of occurrences
k-means	24
Nearest neighbors based techniques	24
Support vector machines based techniques	22
Neural networks based techniques	13
Hierarchical clustering	5
Latent Dirichlet allocation	5
Other clustering-based techniques	5
Boosting-based techniques	3
Fuzzy clustering techniques	3
Logistic regression	2
Motion vector	2
Rocchio's algorithm	2
Adaptive cover tree	1
Ball ranking machines	1
Conditional random fields	1
Differential geometric trajectory cloud	1
Diverse density	1
Expectation maximization	1
Gaussian process regression	1
Hierarchical cellular tree	1
Nearest feature line	1
Probability based learning	1
Pseudo label mining	1
Query generative model	1
Random forest	1
Replicated softmax model	1
Self-paced learning pipeline	1
Static adaptive cover tree	1

Table 6
Machine learning paradigms used in CBVIR papers.

Paradigm	Number of occurrences
Clustering-based	39
Statistical-based	36
Instance-based	25
Connectionist	13
Undefined	12
Symbolic	1

Table 5 gives the reader direct access to the frequency of each ML algorithm in the 89 papers considered in this section. As was the case in the previous tables, Table 5 counts separately the frequency for each table row. One can note that neural networks based techniques are relatively common in the CBVIR literature. All of them were employed for deep learning – Section 4.4.1. Clustering algorithms alternative to k-means are also frequent and illustrate other unsupervised learning approaches useful for video indexing and retrieval.

Table 6 summarizes Table 5 by grouping the algorithms into five ML paradigms (Han and Kamber, 2011; Rich et al., 2008). Algorithms that could not be associated with a paradigm were classified into the undefined category. Clustering-based algorithms lead the table by grouping k-means and its alternatives. Statistical-based and instance-based paradigms come next and include, respectively, SVM and NN, other two commonly used algorithms.

4.4.1. Deep learning and big data

Deep learning, an emerging family of machine learning algorithms (Chollet and Allaire, 2018; Zheng et al., 2018; Guo et al., 2016; Lecun et al., 2015), has recently been applied in CBVIR. We found 13 papers that use this idea for video indexing (Kletz et al., 2018; Liu

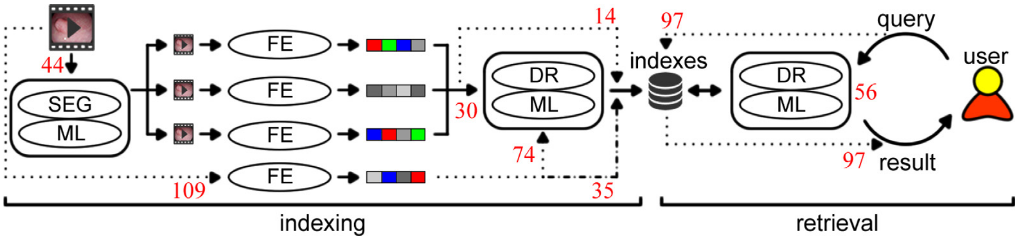

Fig. 4. Content-based video indexing and retrieval workflow used in the publications found. Red numbers indicate the amount of publications in each alternative flow. The use of Segmentation (SEG), Machine Learning (ML), Feature Extraction (FE) and Dimensionality Reduction (DR) concepts is highlighted.

and Sui, 2018; Ma et al., 2018; Petscharnig and Schöffmann, 2018; Markatopoulou et al., 2017; Münzer et al., 2017; Agharwal et al., 2016; Luong et al., 2016; Mühling et al., 2016; Zhang et al., 2016; Jiang et al., 2015; Wattanarachothai and Patanukhom, 2015; Yu et al., 2015). Most of them extract low or high-level features from video frames (Agharwal et al., 2016; Jiang et al., 2015). An example is found in Yu et al. (2015), where both low-level features associated with neural networks and concepts detected from the popular ImageNet model were considered. Other publications learn binary codes for compact video representation (Mühling et al., 2016), such as short hash codes, that can even exploit temporal video properties while providing efficient retrieval (Song et al., 2018; Zhang et al., 2016). Indexing based on ASR and OCR has also benefited from deep learning (Luong et al., 2016; Wattanarachothai and Patanukhom, 2015).

Regarding the deep learning approach, Convolutional Neural Networks (CNN) was used to predict individual concepts, characters (Wattanarachothai and Patanukhom, 2015) and phonemes (Luong et al., 2016). In Mühling et al. (2016), a multi-label CNN predicts multiple concepts simultaneously. CNN popularity can be associated with its ability to process data organized as several arrays, as illustrated by pixel matrices from frames. Recurrent Neural Networks, an alternative approach effective in sequence modeling (Lecun et al., 2015), considers the frame order to describe videos in Zhang et al. (2016). It should be emphasized that most deep learning applications outperformed other methods, such as bag of visual words and video hashing algorithms.

CNN has also been useful in other topics associated with video data, such as action recognition (Wang et al., 2018) and captioning (Gao et al., 2017).

Data mining and CBVIR communities are also increasingly interested in methods dealing with big data (Marx, 2013). An issue investigated in Guo et al. (2015) was the heterogeneity of the multimedia data to be retrieved. The authors propose a method that efficiently processes large media files in low-cost computational nodes working in parallel. A remarkable performance was noted in comparison with alternative approaches, even when the data scale increased. Another scalable method published in Jiang et al. (2015) represents a video by a few concepts consistent with the underlying representation. Searching video from large datasets, which increase in size as time goes by, while providing a personalized experience to the user, was investigated in Feng et al. (2019).

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. Other findings from the systematic review

In this section, systematic review findings regarding the CBVIR workflow, query approaches, retrieval evaluation measures, similarity measures, auxiliary frameworks and video datasets are presented.

5.1. Content-based video indexing and retrieval workflow

Fig. 4 summarizes the workflow used by the selected publications to index and retrieve video based on content. Solid arrows indicate a workflow using the four topics reviewed in this work. Dashed arrows, in turn, indicate alternative flows followed by some publications.

One of the initial steps for video indexing in 44 publications was the segmentation task, implemented with the support of machine learning algorithms (Mühling et al., 2016; Quellec et al., 2014a, 2013; Choi et al., 2013; Quellec et al., 2014b) or other techniques (Section 4.1). In all of them, the segments obtained were submitted for feature extraction (Section 4.2), yielding arrays of video indexes. From the 44 publications, 30 used machine learning and or dimensionality reduction on the extracted indexes (Sections 4.3 and 4.4). In particular, 9 combined both topics, reducing the array size by applying dimensionality reduction techniques based on machine learning (Li et al., 2017; Münzer et al., 2017; Padmakala and AnandhaMala, 2017; Ghosal and Namboodiri, 2014; Choi et al., 2013; Yarmohammadi et al., 2013; Wu et al., 2013; Anh et al., 2012; Huang and Chen, 2011). As indicated in Fig. 4, 35 references did not use segmentation, dimensionality reduction and machine learning to index videos.

Altogether, 56 papers applied machine learning or dimensionality reduction approaches to support video retrieval during query or relevance feedback, as illustrated by Feng et al. (2019), Kletz et al. (2018), Reddy et al. (2018), Valem et al. (2018), Li et al. (2017), Markatopoulou et al. (2017), Mühling et al. (2016), Goméz-Conde and Olivieri (2015), Pereira et al. (2015), Yin et al. (2015), Yu et al. (2015), Charrière et al. (2014), Jones et al. (2014), Quellec et al. (2014a), Rossetto et al. (2014), Shao et al. (2014) and Huurnink et al. (2012). Three out of the 56 papers combined the topics, performing the dimensionality reduction task based on machine learning methods (Reddy et al., 2018; Jones and Shao, 2011; Wang et al., 2011). The remaining publications followed a different path, not using any of these topics for retrieval, as indicated in Fig. 4.

It should be emphasized that 103 papers, nearly 67% of the 153 selected publications, employed dimensionality reduction or machine learning approaches in indexing or retrieval. This finding strengthens the importance of these topics in the context of CBVIR.

5.2. Findings on approaches and measures

As pointed out by 61% of the reviewed papers, Query-by-Example (QbE) has been the most commonly employed approach by which users retrieve videos. In this intuitive approach, a user gives a video as input to obtain one or more similar videos (Schoeffmann et al., 2018; Valem et al., 2018; Goméz-Conde and Olivieri, 2015; Charrière et al., 2014; Cui and Zhu, 2013; Huang and Chen, 2011; Mironica et al., 2011). QbE is followed by image (Beecks et al., 2017; Schoeffmann et al., 2016; Murata et al., 2014), text (Markatopoulou et al., 2017; Marx, 2013; Guo et al., 2012) and sketch (Ghosal and Namboodiri, 2014; Al Kabary and Schuldt, 2014) based categories, which were found in respectively 23%, 22% and 4% of the publications. It should be emphasized that the percentages presented do not sum up to 100%, as it is also possible to find multiple querying approaches in the same paper (Feng et al., 2019; Münzer et al., 2017; Mühling et al., 2016; Guo et al., 2015; Huurnink et al., 2012; Kant, 2012).

Precision and Recall and their variations have been the most popular evaluation measures, as indicated in 61% and 43% of the selected papers. These measures aim to estimate quality, respectively, according to the: (1) ratio of the amount of relevant videos to the total amount of videos obtained after retrieval and (2) ratio of the amount of relevant videos obtained after retrieval to the total amount of relevant videos in a database (Cedillo-Hernandez et al., 2014). They are followed by time-based measures (33%) that quantify, for example, the retrieval or preprocessing time taken to accomplish a specific task (Münzer et al., 2017; Kamde et al., 2016; Charrière et al., 2014; Cui and Zhu, 2013). Accuracy was another popular measure, as illustrated by 20% of the selected references. Besides these findings, one can note that 73% of the reviewed studies apply two or more measures, obtaining a more complete view of the experimental assessments conducted.

As is the case with query approaches and evaluation measures, one can note a variety of similarity measures in the literature. The Euclidean distance is the most frequent choice in the reviewed papers (27%). It is typically applied between a specific query and database videos (Quellec et al., 2013), such that the most similar elements can be retrieved. Other measures found include cosine similarity (Murata et al., 2014), Manhattan distance (Asha and Sreeraj, 2013), histogram-based (Choi et al., 2013) and correlation-based measures (Shen et al., 2015).

5.3. Findings on resources

From the 153 reviewed papers, 54 report the auxiliary frameworks used for video retrieval. Examples of frameworks include the OpenCV computer vision tool, database management systems, the FFmpeg multimedia processing software, the libSVM library for machine learning, the Lucene text search engine and the Tesseract tool for optical character recognition. Most of these frameworks are publicly available, improving the method replicability and saving programming time.

A few papers go beyond and make publicly available their CBVIR systems. Open-source LIvRE (github.com/nospotfer/livre) extends a content-based image retrieval library to support video indexing and retrieval (de Oliveira Barra et al., 2016). Cineast consists in another open-source tool that can be combined with two additional tools (http://vitrivr.org) to achieve indexing and retrieval (Rossetto et al., 2014). VideoVis (https://patentq.njit.edu/oer), a prototype that offers user-friendly clues to search for biomedical educational videos, is proposed in Kamde et al. (2016).

Another relevant resource type consists in video datasets. The datasets associated with the TRECVID meetings have been frequently considered in the relevant literature. As many of them are publicly available (http://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html), it is easier to researchers to compare their proposals with previous results obtained from the same data. The KTH, UCF and Columbia consumer video databases have also been used by researchers (Goméz-Conde and Olivieri, 2015).

Medical datasets regarding the ophthalmology, gynecology and colonoscopy domains are employed in some studies (Schoeffmann et al., 2018; Beecks et al., 2017; Schoeffmann et al., 2016; Charrière et al., 2014; André et al., 2011), illustrating the applicability of retrieval methods for real-world videos. Finally, video datasets built by research authors have been considered in many cases, indicating an alternative when public datasets are not enough to study specific issues.

5.4. Comparison among CBVIR methods

Previous sections highlighted evaluation measures and datasets typically used in the content-based video indexing and retrieval literature. By considering results achieved in the same datasets by the same measures, one can perform an indirect comparison among CBVIR methods. We focused on Mean Average Precision (MAP) and Accuracy results reported in the references, as these were two popular measures addressed by different publications in the same dataset.

The MAP measure is based on other common retrieval evaluation measures, Precision ($Prec$) and Recall (Rec) – Section 5.2. Let rel and $retr$ be, respectively, the set of relevant and retrieved videos returned by a specific query q. The $Prec$ and Rec measures of q are defined by Eqs. (1) and (2), respectively. Let $Prec@i$ be the Precision at cutoff i and ΔRec_i be the change in Recall from videos $i-1$ to i in a ranked list of retrieved videos. The Average Precision of a query q ($AP(q)$) is defined by Eq. (3). MAP, defined by Eq. (4), averages the AP results across all queries in a set of Q queries.

$$Prec = \frac{|rel \cap retr|}{|retr|}. \tag{1}$$

$$Rec = \frac{|rel \cap retr|}{|rel|}. \tag{2}$$

$$AP(q) = \sum_{i=1}^{retr} Prec@i \times \Delta Rec_i. \tag{3}$$

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}. \tag{4}$$

Accuracy in the context of classification represents the number of videos correctly labeled and has been used to support retrieval evaluation (Kohandani Tafresh et al., 2014). Let tp, tn, fp and fn be the number of true positives, true negatives, false positives and false negatives. Then Accuracy is defined by Eq. (5).

$$Acc = \frac{tp+tn}{tp+tn+fp+fn}. \tag{5}$$

Tables 7 and 8 associate the best-reported results with the corresponding datasets.

One can note impressive results in large datasets, such as the MAP value of 0.86 achieved in Yahoo Flickr Creative Common and its 800,000 videos (Jiang et al., 2015). The scalable nature of the method employed in the paper and associated issues explain this finding. Effective CBVIR methods were also observed in datasets from different domains, such as TRECVID 2005 (news) (Wei and Yang, 2013), KTH (human action recognition) (Goméz-Conde and Olivieri, 2015; Jones et al., 2012; Thepade and Yadav, 2015) and UCF (Youtube and

Table 7
Best MAP results reported by some publications per dataset.

	MEDTEST 2013	MEDTEST 2014	Yahoo	TRECVID 2005	TRECVID 2007
Jiang et al. (2015)	0.21	0.19	0.86	–	–
Yu et al. (2015)	0.46	0.42	–	–	–
Agharwal et al. (2016)	–	0.12	–	–	–
Zhang et al. (2016)	–	–	0.60	–	–
Wei and Yang (2013)	–	–	–	0.70	0.51
Memar et al. (2013)	–	–	–	0.17	–
Sang et al. (2015)	–	–	–	–	0.04

Table 8
Best accuracy results reported by some publications per dataset.

	KTH	UCF Youtube	UCF Sports	Hollywood	Colonoscopy
Goméz-Conde and Olivieri (2015)	0.95	1.00	–	–	–
Jones et al. (2014)	0.87	0.55	–	–	–
Ramezani and Yaghmaee (2018a)	0.57	0.46	–	–	–
Jones et al. (2012)	1.00	–	0.90	–	–
Thepade and Yadav (2015)	1.00	–	–	–	–
Jones and Shao (2011)	–	0.60	–	–	–
Jones and Shao (2013)	–	0.55	0.95	0.25	–
Quellec et al. (2014a)	–	–	–	0.75	–
André et al. (2011)	–	–	–	–	0.94
Kohandani Tafresh et al. (2014)	–	–	–	–	0.94

Sports) (Jones and Shao, 2013). A dataset associated with colonoscopy examinations also was associated with relevant Accuracy results (Kohandani Tafresh et al., 2014; André et al., 2011), which is promising for other medical domains.

It should be emphasized that a direct comparison among the reported results is unfeasible, as different preprocessing tasks were adopted by the publications. An example corresponds to the focus on particular video categories chosen in some papers. Tables 7 and 8 are not larger because they did not include papers using different subsets of the same dataset, which directly influenced in the obtained performance. They also discarded publications dealing with different variations of the same evaluation measure. Future work can expand this comparison by implementing some CBVIR methods and applying them in the same datasets, after submitting them to the same preprocessing procedures.

6. 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 (Ji et al., 2011), as indicated in Section 4.3. 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 (Liu and Motoda, 2007), which are publicly available in the Weka framework (www.cs.waikato.ac.nz/ml/weka). Besides traditional linear and nonlinear DR approaches, CBVIR researchers can also take into account recent ideas to preserve the geometric structure of the data inherent to video features (Luo et al., 2018) or deal with streaming data associated with real-time video transmission (Li et al., 2017).

As indicated in Hu et al. (2011), 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 (Tsoumakas et al., 2009), *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 (Mühling et al., 2016), 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 (Halder et al., 2018; Pereira et al., 2018; Zhang and Zhou, 2014).

Another issue involves the study of the values assigned to the parameters of dimensionality reduction techniques and machine learning algorithms. Despite of initial efforts into this direction (Liu and Sui, 2018; Charrière et al., 2017; André et al., 2012), 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 (Chapelle et al., 2002). The use of computationally demanding deep learning approaches also depends on the choice of an appropriate architecture (Zheng et al., 2018).

As noted in Priya and Shanmugam (2013), 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 (Guo et al., 2015; Pranali et al., 2015; Vigneshwari and Juliet, 2015), five references extracting audio-based features and 11 publications using indexes transcribed from speech. Despite of the current limitations on ASR (Spille et al., 2018), such as relatively low performance on under-resourced languages (Besacier et al., 2014), 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 (Cao et al., 2004).

7. 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 (Müller and Unay, 2017; Bhaumik et al., 2016; Priya and Shanmugam, 2013; Hu et al., 2011; Smeaton, 2007; Antani et al., 2002) 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.

CRediT authorship contribution statement

Newton Spolaôr: Methodology, Data curation, Investigation, Writing - original draft, Visualization. **Huei Diana Lee:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Weber Shoity Resende Takaki:** Investigation, Writing - original draft, Writing - review & editing. **Leandro Augusto Ensina:** Writing - review & editing, Resources. **Claudio Saddy Rodrigues Coy:** Conceptualization, Validation. **Feng Chung Wu:** Conceptualization, Validation, Resources.

Acknowledgments

We would like to thank Araucária Foundation for the Support of the Scientific and Technological Development of Paraná, BRAZIL through a Research and Technological Productivity Scholarship for H.D. Lee (grant 028/2019). We also would like to thank PGEEC/UNIOESTE through a postdoctoral scholarship for N. Spolaôr and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 through a MSc. scholarship for L.A. Ensina. These agencies did not have any further involvement in this paper.

References

Abdulhussain, S.H., Ramli, A.R., Saripan, M.I., Mahmmod, B.M., Al-Haddad, S.A.R., Jassim, W.A., 2018. Methods and challenges in shot boundary detection: A review. Entropy 20 (4), 1–42. http://dx.doi.org/10.3390/e20040214.

Agharwal, A., Kovvuri, R., Nevatia, R., Snoek, C.G.M., 2016. Tag-based video retrieval by embedding semantic content in a continuous word space. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1–8. http://dx.doi.org/10.1109/WACV.2016.7477706.

Al Kabary, I., Schuldt, H., 2014. Enhancing sketch-based sport video retrieval by suggesting relevant motion paths. In: International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, United States, pp. 1227–1230. http://dx.doi.org/10.1145/2600428.2609551.

Amiri, A., Abdollahi, N., Jafari, M., Fathy, M., 2011. Hierarchical key-frame based video shot clustering using generalized trace kernel. In: Pichappan, P., Ahmadi, H., Ariwa, E. (Eds.), International Conference on Innovative Computing Technology. In: Communications in Computer and Information Science, vol. 241, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 251–257. http://dx.doi.org/10.1007/978-3-642-27337-7_23.

André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2011. A smart atlas for endomicroscopy using automated video retrieval. Méd. Image Anal. 15 (4), 460–476. http://dx.doi.org/10.1016/j.media.2011.02.003.

André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2012. Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Trans. Méd. Imaging 31 (6), 1276–1288. http://dx.doi.org/10.1109/TMI.2012.2188301.

Anh, T.Q., Bao, P., Khanh, T.T., Thao, B.N.D., Tuan, T.A., Nhut, N.T., 2012. A content based video retrieval analysis system with extensive features by using kullback-leibler. Int. J. Comput. Intell. Syst. 8 (6), 853–858.

Antani, S., Kasturi, R., Jain, R., 2002. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognit. 35 (4), 945–965. http://dx.doi.org/10.1016/S0031-3203(01)00086-3.

Asha, S., Sreeraj, M., 2013. Content based video retrieval using SURF descriptor. In: International Conference on Advances in Computing and Communications. IEEE, Cochin, India, pp. 212–215. http://dx.doi.org/10.1109/ICACC.2013.49.

Ayadi, M.G., Bouslimi, R., Akaichi, J., 2016. A medical image retrieval scheme with relevance feedback through a medical social network. Soc. Netw. Anal. Min. 6 (1), 53. http://dx.doi.org/10.1007/s13278-016-0362-9.

Beecks, C., Kletz, S., Schoeffmann, K., 2017. Large-scale endoscopic image and video linking with gradient-based signatures. In: IEEE Third International Conference on Multimedia Big Data. pp. 17–21. http://dx.doi.org/10.1109/BigMM.2017.44.

Benois-Pineau, J., Precioso, F., Cord, M., 2012. Visual Indexing and Retrieval. Springer Publishing Company, New York, United States.

Besacier, L., Barnard, E., Karpov, A., Schultz, T., 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Commun. 56, 85–100. http://dx.doi.org/10.1016/j.specom.2013.07.008.

Bhaumik, H., Bhattacharyya, S., Nath, M.D., Chakraborty, S.S., 2016. Hybrid soft computing approaches to content based video retrieval: A brief review. Appl. Softw. Comput. 46, 1008–1029. http://dx.doi.org/10.1016/j.asoc.2016.03.022.

Boukhari, A., Serir, A., 2016. Weber binarized statistical image features (WBSIF) based video copy detection. J. Vis. Commun. Image Represent. 34, 50–64. http://dx.doi.org/10.1016/j.jvcir.2015.10.015.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. Springer-Verlag Berlin Heidelberg, Berlin, Germany.

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79. http://dx.doi.org/10.1016/j.neucom.2017.11.077.

Cao, Y., Tavanapong, W., Li, D., Oh, J., de Groen, P.C., Wong, J., 2004. A visual model approach for parsing colonoscopy videos. In: Enser, P., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (Eds.), International Conference on Image and Video Retrieval. In: Lecture Notes in Computer Science, vol. 3115, Springer Berlin Heidelberg, Berlin, Germany, pp. 160–169. http://dx.doi.org/10.1007/978-3-540-27814-6_22.

Cedillo-Hernandez, M., Garcia-Ugalde, F.J., Cedillo-Hernandez, A., Nakano-Miyatake, M., Perez-Meana, H., 2014. Content based video retrival system for mexican culture heritage based on object matching and local-global descriptors. In: International Conference on Mechatronics, Electronics and Automotive Engineering. IEEE, Cuernavaca, Mexico, pp. 38–43. http://dx.doi.org/10.1109/ICMEAE.2014.16.

Chamasemani, F.F., Affendey, L.S., Mustapha, N., Khalid, F., 2018. Surveillance video retrieval using effective matching techniques. In: 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP). pp. 1–5. http://dx.doi.org/10.1109/INFRKM.2018.8464772.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Mach. Learn. 46 (1–3), 131–159.

Charrière, K., Quellec, G., Lamard, M., Coatrieux, G., Cochener, B., Cazuguel, G., 2014. Automated surgical step recognition in normalized cataract surgery videos. In: International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, Chicago, IL, United States, pp. 4647–4650. http://dx.doi.org/10.1109/EMBC.2014.6944660.

Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B., 2017. Real-time analysis of cataract surgery videos using statistical models. Multimedia Tools Appl. 76 (21), 22473–22491. http://dx.doi.org/10.1007/s11042-017-4793-8.

Choi, J., Wang, Z., Lee, S.C., Jeon, W.J., 2013. A spatio-temporal pyramid matching for video retrieval. Comput. Vis. Image Underst. 117 (6), 660–669. http://dx.doi.org/10.1016/j.cviu.2013.02.003.

Chollet, F., Allaire, J.J., 2018. Deep learning with R. Manning, Shelter island, United States.

Cui, H., Zhu, M., 2013. A novel multi-metric scheme using dynamic time warping for similarity video clip search. In: IEEE International Conference on Signal Processing, Communication and Computing. IEEE, KunMing, China, pp. 1–5. http://dx.doi.org/10.1109/ICSPCC.2013.6663926.

Daga, B., 2013. Advances in Computing, Communication, and Control International Conference. Springer Berlin Heidelberg, Berlin, pp. 609–625. http://dx.doi.org/10.1007/978-3-642-36321-4_57, chapter Content Based Video Retrieval Using Color Feature: An Integration Approach.

Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P., 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist. Radiol. Surg. 11 (6), 1081–1089. http://dx.doi.org/10.1007/s11548-016-1371-x.

Ewerth, R., Freisleben, B., 2004. Video cut detection without thresholds. In: Workshop on Signals, Systems and Image Processing. pp. 227–230.

Fan, H., Chang, X., Cheng, D., Yang, Y., Xu, D., Hauptmann, A.G., 2017. Complex event detection by identifying reliable shots from untrimmed videos. In: IEEE International Conference on Computer Vision. pp. 736–744. http://dx.doi.org/10.1109/ICCV.2017.86.

Fan, J., Zhu, X., Xiao, J., 2004. Content-based video indexing and retrieval. In: DiMarco, J. (Ed.), Computer Graphics and Multimedia: Applications, Problems and Solutions. IGI Global, Hershey, PA, United States, pp. 110–144. http://dx.doi.org/10.4018/978-1-59140-196-4.ch007.

Fei, S., Bo, L., Fen, H., Haibo, Z., Jiacheng, C., Yun, R., Lin, G., 2016. A qoe centric distributed caching approach for vehicular video streaming in cellular networks. Wirel. Commun. Mob. Comput. 16 (12), 1612–1624. http://dx.doi.org/10.1002/wcm.2636.

Feng, Y., Zhou, P., Xu, J., Ji, S., Wu, D., 2019. Video big data retrieval over media cloud: A context-aware online learning approach. IEEE Trans. Multimed. 21 (7), 1762–1777. http://dx.doi.org/10.1109/TMM.2018.2885237.

Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T., 2017. Video captioning with attention-based lstm and semantic consistency. IEEE Trans. Multimed. 19 (9), 2045–2055. http://dx.doi.org/10.1109/TMM.2017.2729019.

Gargi, U., Kasturi, R., Strayer, S.H., 2000. Performance characterization of video-shot-change detection methods. IEEE Trans. Circuits Syst. Video Technol. 10 (1), 1–13.

Ghosal, K., Namboodiri, A., 2014. A sketch-based approach to video retrieval using qualitative features. In: Indian Conference on Computer Vision Graphics and Image Processing. ACM, New York, NY, United States, pp. 1–8. http://dx.doi.org/10.1145/2683483.2683537.

Goméz-Conde, I., Olivieri, D.N., 2015. A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system. Expert Syst. Appl. 42 (13), 5472–5490. http://dx.doi.org/10.1016/j.eswa.2015.03.010.

Gonzalez, R.C., Woods, R.E., 2007. Digital Image Processing, third ed. Prentice Hall, Upper Saddle River, NJ, United States.

Gregor, R., Lamprecht, A., Sipiran, I., Schreck, T., Bustos, B., 2015. Empirical evaluation of dissimilarity measures for 3d object retrieval with application to multi-feature retrieval. In: International Workshop on Content-Based Multimedia Indexing. pp. 1–6.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: A review. Neurocomputing 187, 27–48. http://dx.doi.org/10.1016/j.neucom.2015.09.116.

Guo, K., Pan, W., Lu, M., Zhou, X., Ma, J., 2015. An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval. J. Syst. Softw. 102, 207–216. http://dx.doi.org/10.1016/j.jss.2014.09.016.

Guo, X., Zhao, Z., Chen, Y., Cai, A., 2012. An improved system for concept-based video retrieval. In: IEEE International Conference on Network Infrastructure and Digital Content. IEEE, Beijing, China, pp. 391–395. http://dx.doi.org/10.1109/ICNIDC.2012.6418781.

Halder, K., Poddar, L., Kan, M.Y., 2018. Cold start thread recommendation as extreme multi-label classification. In: Companion Proceedings of the Web Conference. pp. 1911–1918. http://dx.doi.org/10.1145/3184558.3191659.

Han, J., Kamber, M., 2011. Data mining: concepts and techniques. Morgan Kaufmann, Burlington, MA, United States.

Hennig, C., Meila, M., Murtagh, F., Rocci, R., 2015. Handbook of cluster analysis, first ed. Chapman and Hall/CRC, Boca Raton, United States.

Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S., 2011. A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. C 41 (6), 797–819. http://dx.doi.org/10.1109/TSMCC.2011.2109710.

Huang, Y.F., Chen, H.W., 2011. Active Media Technology International Conference. Springer Berlin Heidelberg, Berlin, pp. 71–82. http://dx.doi.org/10.1007/978-3-642-23620-4_11, chapter A Multitype Indexing CBVR System Constructed with MPEG-7 Visual Features.

Huang, C.L., Liao, B.Y., 2001. A robust scene-change detection method for video segmentation. IEEE Trans. Circuits Syst. Video Technol. 11 (12), 1281–1288. http://dx.doi.org/10.1109/76.974682.

Huurnink, B., Snoek, C.G.M., de Rijke, M., Smeulders, A.W.M., 2012. Content-based analysis improves audiovisual archive retrieval. IEEE Trans. Multimedia 14 (4), 1166–1178. http://dx.doi.org/10.1109/TMM.2012.2193561.

Inoue, N., Shinoda, K., 2016. Semantic indexing for large-scale video retrieval. ITE Trans. Media Technol. Appl. 4 (3), 209–217. http://dx.doi.org/10.3169/mta.4.209.

Jackson, J.E., 2003. A user's guide to principal components. Wiley-Interscience, New York, United States.

Ji, X., Han, J., Hu, X., Li, K., Deng, F., Fang, J., Guo, L., Liu, T., 2011. Retrieving video shots in semantic brain imaging space using manifold-ranking. In: IEEE International Conference on Image Processing. IEEE, Brussels, Belgium, pp. 3633–3636. http://dx.doi.org/10.1109/ICIP.2011.6116505.

Jiang, L., Yu, S.I., Meng, D., Yang, Y., Mitamura, T., Hauptmann, A.G., 2015. Fast and accurate content-based semantic search in 100m internet videos. In: ACM International Conference on Multimedia. pp. 49–58. http://dx.doi.org/10.1145/2733373.2806237.

Jones, S., Shao, L., 2011. Action retrieval with relevance feedback on youtube videos. In: International Conference on Internet Multimedia Computing and Service. pp. 42–45. http://dx.doi.org/10.1145/2043674.2043687.

Jones, S., Shao, L., 2013. Content-based retrieval of human actions from realistic video databases. Inform. Sci. 236, 56–65. http://dx.doi.org/10.1016/j.ins.2013.02.018.

Jones, S., Shao, L., Du, K., 2014. Active learning for human action retrieval using query pool selection. Neurocomputing 124, 89–96. http://dx.doi.org/10.1016/j.neucom.2013.07.031.

Jones, S., Shao, L., Zhang, J., Liu, Y., 2012. Relevance feedback for real-world human action retrieval. Pattern Recognit. Lett. 33 (4), 446–452. http://dx.doi.org/10.1016/j.patrec.2011.05.001.

Kamde, P.M., Shiravale, S., Algur, S.P., 2013. Entropy supported video indexing for content based video retrieval. Int. J. Comput. Appl. 62 (17), 1–6.

Kamde, P.M., Shiravale, S., Algur, S.P., 2016. A new visual navigation system for exploring biomedical open educational resource (OER) videos. J. Amer. Med. Inf. Assoc. 23 (e1), e34–e41. http://dx.doi.org/10.1093/jamia/ocv123.

Kant, S., 2012. Activity-based exploitation of full motion video (fmv). Proc. SPIE 8386, http://dx.doi.org/10.1117/12.920280, 83860D–83860D–11.

Kitchenham, B.A., Charters, S., 2007. Guidelines for performing systematic literature reviews in software engineering. Evidence-based Software Engineering Technical Report.

Kletz, S., Leibetseder, A., Schoeffmann, K., 2018. Evaluation of visual content descriptors for supporting ad-hoc video search tasks at the video browser showdown. In: Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.S., Gabbouj, M., Elgammal, A. (Eds.), MultiMedia Modeling. Springer International Publishing, Cham, pp. 203–215.

Kohandani Tafresh, M., Linard, N., André, B., Ayache, N., Vercauteren, T., 2014. Semi-automated query construction for content-based endomicroscopy video retrieval. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, Cham, pp. 89–96.

Kulkarni, P., Patil, B., Joglekar, B., 2015. An effective content based video analysis and retrieval using pattern indexing techniques. In: International Conference on Industrial Instrumentation and Control. IEEE, Pune, India, pp. 87–92. http://dx.doi.org/10.1109/IIC.2015.7150717.

Kumar, C.R., Suguna, S., 2016. Visual semantic based 3d video retrieval system using hdfs. Data Min. Knowl. Discov. 10 (8), 3806–3825.

Kumar, C.R., Sujatha, S.N.N., 2014. Star: Semi-supervised-clustering technique with application for retrieval of video. In: International Conference on Intelligent Computing Applications. IEEE, Coimbatore, India, pp. 223–227. http://dx.doi.org/10.1109/ICICA.2014.55.

Lakshmi Rupa, G., Gitanjali, J., 2011. A video mining application for image retrieval. Int. J. Comput. Appl. 20 (3), 46–51. http://dx.doi.org/10.5120/2410-3214.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nat 521, 436–444.

Lelescu, D., Schonfeld, D., 2001. Video skimming and summarization based on principal component analysis. In: Al-Shaer, E.S., Pacifici, G. (Eds.), Management of Multimedia on the Internet. In: Lecture Notes in Computer Science, vol. 2216, Springer Berlin Heidelberg, Berlin, Germany, pp. 128–141.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature selection: A data perspective. ACM Comput. Surv. 50 (6), 94:1–94:45. http://dx.doi.org/10.1145/3136625.

Li, K., Li, S., Oh, S., Fu, Y., 2017. Videography-based unconstrained video analysis. IEEE Trans. Image Process. 26 (5), 2261–2273. http://dx.doi.org/10.1109/TIP.2017.2678800.

Li, L.J., Socher, R., Fei-Fei, L., 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, United States, pp. 2036–2043. http://dx.doi.org/10.1109/CVPR.2009.5206718.

Lian, S., Nikolaidis, N., Sencar, H.T., 2010. Content-based video copy detection – a survey. In: Sencar, H.T., Velastin, S., Nikolaidis, N., Lian, S. (Eds.), Intelligent Multimedia Analysis for Security Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 253–273. http://dx.doi.org/10.1007/978-3-642-11756-5_12.

Liang, B., Xiao, W., Liu, X., 2012. Design of video retrieval system using mpeg-7 descriptors. Procedia Eng. 29, 2578–2582. http://dx.doi.org/10.1016/j.proeng.2012.01.354.

Liu, H., Motoda, H., 2007. Computational methods of feature selection. Chapman and Hall/CRC, Boca Raton, United States.

Liu, A.A., Shao, Z., Wong, Y., Li, J., Su, Y.T., Kankanhalli, M., 2017. LSTM-based multi-label video event detection. Multimedia Tools Appl. http://dx.doi.org/10.1007/s11042-017-5532-x.

Liu, Y., Sui, A., 2018. Research on feature dimensionality reduction in content based public cultural video retrieval. In: IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). pp. 718–722. http://dx.doi.org/10.1109/ICIS.2018.8466379.

Loganathan, D., Jamal, J., Nijanthan, P., Balamurugan, V.K., 2012. Advances in Communication, Network, and Computing International Conference. Springer Berlin Heidelberg, Berlin, pp. 351–357. http://dx.doi.org/10.1007/978-3-642-35615-5_56, chapter Enhanced Video Indexing and Retrieval Based on Face Recognition through Combined Detection and Fast LDA.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. IEEE, Kerkyra, Greece, pp. 1150–1157. http://dx.doi.org/10.1109/ICCV.1999.790410.

Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A.G., Zheng, Q., 2018. Adaptive unsupervised feature selection with structure regularization. IEEE Trans. Neural Netw. Learn. Syst. 29 (4), 944–956. http://dx.doi.org/10.1109/TNNLS.2017.2650978.

Luong, T.H., Pham, N.M., Vu, Q.H., 2016. Vietnamese multimedia agricultural information retrieval system as an info service. In: Murakami, Y., Lin, D. (Eds.), International Workshop on Worldwide Language Service Infrastructure. In: Lecture Notes in Computer Science, vol. 9442, Springer International Publishing, Cham, Switzerland, pp. 147–160. http://dx.doi.org/10.1007/978-3-319-31468-6_11.

Ma, C., Gu, Y., Gong, C., Yang, J., Feng, D., 2018. Unsupervised video hashing via deep neural network. Neural Process. Lett. 47 (3), 877–890. http://dx.doi.org/10.1007/s11063-018-9812-x.

Markatopoulou, F., Galanopoulos, D., Mezaris, V., Patras, I., 2017. Query and keyframe representations for ad-hoc video search. In: ACM on International Conference on Multimedia Retrieval. pp. 407–411. http://dx.doi.org/10.1145/3078971.3079041.

Maron, O., Lozano-Pérez, T., 1998. A framework for multiple-instance learning. In: Conference on Advances in Neural Information Processing Systems. pp. 570–576.

Marx, V., 2013. Biology: The big challenges of big data. Nat 498, 255–260.

Memar, S., Affendey, L.S., Mustapha, N., Doraisamy, S.C., Ektefa, M., 2013. An integrated semantic-based approach in concept based video retrieval. Multimedia Tools Appl. 64 (1), 77–95. http://dx.doi.org/10.1007/s11042-011-0848-4.

Mironica, I., Vertan, C., Ionescu, B., 2011. A relevance feedback approach to video genre retrieval. In: IEEE International Conference on Intelligent Computer Communication and Processing. IEEE, Cluj-Napoca, Romania, pp. 327–330. http://dx.doi.org/10.1109/ICCP.2011.6047890.

Mitrović, D., Zeppelzauer, M., Zaharieva, M., Breiteneder, C., 2011. Retrieval of visual composition in film. In: International Workshop on Image Analysis for Multimedia Interactive Services. TU Delft, Delft, The Netherlands, pp. 1–4.

Mühling, M., Meister, M., Korfhage, N., Wehling, J., Hörth, A., Ewerth, R., Freisleben, B., 2016. Content-based video retrieval in historical collections of the german broadcasting archive. In: Fuhr, N., Kovács, L., Risse, T., Nejdl, W. (Eds.), International Conference on Theory and Practice of Digital Libraries. In: Lecture Notes in Computer Science, vol. 9819, Springer International Publishing, Cham, pp. 67–78. http://dx.doi.org/10.1007/978-3-319-43997-6_6.

Müller, M., 2007. Information Retrieval for Music and Motion. Springer, New York, United States.

Müller, H., Unay, D., 2017. Retrieval from and understanding of large-scale multi-modal medical datasets: A review. IEEE Trans. Multimedia 19 (9), 2093–2104. http://dx.doi.org/10.1109/TMM.2017.2729400.

Münzer, B., Primus, M.J., Hudelist, M., Beecks, C., Hürst, W., Schoeffmann, K., 2017. When content-based video retrieval and human computation unite: Towards effective collaborative video search. In: IEEE International Conference on Multimedia Expo Workshops. pp. 214–219. http://dx.doi.org/10.1109/ICMEW.2017.8026262.

Münzer, B., Schoeffmann, K., Böszörmenyi, L., 2013. Relevance segmentation of laparoscopic videos. In: IEEE International Symposium on Multimedia. pp. 84–91. http://dx.doi.org/10.1109/ISM.2013.22.

Münzer, B., Schoeffmann, K., Böszörmenyi, L., 2018. Content-based processing and analysis of endoscopic images and videos: a survey. Multimedia Tools Appl. 77 (1), 1323–1362. http://dx.doi.org/10.1007/s11042-016-4219-z.

Murata, M., Nagano, H., Mukai, R., Kashino, K., Satoh, S., 2014. Bm25 with exponential IDF for instance search. IEEE Trans. Multimedia 16 (6), 1690–1699. http://dx.doi.org/10.1109/TMM.2014.2323945.

Ngo, T.T., Vo, D., 2014. A novel content based scene retrieval using multi-frame features. In: International Conference on Advanced Technologies for Communications. IEEE, Hanoi, Vietnam, pp. 105–108. http://dx.doi.org/10.1109/ATC.2014.7043365.

de Oliveira Barra, G., Lux, M., i Nieto, X.G., 2016. Large scale content-based video retrieval with livre. In: International Workshop on Content-Based Multimedia Indexing. IEEE, Bucharest, Romania, pp. 1–4. http://dx.doi.org/10.1109/CBMI.2016.7500266.

Padmakala, S., AnandhaMala, G., 2017. Interactive video retrieval using semantic level features and relevant feedback. Int. Arab J. Inf. Technol. 14 (5), 764–773.

Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.C., 2018. Categorizing feature selection methods for multi-label classification. Artif. Intell. Rev. 49 (1), 57–78. http://dx.doi.org/10.1007/s10462-016-9516-4.

Pereira, M.H.R., de Souza, C.L., Pádua, F.L.C., Silva, G.D., de Assis, G.T., Pereira, A.C.M., 2015. SAPTE: A multimedia information system to support the discourse analysis and information retrieval of television programs. Multimedia Tools Appl. 74 (23), 10923–10963. http://dx.doi.org/10.1007/s11042-014-2311-9.

Petković, M., Jonker, W., 2004. Content-Based Video Retrieval - a database perspective, first ed. Springer US, New York, NY, United States.

Petscharnig, S., Schöffmann, K., 2018. Binary convolutional neural network features off-the-shelf for image to video linking in endoscopic multimedia databases. Multimedia Tools Appl. 77 (21), 28817–28842. http://dx.doi.org/10.1007/s11042-018-6016-3.

Pouyanfar, S., Yang, Y., Chen, S.C., Shyu, M.L., Iyengar, S.S., 2018. Multimedia big data analytics: A survey. ACM Comput. Surv. 51 (1), 10:1–10:34. http://dx.doi.org/10.1145/3150226.

Pranali, B., Anil, W., Kokhale, S., 2015. Inhalt based video recuperation system using ocr and asr technologies. In: International Conference on Computational Intelligence and Communication Networks. IEEE, Jabalpur, India, pp. 382–386. http://dx.doi.org/10.1109/CICN.2015.315.

Praveena, M.D.A., Bharathi, B., 2017. A survey paper on big data analytics. In: International Conference on Information Communication and Embedded Systems. pp. 1–9.

Primus, M.J., Schoeffmann, K., Böszörmenyi, L., 2013. Segmentation of recorded endoscopic videos by detecting significant motion changes. In: International Workshop on Content-Based Multimedia Indexing. pp. 223–228. http://dx.doi.org/10.1109/CBMI.2013.6576587.

Priya, R., Shanmugam, T.N., 2013. A comprehensive review of significant researches on content based indexing and retrieval of visual information. Front. Comput. Sci. 7 (5), 782–799. http://dx.doi.org/10.1007/s11704-013-1276-6.

Puthenputhussery, A., Chen, S., Lee, J., Spasovic, L., Liu, C., 2017. Learning and recognition methods for image search and video retrieval. In: Liu, C. (Ed.), Recent Advances in Intelligent Image Search and Video Retrieval. In: Intelligent Systems Reference Library, vol. 121, Springer International Publishing, Cham, pp. 21–43.

Qin, Z., Shelton, C.R., 2017. Event detection in continuous video: An inference in point process approach. IEEE Trans. Image Process. 26 (12), 5680–5691. http://dx.doi.org/10.1109/TIP.2017.2745209.

Quellec, G., Charrière, K., Lamard, M., Droueche, Z., Roux, C., Cochener, B., Cazuguel, G., 2014a. Real-time recognition of surgical tasks in eye surgery videos. Med. Image Anal. 18 (3), 579–590. http://dx.doi.org/10.1016/j.media.2014.02.007.

Quellec, G., Lamard, M., Cazuguel, G., Droueche, Z., Roux, C., Cochener, B., 2011. 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. IEEE, Boston, MA, United States, pp. 4465–4468. http://dx.doi.org/10.1109/IEMBS.2011.6091107.

Quellec, G., Lamard, M., Cochener, B., Cazuguel, G., 2014b. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans. Med. Imaging 33 (12), 2352–2360. http://dx.doi.org/10.1109/TMI.2014.2340473.

Quellec, G., Lamard, M., Droueche, Z., Cochener, B., Roux, C., Cazuguel, G., 2013. 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. Springer Berlin Heidelberg, Berlin, Germany, pp. 10–20. http://dx.doi.org/10.1007/978-3-642-36678-9_2.

Raieli, R., 2013. Multimedia Information Retrieval: Theory and Techniques, first ed. Chandos Publishing, Oxford, United Kingdom.

Ramezani, M., Yaghmaee, F., 2018a. Motion pattern based representation for improving human action retrieval. Multimedia Tools Appl. 77 (19), 26009–26032. http://dx.doi.org/10.1007/s11042-018-5835-6.

Ramezani, M., Yaghmaee, F., 2018b. Retrieving human action by fusing the motion information of interest points. Int. J. Artif. Intell. Tools 27 (3), http://dx.doi.org/10.1142/S0218213018500082.

Ranjith Kumar, C.M., Suguna, S., 2018. A powerful and lightweight 3d video retrieval using 3d images over hadoop mapreduce. In: Hemanth, D.J., Smys, S. (Eds.), Computational Vision and Bio Inspired Computing. Springer International Publishing, Cham, pp. 744–757.

Reddy, V., Varma, P.S., Govardhan, A., 2018. Action model prediction and analysis for CBMR application. In: International Conference on Computing Methodologies and Communication. pp. 1015–1020. http://dx.doi.org/10.1109/ICCMC.2018.8487504.

Rich, E., Knight, K., Nair, S.B., 2008. Artificial intelligence, third ed. Tata McGraw-Hill Education, Delhi, India.

Rossetto, L., Giangreco, I., Schuldt, H., 2014. Cineast: A multi-feature sketch-based video retrieval engine. In: IEEE International Symposium on Multimedia. IEEE, Taichung, Taiwan, pp. 18–23. http://dx.doi.org/10.1109/ISM.2014.38.

Rouhi, A.H., Thom, J.A., 2014. A compressed-domain robust descriptor for near duplicate video copy detection. In: International Conference on Image and Vision Computing New Zealand. pp. 130–135. http://dx.doi.org/10.1145/2683405.2683417.

Sadlier, D.A., Marlow, S., O'Connor, N., Murphy, N., 2002. Automatic tv advertisement detection from mpeg bitstream. Pattern Recognit. 35 (12), 2719–2726. http://dx.doi.org/10.1016/S0031-3203(01)00251-5.

Safadi, B., Sahuguet, M., Huet, B., 2014. When textual and visual information join forces for multimedia retrieval. In: International Conference on Multimedia Retrieval. ACM, New York, NY, USA, p. 265.

Sang, M., Sun, Z., Jia, K., 2015. Semantic similarity based video reranking. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN). pp. 1420–1423. http://dx.doi.org/10.1109/CICN.2015.274.

Schoeffmann, K., Beecks, C., Lux, M., Uysal, M.S., Seidl, T., 2016. Content-based retrieval in videos from laparoscopic surgery. Proc. SPIE 9786, 9786–1–9786–10. http://dx.doi.org/10.1117/12.2216864.

Schoeffmann, K., Del Fabro, M., Szkaliczki, T., Böszörmenyi, L., Keckstein, J., 2015. Keyframe extraction in endoscopic video. Multimedia Tools Appl. 74 (24), 11187–11206. http://dx.doi.org/10.1007/s11042-014-2224-7.

Schoeffmann, K., Husslein, H., Kletz, S., Petscharnig, S., Muenzer, B., Beecks, C., 2018. Video retrieval in laparoscopic video recordings with dynamic content descriptors. Multimedia Tools Appl. 77 (13), 16813–16832. http://dx.doi.org/10.1007/s11042-017-5252-2.

SenGupta, A., Thounaojam, D.M., Singh, K.M., Roy, S., 2015. Video shot boundary detection: A review. In: IEEE International Conference on Electrical, Computer and Communication Technologies. IEEE, Coimbatore, India, pp. 1–6. http://dx.doi.org/10.1109/ICECCT.2015.7226084.

Shao, L., Jones, S., Li, X., 2014. Efficient search and localization of human actions in video databases. IEEE Trans. Circuits Syst. Video Technol. 24 (3), 504–512. http://dx.doi.org/10.1109/TCSVT.2013.2276700.

Sharma, R., Mummareddy, S., Hershey, J., Jung, N., 2013. Method and system for analyzing shopping behavior in a store by associating RFID data with video-based behavior and segmentation data. Patent. US 8380558.

Shen, X., Zhang, L., Wang, Z., Feng, D., 2015. Spatial-temporal correlation for trajectory based action video retrieval. In: IEEE International Workshop on Multimedia Signal Processing. IEEE, Xiamen, China, pp. 1–6. http://dx.doi.org/10.1109/MMSP.2015.7340811.

Smeaton, A.F., 2007. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. Inf. Syst. 32 (4), 545–559. http://dx.doi.org/10.1016/j.is.2006.09.001.

Song, J., Yang, Y., Huang, Z., Shen, H.T., Luo, J., 2013. Effective multiple feature hashing for large-scale near-duplicate video retrieval. IEEE Trans. Multimedia 15 (8), 1997–2008. http://dx.doi.org/10.1109/TMM.2013.2271746.

Song, J., Zhang, H., Li, X., Gao, L., Wang, M., Hong, R., 2018. Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Trans. Image Process. 27 (7), 3210–3221. http://dx.doi.org/10.1109/TIP.2018.2814344.

Spille, C., Kollmeier, B., Meyer, B.T., 2018. Comparing human and automatic speech recognition in simple and complex acoustic scenes. Comput. Speech Lang. 52, 123–140. http://dx.doi.org/10.1016/j.csl.2018.04.003.

Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D., 2016. A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing 180, 3–15. http://dx.doi.org/10.1016/j.neucom.2015.07.118.

Tao Shen, H., Liu, J., Huang, Z., Ngo, C.W., Wang, W., 2013. Near-duplicate video retrieval: Current research and future trends. ACM Comput. Surv. 45, 44:1–44:23.

Thepade, S.D., Yadav, N., 2015. Novel efficient content based video retrieval method using cosine-haar hybrid wavelet transform with energy compaction. In: International Conference on Computing Communication Control and Automation. pp. 615–619. http://dx.doi.org/10.1109/ICCUBEA.2015.126.

Tsoumakas, G., Katakis, I., Vlahavas, I., 2009. Mining multi-label data. Data Min. Knowl. Discov. Handb. 1–19.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2017. Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36 (1), 86–97. http://dx.doi.org/10.1109/TMI.2016.2593957.

Ushapreethi, P., Lakshmipriya, G.G., 2017. Survey on video big data: Analysis methods and applications. Int. J. Appl. Eng. Res. 12, 2221–2231.

Valem, L.P., aes Pedronette, D.C.G., Almeida, J., 2018. Unsupervised similarity learning through cartesian product of ranking references. Pattern Recognit. Lett. 114, 41–52. http://dx.doi.org/10.1016/j.patrec.2017.10.013, Data Representation and Representation Learning for Video Analysis.

Varytimidis, C., Rapantzikos, K., Loukas, C., Kollias, S., 2016. Surgical video retrieval using deep neural networks. In: Proceedings of Workshop and Challenges on Modeling and Monitoring of Computer Assisted Interventions. pp. 1–11.

Vigneshwari, G., Juliet, A.N.M., 2015. Optimized searching of video based on speech and video text content. In: International Conference on Soft-Computing and Networks Security. IEEE, Coimbatore, India, pp. 1–4. http://dx.doi.org/10.1109/ICSNS.2015.7292369.

Wang, L., Bao, Y., Li, H., Fan, X., Luo, Z., 2017. Compact CNN based video representation for efficient video copy detection. In: Amsaleg, L., Gumundsson, G.ó., Gurrin, C., Jónsson, B.ó., Satoh, S. (Eds.), MultiMedia Modeling. Springer International Publishing, Cham, pp. 576–587.

Wang, R.B., Chen, H., Yao, J.L., Guo, Y.T., 2016. Video copy detection based on temporal contextual hashing. In: IEEE Second International Conference on Multimedia Big Data. pp. 223–228. http://dx.doi.org/10.1109/BigMM.2016.12.

Wang, X., Gao, L., Wang, P., Sun, X., Liu, X., 2018. Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Trans. Multimed. 20 (3), 634–644. http://dx.doi.org/10.1109/TMM.2017.2749159.

Wang, L., Song, D., Elyan, E., 2011. Video retrieval based on words-of-interest selection. In: European Conference on Advances in Information Retrieval. pp. 687–690.

Wattanarachothai, W., Patanukhom, K., 2015. Key frame extraction for text based video retrieval using maximally stable extremal regions. In: International Conference on Industrial Networks and Intelligent Systems. European Alliance for Innovation, Begijnhoflaan, Belgium, pp. 29–37. http://dx.doi.org/10.4108/icst.iniscom.2015.258410.

Wei, X.Y., Yang, Z.Q., 2013. Coaching the exploration and exploitation in active learning for interactive video retrieval. IEEE Trans. Image Process. 22 (3), 955–968. http://dx.doi.org/10.1109/TIP.2012.2222902.

Wu, G.L., Kuo, Y.H., Chiu, T.H., Hsu, W.H., Xie, L., 2013. Scalable mobile video retrieval with sparse projection learning and pseudo label mining. IEEE Multimedia 20 (3), 47–57. http://dx.doi.org/10.1109/MMUL.2013.13.

Xu, Z., Yang, Y., Hauptmann, A.G., 2015. A discriminative CNN video representation for event detection. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1798–1807.

Yang, H., Meinel, C., 2014. Content based lecture video retrieval using speech and video text information. IEEE Trans. Learn. Technol. 7 (2), 142–154.

Yarmohammadi, H., Rahmati, M., Khadivi, S., 2013. Content based video retrieval using information theory. In: Iranian Conference on Machine Vision and Image Processing. IEEE, Zanjan, Iran, pp. 214–218. http://dx.doi.org/10.1109/IranianMVIP.2013.6779981.

Yin, Y., Seo, B., Zimmermann, R., 2015. Content vs. context: Visual and geographic information use in video landmark retrieval. ACM Trans. Multimedia Comput. Commun. Appl. 11 (3), 39:1–39:21. http://dx.doi.org/10.1145/2700287.

Younessian, E., Rajan, D., 2012. 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. In: Lecture Notes in Computer Science, vol. 7131, Springer Berlin Heidelberg, Berlin, Germany, pp. 186–195.

Yu, S.I., Jiang, L., Xu, Z., Yang, Y., Hauptmann, A.G., 2015. Content-based video search over 1 million videos with 1 core in 1 second. In: ACM on International Conference on Multimedia Retrieval. ACM, New York, NY, USA, pp. 419–426. http://dx.doi.org/10.1145/2671188.2749398.

Yu Cao, Tavanapong, W., Kihwan Kim, Wong, J., JungHwan Oh, de Groen, P.C., 2004. A framework for parsing colonoscopy videos for semantic units. In: IEEE International Conference on Multimedia and Expo. pp. 1879–1882. http://dx.doi.org/10.1109/ICME.2004.1394625.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., Zhang, B., 2007. A formal study of shot boundary detection. IEEE Trans. Circuits Syst. Video Technol. 17 (2), 168–186. http://dx.doi.org/10.1109/TCSVT.2006.888023.

Zha, Z.J., Wang, M., Zheng, Y.T., Yang, Y., Hong, R., Chua, T.S., 2012. Interactive video indexing with statistical active learning. IEEE Trans. Multimedia 14 (1), 17–27. http://dx.doi.org/10.1109/TMM.2011.2174782.

Zhang, H., Wang, M., Hong, R., Chua, T.-S., 2016. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing. In: ACM on Multimedia Conference. pp. 781–790. http://dx.doi.org/10.1145/2964284.2964308.

Zhang, M.L., Zhou, Z.H., 2014. A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. 26 (8), 1819–1837. http://dx.doi.org/10.1109/TKDE.2013.39.

Zhao, F., Huang, Y., Wang, L., Tan, T., 2013. Discovering compact topical descriptors for web video retrieval. In: IEEE International Conference on Image Processing. pp. 2679–2683. http://dx.doi.org/10.1109/ICIP.2013.6738552.

Zheng, L., Yang, Y., Tian, Q., 2018. SIFT meets CNN: A decade survey of instance retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 40 (5), 1224–1244. http://dx.doi.org/10.1109/TPAMI.2017.2709749.

Zhi-Hua, Z., 2012. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, Boca Raton, United States.

Zhu, N., He, W., Hua, Y., Chen, Y., 2015. Marlin: Taming the big streaming data in large scale video similarity search. In: IEEE International Conference on Big Data (Big Data). pp. 1755–1764. http://dx.doi.org/10.1109/BigData.2015.7363947.