International Journal of Theoretical & Applied Computational Intelligence
https://ijtaci.com

Review article

Content-Based Video Retrieval: A Comprehensive Review of Methods, Frameworks and Trends

Ajaz Khan Baig[1], Uzair Ishtiaq[2*], Zubair Ishtiaque[3] and Uzair Iqbal[4]

[1]Department of Computer Sciences, IBADAT International University, Islamabad, Pakistan

[2]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[3]Department of Analytical, Biopharmaceutical and Medical Sciences, Atlantic Technological University, H91 T8NW Galway, Ireland

[4]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

*Corresponding Author: Uzair Ishtiaq. Email: uzair@um.edu.my

https://orcid.org/0000-0002-6015-3821

Received: 11/9/2025; Revised: 7/12/2025; Accepted: 02/01/2026; Published: 10/01/2026.

https://doi.org/10.65278/IJTACI.2026.2

Abstract: Multimedia retrieval and delivery play a critical role in numerous application domains, addressing the challenge of identifying, filtering, and managing exponentially growing volumes of multimedia data. Retrieval is the process by which a user or system identifies needed content, prompting a retrieval system to return relevant data. Dynamic videos possess distinct characteristics that differ from static images, containing content structured across shot, scene, and program levels. Effectively retrieving such content requires an automatic system capable of operating at each of these levels. Content-Based Video Retrieval (CBVR) is the process of efficiently searching for and retrieving video content while ensuring high relevance of the results. This review provides a tutorial and a comprehensive overview of the techniques used for content-based video retrieval. It focuses on video retrieval methodologies alongside an overview of video indexing. A generic CBVR framework is discussed in detail. The core CBVR section includes a literature review and discussion to explore prevailing trends in retrieval systems. The TRECVID benchmark for evaluating different retrieval systems is analyzed. Furthermore, key techniques such as similarity measures and relevance feedback for video retrieval are examined. Finally, future research directions are identified and discussed.

Keywords: Content Based Video Retrieval; Video Indexing; Video

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is cited

Mining; Key Frame Extraction; Deep Learning

1. Introduction

Multimedia content encompasses images, text descriptions, and identifiable objects within images or videos. The field of multimedia retrieval involves two primary tasks: the indexing and retrieval of static images, and the more complex challenge of handling video content, which is structured hierarchically into frames, shots, scenes, clips, and full programs. Compared to static images, the dynamic nature of videos introduces significant complexity in indexing and retrieval. This complexity arises from distinct characteristics such as vast volumes of raw data, information richness that surpasses a single image, and a lack of inherent structure [1]. Video retrieval systems are designed to locate videos within existing archives or from sources like web-based platforms and social media. The process typically involves first indexing the video content, after which a retrieval system can respond to user queries with relevant matches. Historically, video databases have been small enough for manual indexing via keyword annotation (metadata). However, the exponential growth of multimedia data has rendered manual methods infeasible, creating a pressing need for automated indexing and retrieval systems. Content-Based Video Retrieval (CBVR) addresses this need by providing automated methods to store, index, describe, organize, and efficiently retrieve video and multimedia content. CBVR systems can scale to handle both small collections and large archives. Advanced algorithms analyze video content to perform indexing at multiple levels: shot, scene, and program (or clip). Corresponding retrieval systems, developed for both closed archives and open web content, process user queries to return the most relevant results at the appropriate granularity. Recent advances in CBVR demonstrate the effectiveness of integrating multiple modalities, such as visual features, audio speech, subtitles, and object recognition, often leveraging distributed and deep-learning methods. These approaches are crucial for managing large-scale video collections while improving retrieval accuracy and speed [2].

The applications of visual content-based indexing and retrieval are broad, spanning large institutional archives and web-based video repositories. Web videos often include auxiliary textual information (e.g., titles, descriptions) that can aid retrieval. Key application domains include rapid video library browsing, e-commerce analysis, remote instruction, digital museums, news event analysis, intelligent web video management, and video surveillance [3]. To understand CBVR, it is essential to define its basic units. An image is an electronic representation of a visual subject. A frame is a single, electronically coded still image within a video sequence. A shot is an uninterrupted sequence of frames captured from a single camera operation. Multiple shots are combined to form a scene, which depicts action in a continuous time and location. In retrieval, an image serves as the fundamental unit; images are indexed based on metadata, depicted objects, background, and other features, enabling the search and retrieval of individual static images from a large pool. Video retrieval extends this concept by searching based on information associated with the entire video or its components, such as keywords, tags, or descriptions linked to its constituent frames. The process involves searching indexed videos within large databases. The content within video frames, including objects, backgrounds, color, and texture, can be categorized into two feature types: low-level and high-level [4]. Low-level features consist of directly computable attributes like color histograms, texture patterns, and associated text. High-level features involve semantic concepts such as recognized objects, motion patterns, specific instruments, or events. In CBVR systems, objects within video frames are extracted, and their content is analyzed to enable retrieval (see Figure 1).

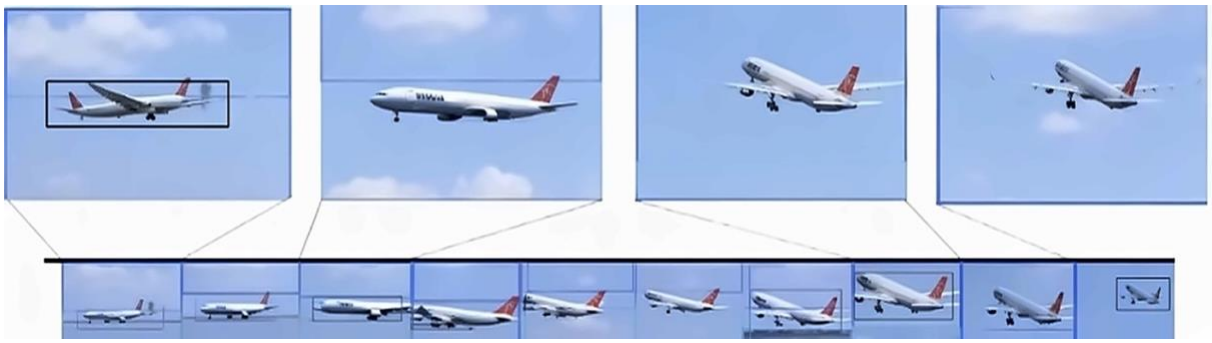

Figure 1: Object behavior in frames

Video content is analyzed to extract and associate meaningful information, enabling automatic indexing using specialized tools. This indexing operates at multiple hierarchical levels: frame, shot, scene, and program.

The field of content-based video indexing and retrieval boasts a broad range of applications, attracting significant research interest worldwide. Since 2001, the National Institute of Standards and Technology (NIST) has hosted the annual Text Retrieval Conference (TREC) Video Retrieval Evaluation (TRECVID) [5]. This initiative provides a common platform to assess and accelerate progress in video analysis and retrieval. TRECVID supplies a large-scale, standardized video dataset, allowing developers to benchmark the effectiveness and efficiency of their CBVR algorithms across various domains. Video standards, such as MPEG and TV-Anytime, aim to ensure compatibility and define interfaces for video concepts. These standards facilitate the development of efficient and accurate video retrieval algorithms by promoting a common framework. Videos contain rich multimodal information, which can be extracted from different channels, including metadata, audio data, and visual information. In metadata, textual tags are associated with a video (e.g., title, summary, date, actors, format, size, copyright). In audio data the characteristics are extracted from the auditory channel. However, in visual information, the features are extracted from the visual channel, including objects, backgrounds, colors, and textures. The volume of video data in the digital universe is growing exponentially, as is the associated descriptive information. This expansion creates a pressing need for more sophisticated indexing and retrieval tools to manage increasingly complex video collections. Consequently, continued research and development in this field is essential. This review focuses specifically on visual content, such as objects, colors, textures, and backgrounds for video retrieval. To establish a foundational understanding, a significant portion is dedicated to video indexing, as it is the critical prerequisite for any retrieval system. A generic framework for video indexing is presented and discussed in subsequent sections. The rationale is to build a clear conceptual bridge from indexing (the essential first step) to retrieval (the subsequent step). The primary contribution of this study is to provide a comprehensive exploration of video retrieval, encompassing its fundamental terminologies, core concepts, prevalent methods, key applications, and future research challenges.

2. Literature review of CBVR techniques

The effectiveness of Content-Based Video Retrieval (CBVR) hinges critically on the underlying similarity measures used to match user queries with video content. A typical retrieval process begins with a user submitting a query via a system interface. The system parses this query and compares it against indexed videos in a database or web repository. Successful matches return the most relevant videos, often accompanied by related results, while unsuccessful searches yield a null response. The nature of the query

itself, whether a single input or a combination, varies by system design, with significant research dedicated to making query interfaces more intuitive and effective.

This section reviews the core technical pillars of CBVR systems as presented in the literature: (1) query formulation and interfaces, (2) similarity measurement techniques, (3) relevance feedback mechanisms, and (4) video summarization for efficient browsing.

2.1 Query Formulation and Interfaces

Query formulation is the critical first step in engaging a retrieval system. Research has evolved from simple single-modality queries towards more sophisticated, multi-modal, and adaptive interfaces to bridge the "semantic gap" between user intent and low-level video features. Early and foundational work focused on specific query types. For instance, Hu et al. [6] introduced a trajectory-based interface where users could sketch a shape or trajectory, optionally augmented with text, to retrieve related images or videos. Similarly, Sivic and Zisserman [7] demonstrated query-by-object, where a user-provided image of an object retrieves all videos containing it. Moving towards natural language, Aytar et al. [8] parsed English queries to match semantic concepts within videos, addressing the significant challenge of understanding user intent. A key limitation of single-query interfaces is their often narrow or immature search scope, which can fail to capture the user's precise need. Consequently, research advanced towards combination-based and adaptive query systems. Kennedy et al. [9] developed a framework that automatically discovers useful query combinations from training data, classifying them for a uni-modal search. Yan and Hauptmann [10] further advanced this by fusing multiple search tools and implementing a query-class-dependent retrieval system, where query classes are learned from training data to guide the search process adaptively. This line of work includes methods to distinguish between semantic classes (e.g., person vs. non-person queries).

2.2 Similarity measurement techniques

Once a query is formulated, the system must compute its similarity to candidate videos. Similarity measures typically fall into four categories: feature-based, text-based, ontology-based, and combination-based (hybrid) matching. Feature-based matching compares low-level or object-level visual descriptors. Browne and Smeaton [11] utilized a query-by-example approach based on low-level features. Other methods have incorporated stationary key-frame features, motion features, and object-based features for matching [12]. Snoek et al. [13] normalized the semantic concepts derived from videos and the query text provided by the user, computing similarities between these conceptual representations.

2.3 Relevance feedback and deep learning–based refinement in video retrieval

To refine imperfect initial results, relevance feedback has become an essential component of interactive CBVR systems. It involves the user providing feedback on initial results, which the system uses to adjust its model. Feedback can be explicit (direct user input), implicit (inferred from user behavior), or pseudo-relevance (automatically generated). In explicit feedback approaches, Le [14] required users to manually select positive video samples, using these to enhance subsequent results. Another common method allows users to rank samples as relevant or non-relevant, with the system adjusting feature weights accordingly. Some authors used explicit feedback to re-weight features and optimize the query point for refinement [15]. Implicit feedback infers intent indirectly. Joachims [16] refined queries using click-through data ranked by the search algorithm. Ghosh et al. [17] personalized retrieval results based on such implicit click-through feedback, while others have presented models to adapt video retrieval based on simulated implicit feedback. Pseudo-relevance feedback automates the process. Khan et al. [18] demonstrated the effectiveness of fused deep features for human action recognition, while Sivic et al. [19] and Luan et al. [20] provided foundational methods for person spotting and interactive video retrieval. More recently, Nallappan and Velswamy [21]

utilized deep neural networks for content-based video retrieval and anomaly detection. Ragedhaksha et al. [22] revealed the use of deep learning approaches for object detection and anomaly recognition in surveillance videos, supporting enhanced aviation safety and security. Authors in [23] combined vision transformers with a genetic algorithm to refine retrieval results without user interaction. Luan et al. [20] also used an active learning approach where the system iteratively selects the most informative videos for user judgment until the set of unlabeled relevant videos is minimized, effectively learning a classification boundary.

2.4 Video summarization techniques for efficient browsing

For efficient navigation of large video collections, video summarization creates condensed overviews by removing redundant content and presenting key information in a readable or visual format [24, 25]. This allows users to quickly assess relevance from a large pool. Techniques vary in their approach. Guironnet et al. [26] utilized camera motion features to detect distinctive key frames for summarization. Xie and Wu [27] generated abstracts for news videos by selecting key frames based on static features and using clustering to differentiate scenes. Xiao et al. [28] developed an algorithm to remove similar segments and select the most informative parts of a video shot for "skimming." Gong [29] presented a bimodal technique, separating and summarizing audio and video content independently using representative audio tracks and visually distinct frames, then reintegrating them via a bipartite graph for a cohesive abstract. The TRECVID 2007 rushes summarization task [30] prompted algorithms that used a top-down approach: first filtering out uninformative shots, then minimizing redundancy within shots using key-frame features.

3. A generic framework for content-based video retrieval

Following the taxonomy of Hu et al. [31], visual CBVR includes several layers as shown in Figure 2. This section describes the generic CBVR framework. It has two major components: an indexing component, which consists of structure analysis, feature extraction, data mining, classification, and annotation, and a retrieval component containing query, browsing, and results.

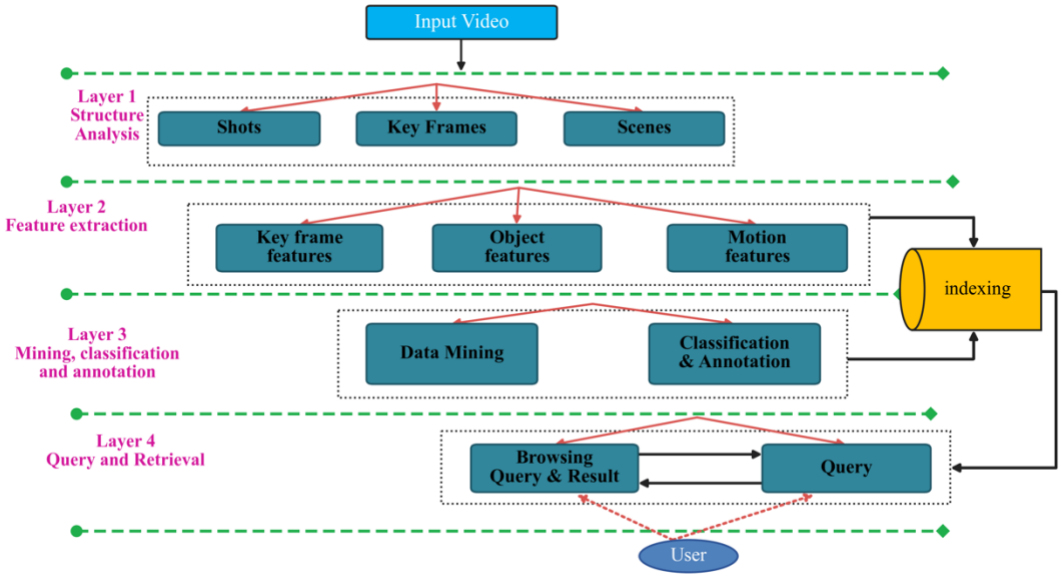

Figure 2: Content Based Video Retrieval (CBVR) model

A canonical framework for Content-Based Video Retrieval (CBVR) systems can be decomposed into several interconnected processing layers. This structure, illustrated in Figure 2, moves from raw video data to a searchable index and, finally, to an interactive retrieval interface. The core stages are described below.

- **Video content analysis and structuring:** The initial layer involves parsing the raw video stream into meaningful structural units. This is achieved through video segmentation, which partitions the video temporally at the shot and scene levels using algorithms for shot-boundary detection. Consecutive frames with consistent visual content are grouped into shots, and semantically related shots are clustered into scenes. From these units, representative key frames are extracted to serve as concise visual summaries. This structural analysis often leverages low-level features such as color histograms, edge information, video change ratios, and motion vectors to identify boundaries and transitions.
- **Feature extraction:** Following segmentation, discriminative features are extracted from the identified units (shots, scenes, and key frames). These features form the descriptive basis for all subsequent operations. They are broadly categorized into:
  - **Static frame features:** Attributes of individual key frames, such as color, texture, and shape.
  - **Motion features:** Dynamics within shots, including optical flow, motion trajectories, and camera motion.
  - **Object and semantic features:** Detected entities (e.g., faces, vehicles, buildings) and their spatial relationships.
  - **Local features:** Distinctive interest points and descriptors (e.g., SIFT, SURF) that are robust to transformations.
- **Indexing and video data mining:** The extracted features are used to construct a searchable index that is a critical step enabling efficient retrieval from large databases. This process involves video data mining, where patterns and knowledge are derived from the feature space. The indexing itself can be:
  - **High-Dimensional Indexing:** Organizing the numerical feature vectors using specialized data structures (e.g., k-d trees, hash-based methods) for fast similarity search.
  - **Semantic Indexing:** Associating higher-level concepts or labels (e.g., "outdoor," "meeting," "sports") with video units based on the mined knowledge, bridging the gap between low-level features and user semantics.
- **Video classification and annotation**: To further organize the database and support semantic search, videos or their segments are often categorized. Video classification assigns predefined labels (e.g., genre, topic) to entire videos or major segments based on their extracted features. Video annotation is a finer-grained process, attaching descriptive metadata (tags, keywords, or bounding boxes) to specific objects, events, or temporal regions within a video. While related, annotation provides a more detailed and localized description than broad-scale classification.
- **Query and retrieval:** This is the user-facing layer of the framework. A query, which can be an example video, a sketch, keywords, or a combination, is submitted through an interface. The system parses the query, converts it into a comparable feature representation, and employs similarity measures to search the indexed database. Upon finding matches, the most relevant videos are retrieved and presented. Results are frequently delivered as a video summary (e.g., a set of key frames or a skim) to facilitate efficient user browsing. To refine imperfect results, the framework often incorporates relevance feedback mechanisms, where user input on initial results is used to optimize subsequent search iterations through query refinement or feature re-weighting.

  The following sections provide the details of the components of CBVR.

3.1 Video content analysis and structure

Video content is inherently hierarchical, organized into three fundamental levels: frames, shots, and scenes, as illustrated in Figure 3. Understanding this structure is essential for effective CBVR. The frame

is the most basic unit, a single static image sampled at a specific time point. A shot is defined as an uninterrupted sequence of frames captured from a single camera operation. It represents a continuous action in time and space and serves as the primary unit for retrieval in many CBVR systems. A scene is a higher-level semantic unit composed of multiple consecutive shots that share a common location, narrative, or thematic context. This hierarchical relationship can be summarized as follows: a video is composed of scenes; scenes are composed of shots; and shots are composed of frames. In the processing pipeline, this structure dictates the segmentation strategy. The initial step involves shot boundary detection to parse the video stream into shots. From each detected shot, representative key frames are extracted to summarize its visual content. Subsequently, scene segmentation techniques group visually or semantically similar shots into scenes. This multi-level segmentation enables indexing and retrieval at different granularities, whether a user searches for a specific object within a frame, an action within a shot, or an event spanning a scene [32].

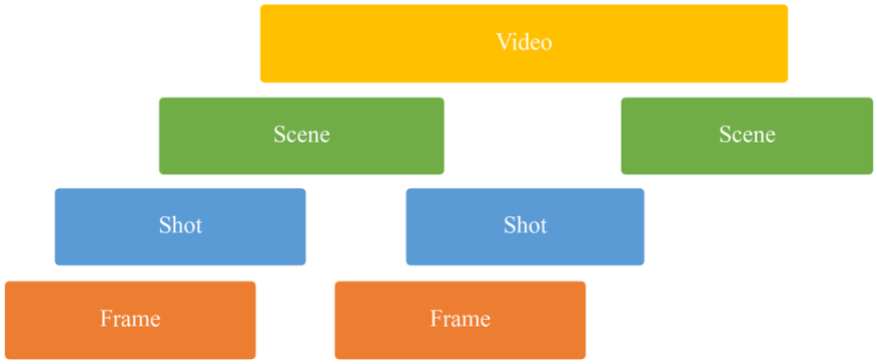

Figure 3: Hierarchical structure of video, scenes, shots, and frames

Following the hierarchical segmentation of a video into its constituent shots and frames, the system is prepared to process user queries. A user initiates a search by submitting a query, which can take various forms such as descriptive text, an example object image, a sketch, or a combination of these modalities. The retrieval system decodes this input, converting it into a compatible feature representation. This representation is then matched against the indexed database of video content, searching at the shot, key frame, or semantic concept level based on the query type. When a sufficient match is identified, the system retrieves and presents the most relevant video(s) to the user. Furthermore, to aid exploration, the results are often accompanied by a set of related videos, facilitating a more comprehensive browsing experience.

3.1.1 Shot boundary detection

A video shot is defined as an uninterrupted sequence of frames captured from a single, continuous camera operation, bounded by a start (record) and stop point. This temporal segment forms the basic structural unit for organizing and retrieving video content. The transitions marking shot boundaries are categorized into two primary types [33, 34]. Abrupt Transitions (Cuts) refer to the instantaneous changes from one shot to the next and Gradual Transitions are the smoother changes that occur over several frames, including fades (in/out), dissolves, and wipes. Given the strong visual coherence among frames within a shot, it is widely considered the fundamental element for subsequent video retrieval operations. Detecting these boundaries is a multi-step process that involves extracting features from frames, measuring their similarities, and applying a decision mechanism. The standard pipeline consists of the following stages:

- **Feature extraction**: Discriminative features are extracted from each frame to facilitate comparison. These features are broadly categorized into two levels. Low-Level Features are computable directly from pixel data, such as color histograms, edge change ratios, and texture descriptors. In contrast, High-Level (Semantic) Features involve more complex interpretation, including detected objects, object motion trajectories, background scene classification, and identified instruments or entities.
- **Similarity measurement:** The similarity between frames or groups of frames is quantified to identify points of significant visual change. The two predominant measurement schemes are pairwise similarity and window-based similarity. The former computes a distance metric (e.g., Euclidean distance, Cosine dissimilarity) directly between consecutive frames using their extracted feature vectors. A sharp drop in similarity suggests a potential cut. The latter measures aggregate similarity within a sliding temporal window of frames. This method is more effective for detecting gradual transitions, as it compares the content of a small segment against adjacent segments.
- **Boundary decision:** The final step classifies frame pairs or regions as either a boundary or non-boundary based on the computed similarities. Contemporary approaches fall into two main categories, namely Threshold-Based Methods and Machine Learning-Based Methods. In threshold-based methods, a shot change is declared when the inter-frame similarity falls below a predefined threshold. In contrast, Machine Learning-Based Methods frame SBD as a pattern classification or clustering problem. Further, in Supervised Learning, classifiers like Support Vector Machines (SVM), Adaboost and neural networks are trained on labeled datasets using frame features (e.g., color, motion) to distinguish boundary frames from non-boundary frames [35]. Unsupervised Learning identifies boundaries by grouping frames without pre-existing labels. Common techniques include Similarity Clustering (for pairwise frame similarities; clusters of low similarity values indicate boundaries) and Content Clustering (which treats each shot as a cluster of visually similar frames). Algorithms like K-means and Fuzzy C-means group frames in feature space, with transitions between clusters denoting shot boundaries.

3.1.2 Key frame extraction

Key frame extraction is the foundational process of selecting frames that best summarize shot content, aiming for distinctiveness and minimal redundancy. Historically, features such as color histograms, edge descriptors, and MPEG-7 motion activity descriptors have been used. The methodologies can be broadly categorized as follows [36]. However, a mere listing of categories obscures the critical engineering trade-offs involved in choosing an algorithm for a specific CBVR application. The following list describes the categories, while Table 1 provides a synthesized critical comparison of the key frame extraction paradigms.

- **Sequential comparison between frames:** Subsequent frames are sequentially compared to a previously extracted key frame. When a frame is found to be sufficiently different (based on a feature distance threshold), it is selected as the next key frame. This process continues to the end of the video.
- **Global contrast among frames:** These algorithms use the global differences between all frames in a shot, distributing key frames by minimizing a predefined objective function. Common objectives include: (1) Maximum Coverage, which maximizes the representativeness of key frames for all frames; (2) Minimum Correlation, which extracts key frames to minimize the sum of correlations between them; and (3) Minimum Reconstruction Error, which selects frames that best reconstruct all other frames (particularly useful for animated or highly structured videos).
- **Reference frame:** A single reference frame is first generated, often using aggregated features like an average histogram. Frames in the shot are then compared to this reference, and those exceeding a

difference threshold are selected as key frames. These algorithms are notably straightforward to implement.

- **Clustering:** Frames are clustered in a feature space (using global or local features), and the frames closest to the cluster centroids are chosen as key frames. Past implementations have used Fuzzy C-Means and agglomerative hierarchical clustering (e.g., complete-link method).
- **Curve simplification:** The sequence of frames is treated as a curve in a multi-dimensional feature space. Key frame extraction is formulated as a curve simplification problem, where the goal is to select the minimal set of points (frames) that preserve the essential shape (content evolution) of the curve.
- **Objects/events:** These algorithms are semantically driven, extracting key frames based on the detection of specific objects or the occurrence of predefined events (e.g., a goal in soccer, a scene change). Features like object shape and color are used to link key frames to high-level semantic information.

Table 1: Critical comparison of key frame extraction paradigms

Paradigm	Strengths	Limitations & Why They Fail
Sequential comparison	Extremely simple, fast and having online processing capability.	• Highly sensitive to threshold choice. Fails with gradual transitions or highly variable content. • Error propagation: A poor key frame selection skews all subsequent comparisons. • Redundant for static shots; sparse for dynamic ones.
Global contrast	Optimal for the chosen objective while providing a non-redundant, representative summary.	• Offline only: Requires the entire shot/video. • Computationally expensive (often $O(n^2)$ or NP-hard approximations). • Fails if the objective function does not match the real-world notion of "importance."
Reference frame	Simple to implement and understand and faster than global methods.	• Crude representation: A single reference poorly models complex shot content. Fails catastrophically in shots with bimodal or multi-modal content (e.g., a pan across two distinct scenes).
Clustering	Robust to noise and minor variations, controllable output via the number of clusters (k).	• Requires predefining 'k' (number of key frames), which is often unknown. • Feature sensitivity: Quality depends entirely on the chosen feature space. Fails if features don't correlate with perceptual importance. • Offline and iterative.
Curve simplification	Intuitively models the temporal flow of a shot, good at capturing progressive changes (zooms, pans).	• Complex to implement and parameterize. • May select frames from smooth transitions rather than stable, informative segments.
Objects/events	High semantic relevance where key frames are tied directly to meaningful content, excellent for domain-specific retrieval.	• Not generic. Requires pre-defined event/object detectors. • Fails completely outside its trained domain. • Heavily dependent on the accuracy of the underlying detector.

3.1.3 Scene segmentation

Scene segmentation represents the process of grouping contiguous shots that share a higher-level semantic meaning into a coherent narrative unit, forming a scene. This grouping relies on integrating various information streams, including visual, auditory, conceptual, and semantic cues derived from the video content [37]. The methodologies for scene segmentation can be categorized according to how shots are represented and processed. From a representational perspective, key approaches include the key frame-based method, where each shot is represented by extracted key frame features and scenes are formed by clustering shots with visually similar key frames, though this static representation often fails to capture

temporal dynamics. The audio-visual integrated approach detects scene boundaries by identifying concurrent changes in both the visual and audio channels, treating scenes as multimodal segments. Alternatively, the background-based approach operates on the supposition that shots sharing a similar background belong to the same scene, matching frames based on background content and comparing key frames across potential scenes.

From the perspective of processing methodology, scene segmentation techniques are further divided into four categories. The integration-based, or bottom-up, approach merges similar shots iteratively to form scenes. Conversely, the splitting-based, or top-down, approach partitions the entire video into distinct scenes, ensuring all shots belonging to a semantic unit are contained together, where the misplacement of a single shot can render the division ineffective. The statistical-based approach constructs probabilistic or statistical models of shot sequences to infer scene boundaries. Finally, the shot boundary classification approach adopts a hierarchical strategy: it first extracts standard shot boundaries and then uses advanced features from these boundaries to classify them as either scene boundaries or non-scene boundaries, effectively building scenes from a refined analysis of transitions.

3.2 Feature extraction

Visual video indexing and retrieval fundamentally depend on the features extracted from the structural analysis of video content, making feature extraction the core technical component of these systems. These features are broadly categorized into two semantic levels. High-level, or semantic, features involve an interpretative understanding of the video's content, such as recognized objects, specific instruments, patterns of motion, and scene background. In contrast, low-level, or static, features consist of directly computable perceptual attributes derived from pixel data, including color histograms, texture patterns, and shape descriptors [31]. The effectiveness of any retrieval system hinges on the discriminative power of these extracted features to bridge the gap between raw visual data and user-centric semantic queries. In this study, only visual features are included that focus on visual indexing and retrieval. Table 2 provides a comparative analysis of video feature extraction paradigms. These features are as follows:

3.2.1 Static features of key frames

Building upon the foundational role of features, the specific techniques for indexing and retrieval often operate on key frames, which serve as concise visual representatives of video content. Consequently, established image retrieval methodologies are employed to extract discriminative features from these key frames for subsequent indexing. These features are typically categorized into three primary types aligned with human visual perception: color-based, texture-based, and shape-based approaches [31]. The color-based approach quantifies chromatic information using descriptors such as color histograms, color moments, or mixture of Gaussian models, operating within color spaces like RGB or HSV. The texture-based approach characterizes the surface patterns and granularity of objects or regions, independent of color and intensity. Common descriptors for this include co-occurrence matrices and wavelet-transformed texture features. Finally, the shape-based approach focuses on the geometric form of objects within the key frame, often initiated by detecting object contours or edges. A histogram of edge orientations or distributions is then commonly used to describe and encode the shape characteristics for retrieval. These feature categories provide a robust and interpretable basis for matching visual content in video retrieval systems.

3.2.2 Object features

These extracted features, including the color, texture, and spatial dimensions of image regions corresponding to objects, form the basis for searching videos that contain visually similar objects. Object-level features specifically target identifiable entities within a frame, such as a person, car, tree, or airplane. However, a significant limitation in utilizing object features lies in the inherent difficulty of both accurately identifying the objects themselves and, more crucially, interpreting their spatial and semantic relationships with other objects in the scene. This process is not only computationally complex but also time-consuming, presenting a major challenge for real-time or large-scale video retrieval systems.

3.2.3 Motion features

Motion features provide a critical dimension that distinguishes dynamic video from static imagery, capturing temporal variations that often correspond more closely to semantic concepts and user intent. These features primarily describe movement within the visual field, categorized as either background motion (typically induced by camera operations like pans, tilts, and zooms) or foreground motion (the movement of objects within the scene). Methodologies for characterizing motion can be grouped into three principal types. Statistic-based features model the distribution of motion vectors across frames, quantifying global or local motion patterns through statistical measures. Trajectory-based features focus on modeling the path of specific moving objects, constructing a trajectory curve by tracking the same object across successive frames. Finally, object relationship-based features aim to describe the evolving spatial and interaction relationships between objects across time. While these relational features offer rich semantic potential, they present a significant technical challenge: tracking and consistently interpreting the complex relationships among objects across different key frames is exceptionally difficult, extending beyond the simpler task of analyzing relationships within a single, static frame.

Table 2: Comparative analysis of video feature extraction paradigms

Feature Category	Strengths	Limitations and Accuracy Implications	Computational Cost
Static Key-Frame features	• Intuitive & Aligned with Human Perception. • Simple and fast to compute (especially color histograms). • Robust to small deformations and noise (texture).	• Large Semantic Gap: Low-level features poorly correlate with high-level concepts. Low accuracy for semantic queries. • Sensitive to viewing conditions (lighting, scale, rotation). • Shape features require accurate segmentation, which is itself a hard problem.	Low to Medium
Object-Level features	• Higher Semantic Relevance. Directly represents entities users query for (objects). • Enables relational and attribute-based search (e.g., "red car next to a building").	• Heavily dependent on object detector accuracy. Fails completely if objects are not detected. • Computationally expensive at runtime. • Difficulty in modeling object relationships across frames (as noted in the text).	High (cost of detection / segmentation models)
Motion features	• Defining Characteristic of Video. Captures temporal dynamics absent in images. • Closer to action/event semantics (e.g., "running," "converging"). • Trajectories are compact and descriptive.	• Complex to extract reliably. Tracking is error-prone (occlusion, drift). • Camera motion can dominate and obscure object motion. • Relational features are highly complex and fragile across shots. Accuracy is variable and context-dependent.	Medium to Very High (trajectory tracking and relational modeling are costly)

3.3 Video data mining

Video mining utilizes the features detected during the structural analysis to uncover higher-order patterns and semantics essential for retrieval. Its primary tasks include identifying patterns in video concepts, modeling object behavior, determining scene uniqueness, detecting event patterns and their associations, and extracting other semantic relationships. The specific data mining techniques employed are inherently dependent on the application context and the underlying video indexing scheme.

Several core strategies are central to video data mining. Object mining involves clustering different instances of the same object that appear across various parts of a video or different scenes. This task is challenging due to variations in an object's appearance. A common approach uses spatial neighborhood techniques to cluster frames from specific domains, subsequently mining frequently appearing objects from these clusters. Special pattern detection applies to actions and events for which prior models exist, such as human actions, sporting events, or traffic incidents. However, detecting patterns for an entire video is difficult due to the potential multiplicity of events, and the technique is hard to generalize across all video domains. In contrast, pattern discovery employs algorithms, often semi-supervised or supervised, to automatically detect previously unknown patterns. These discovered patterns can then model new applications and help explore novel data structures, with applications in unusual event detection and associating clusters or text with patterns for retrieval.

Further strategies include video association mining, which finds relationships between concurrent events, such as the co-occurrence of specific objects, to detect inherent and frequent associative patterns. Tendency mining analyzes and detects trends within a particular event by tracking its current progression, mining these trends based on object characteristics. Finally, preference mining focuses on extracting user preferences from various video genres like news, movies, or sports.

The subsequent processes of video classification and annotation are heavily reliant on this foundational video structural analysis and extracted features, particularly key frames. Classification aims to group videos into recognizable categories, such as news, entertainment, or sports, to facilitate organization and search. This is performed based on extracted key frames alongside higher-level semantic and conceptual information derived from the video content.

3.4 Video classification

This process involves extracting meaningful information from videos using their underlying features to classify the content into predefined categories or classes. Employing data mining techniques for this classification is highly efficient for video retrieval and constitutes a critical component for navigating large databases. The content used for classification encompasses both semantic information and video production effects. Semantic content includes broad categories such as video genre, specific events, and recognized objects, where genre represents a wider and coarser level of detection.

Edit effect classification focuses on production elements derived from camera motions and the composition of shots and scenes. While these effects are not inherent semantic parts of the video narrative, they provide valuable context for understanding content and can aid in conceptual classification. This information proves particularly useful for retrieval in scenarios where other semantic indexes are incomplete or unavailable.

Video genre classification categorizes videos into types such as news, movies, or talk shows. This task is approached through several methodologies. A statistical approach models various genres by first

analyzing low-level properties like color, camera motion, and texture, then deriving more abstract attributes such as pacing and specific camera movements, which are finally mapped to genre labels. Alternatively, a rule-based approach applies heuristic rules that connect domain knowledge to low-level features. Machine learning offers a data-driven method, where classifiers are trained on labeled samples using low-level features to automatically categorize video genres.

Event and object classification targets human-recognizable occurrences that signify the video's narrative significance. Classifying these events, based on visible occurrences, is instrumental for content-based categorization and is closely linked to event detection in video data mining. The object serves as the fundamental unit for event classification, with the process often beginning by classifying individual objects, facial recognition being one of the most common applications, to build an understanding of the larger event context.

3.5 Video annotation

Video annotation is the process of assigning predefined semantic labels, such as "person," "chair," or "tree", to specific shots or segments within a video. While video classification typically categorizes an entire video into a broad genre, annotation operates at a finer granularity, applying labels to scenes, events, or individual shots. The fundamental methodology for both annotation and classification involves extracting low-level features from the visual content and then employing trained classifiers to map these features to specific concepts or labeled groups. A single video segment can be annotated with multiple, overlapping concepts.

Annotation strategies are generally categorized into three types. The first is isolated concept-based annotation, where the process relies on training individual concept detectors. Each concept from a predefined visual lexicon is treated independently, with separate binary classifiers detecting specific semantic elements without considering the relationships between them. The second strategy is context-based annotation, which aims to improve detection accuracy by leveraging contextual relationships. This approach either refines the output of individual binary classifiers or infers higher-level concepts by fusing already detected lower-level semantics using contextual fusion strategies; video ontologies are often employed here to model and detect these conceptual relationships. The third strategy is integration-based annotation, a more holistic approach that models both individual concepts and their inherent correlations simultaneously. In this method, learning and categorization occur in parallel by utilizing the entire set of training samples, allowing the model to capture the complex interdependencies within the video's semantic structure.

3.6 Queries and retrieval

Video retrieval is executed once the video content has been indexed, utilizing these indices to respond to user queries. The process begins when a user submits a query through an interface; the system employs similarity measures to compare the query against the indexed database and retrieves the most relevant candidate videos. To further refine imperfect results, relevance feedback mechanisms are often integrated to optimize subsequent search iterations. Two prominent interface designs exemplify this process: the InfoMedia interface supports filtering and sorting based on visual concepts, allowing users to enter descriptors related to objects, colors, textures, or topics to narrow down a large pool to a more manageable set of candidate videos. In contrast, the MediaMill interface employs a multi-modal strategy, combining query-by-example with textual keywords and maintaining a history of previous searches to inform the current one.

The query itself serves as the fundamental input for retrieving videos from a large database, and retrieval is the process of searching for the desired content using this input. Various query types have been developed to accommodate different user intents and information availability, all serving the same ultimate purpose. Query-by-example allows a user to provide a sample image or video clip; the system extracts low-level static features from this example to find visually similar content. Query-by-sketch enables users to draw a simple sketch, which the system then matches against previously extracted key frames. Query-by-object involves submitting an image of a specific object (e.g., a car or chair), prompting the system to retrieve all video segments containing that object. Query-by-keywords leverages textual metadata, transcripts, or assigned visual semantics, allowing users to search with simple descriptive terms. Query-by-natural-language permits users to express their needs in everyday language, which the system must then parse into actionable semantic concepts, which is a common yet challenging approach used in general web search. Finally, combination-based or multi-modal querying integrates two or more of the above types (e.g., text plus a sketch) to overcome the limitations of any single method, providing a more flexible and powerful interface for retrieving content from complex multimedia collections.

3.6.1 Similarity measure

Measuring video similarity is a core function that plays a significant role in the effectiveness of any Content-Based Video Retrieval (CBVR) system. The methods for determining similarity are broadly categorized into feature matching, text matching, ontology matching, and combination-based matching, with the optimal choice being highly dependent on the specific application context. The most direct form is feature-based similarity matching, which typically calculates the average distance between corresponding low-level or high-level features extracted from video frames or shots. For instance, an example-based query utilizes low-level visual features to find relevant videos. This matching can be performed at various granularities, such as the scene, shot, or program level, and may incorporate motion features, object features, or key frame descriptors based on the nature of the user's query. Object features, event characteristics, or static visual attributes can all serve as the basis for comparison.

In contrast, text matching offers a simpler paradigm, where a user's textual query is matched against video descriptions, metadata, or derived semantic concepts. The text and concepts are often normalized to ensure optimal matching, a method exemplified by general-purpose search engines like Google, where a novice user can enter a phrase and receive a set of best-matched videos to choose from. To leverage the strengths of multiple modalities, combination-based approaches integrate different matching strategies. These methods pull from conceptual semantics by learning from training collections, utilizing frameworks such as query-class-dependent mixture models. This fusion makes combination-based queries particularly powerful for multimodal searches, where a single input might combine text, an example, and a sketch to retrieve more accurate and comprehensive results.

3.6.2 Relevance feedback

Relevance feedback is a critical technique for refining video search results by incorporating user input or automated analysis to rerank retrieved videos, thereby optimizing subsequent searches. This refinement process typically involves query point optimization, feature weight adjustment, and the integration of supplemental information, effectively bridging the semantic gap between low-level visual features and high-level user relevance. The technique can also reflect a user's confidence in initial results, as illustrated in Figure 4. Its utility extends to various applications; for example, in a torrent download system, user feedback after downloading a video can directly influence its future ranking and prominence for other users.

Relevance feedback mechanisms are generally classified into three distinct types based on the source of the feedback signal.

Explicit relevance feedback requires users to directly and consciously identify which videos in the initial result set are relevant, often by selecting or rating items. While this method can yield highly accurate refinements due to clear user intent, it demands considerable time, effort, and active engagement from the user. Implicit relevance feedback, in contrast, infers user preferences indirectly by analyzing behavioral data, such as click-through patterns from search engine logs, where clicking on a video in a ranked list is interpreted as a positive relevance signal. Pseudo-relevance feedback automates the process entirely by having the system assume that the top-ranked items from an initial search are positive examples and lower-ranked items are negative examples. These assumed labels are then used to automatically refine the query or model without any user interaction, operating on the principle that items closer to the original query in the feature space are likely relevant, while those farther away are not.

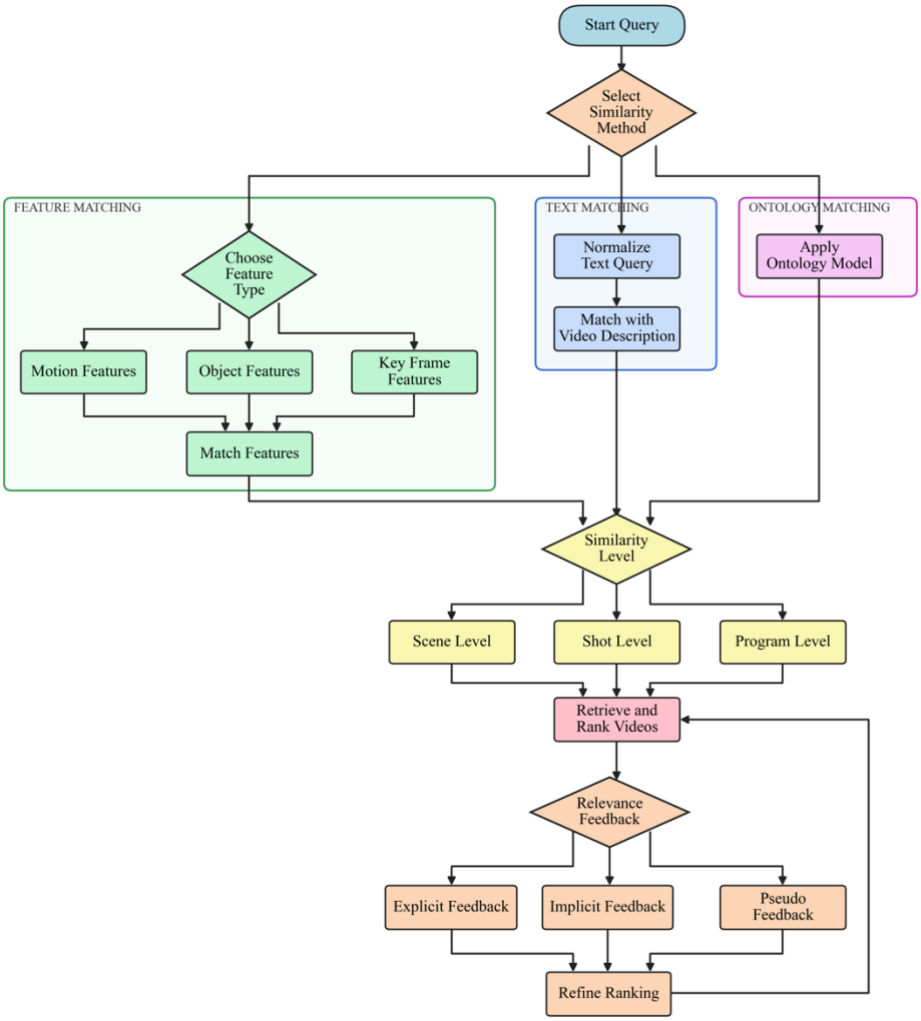

Figure 4: Similarity matching and relevance feedback workflow

3.6.3 Video summarization and browsing

Video summarization is the process of removing redundant components from a video to create a condensed abstract or summary representation. This technique is particularly valuable for managing

extensive video archives, as a concise summary can effectively represent the entirety of a longer video, allowing users to efficiently browse and identify desired content. For instance, a user considering whether to download a full movie may first view its trailer, a quintessential form of video summarization, to make an informed decision. In this way, summaries serve as a vital browsing interface for large video collections. The methodologies for generating these summaries generally fall into two fundamental categories, each addressing the challenge of distilling essential content in different ways.

- **Static video abstracts:** This first category involves generating a summary composed of a curated collection of key frames extracted from the source video. Representative frames are selected from each significant scene and organized sequentially, often accompanied by descriptive text, to form a coherent abstract. Common implementations of this approach include the video table of contents, the storyboard, and the graphical video summary. Historically, numerous algorithms have been developed to automate this process of selecting and arranging key frames. The primary advantage of this static summary is its immediacy; a compact visual representation can be displayed almost instantaneously, eliminating synchronization issues during search and navigation. Furthermore, it allows for non-linear browsing, as users can freely examine the key frames in any order to quickly assess the video's content.
- **Dynamic video skim:** Video skimming involves generating a concise summary by extracting and combining key segments from a source video, analogous to a movie trailer. The resulting skimmed video is a significantly shorter representation of the original content, preserving its essential narrative or informational core. These skims are valuable for efficient video browsing, editing, and content review, offering a time-saving, engaging, and informative alternative for users. Audio can be incorporated to enhance the comprehensibility of skimmed videos. Video skimming methodologies can be broadly categorized as follows:
- **Redundancy removal:** This strategy eliminates redundant segments and retains the most informative ones to form a video skim. Shot evaluation algorithms are typically employed to identify and extract the most significant segments by discarding repetitive content.
- **Object or event detection:** This approach leverages semantic cues, such as specific objects, actions, or events, for skimming. Video segments are selected based on the presence of these predefined semantics. The detected objects or events are then ranked according to their importance to create the final skim.
- **Multimodal integration:** For structured content like news programs and documentaries, the accompanying audio (e.g., speech) provides a valuable resource for summarization. When captions or transcripts are available, a text summary can be generated and integrated with the visual summary to produce a more informative video skim.

4. Deep learning revolution: From CNNs to video transformers and multimodal retrieval

While the foundational methodologies described in previous sections have established the core paradigms of CBVR, the field has undergone a radical transformation since the mid-2010s, driven by deep learning. This shift has moved the focus from hand-crafted feature engineering to learning hierarchical, task-optimized representations directly from data. The evolution has progressed from Convolutional Neural Networks (CNNs) for spatial feature extraction to sophisticated architectures capable of modeling long-range temporal dependencies and understanding multimodal queries (see Table 3). This section reviews these modern approaches, which now define the state-of-the-art.

Table 3: Key modern studies in deep learning for video retrieval

Study (Year)	Architecture	Key Insight for CBVR

[38] ViViT: A Video Vision Transformer (2021)	Adapts the Vision Transformer (ViT) to video via factorized (spatial + temporal) or joint spatio-temporal self-attention.	Demonstrates a scalable blueprint for applying pure transformer architectures to video, enabling long-range spatio-temporal modeling.
[39] TimeSformer (2021)	Introduces "divided space-time attention" — applying spatial and temporal attention separately — for computational efficiency.	Makes video transformers more practical by significantly reducing the quadratic complexity of full spatio-temporal attention.
[40] Video Swin Transformer (2022)	Applies a hierarchical, shifted-window approach (from Swin Transformer) to video, building multi-scale representations.	Achieves state-of-the-art performance by efficiently modeling video at multiple resolutions via local window attention.
[41] Spatiotemporal Contrastive Learning (2021)	Contrastive: Learns by contrasting different views of a video.	Establishes self-supervised learning (SSL) as the dominant pre-training paradigm.
[42] VideoMAE (2022)	VideoMAE: Uses a high-ratio masked autoencoding pretext task.	Enables learning from vast unlabeled video data.
[43] Frozen in Time (2021)	Dual-encoder architecture (video + text transformers) trained with contrastive loss on large-scale video-text datasets (WebVid-2M).	Sets the standard for scalable text-to-video retrieval, allowing efficient search via nearest neighbor lookup in a shared embedding space.
[44] CLIP4Clip (2022)	Empirical study adapting the CLIP image-text model to video via temporal pooling strategies (mean, sequential, etc.).	Shows the powerful transferability of web-scale image-text foundation models to video tasks with minimal architectural change.

4.1 Ascendancy of video transformers

The limitations of CNNs in capturing long-range dependencies and the sequential processing constraints of RNNs have been addressed by the advent of transformer architectures adapted for video. Building on the success of the Vision Transformer (ViT) in images, ViViT: A Video Vision Transformer [38] pioneers the extension to video via factorized spatial and temporal attention mechanisms, providing a scalable framework for joint spatio-temporal modeling. Efficiency quickly became a critical concern, leading to innovations like the TimeSformer [39], which introduces divided space-time attention to reduce the quadratic complexity of naive self-attention, making transformer models practical for longer clips. Further architectural refinements are introduced by the Video Swin Transformer [40], which applies a hierarchical, shifted-window approach to video, enabling the efficient computation of multi-scale spatio-temporal features and achieving leading performance on major action recognition benchmarks. These models collectively demonstrate that self-attention is a powerful primitive for video understanding, effectively capturing both global context and fine-grained motion patterns.

4.2 Self-supervised learning for scalable representation learning

A major bottleneck for supervised deep learning is the cost and scale of labeled video data. Self-supervised learning (SSL) has emerged as the dominant paradigm for pre-training video backbones on vast amounts of unlabeled web video. Early contrastive methods, such as those presented in Spatiotemporal Contrastive Video Representation Learning [41], learn useful features by contrasting different augmentations of the same video. A more recent and influential direction is masked autoencoding, inspired by BERT. VideoMAE [42] demonstrates that masking a high proportion of spatio-temporal patches and training a model to reconstruct them is an exceptionally effective pre-task for video, due to the strong spatial-temporal redundancy inherent in video data. Models pre-trained with VideoMAE achieve superior performance on downstream tasks, including retrieval, with minimal fine-tuning. This trend towards SSL

underscores a move away from task-specific feature design and towards learning general-purpose, transferable video representations from the structure of the data itself.

4.3 Large-scale multimodal video-language retrieval

The most significant user-facing advancement in modern CBVR is the shift from low-level visual search to cross-modal retrieval using natural language. This requires jointly modeling video and text in a shared embedding space. The Frozen in Time [43] model establishes a standard and scalable dual-encoder architecture, where separate video and text encoders are trained with a contrastive loss on large-scale datasets like WebVid-2M. This allows for efficient billion-scale retrieval via approximate nearest neighbor search. Concurrently, the community has leveraged powerful image-text foundation models. CLIP4Clip [44] provides an extensive empirical study showing that image-text models like CLIP can be effectively adapted for video retrieval through simple temporal pooling strategies, highlighting the transferability of knowledge from web-scale image-text pre-training. These approaches have made natural language video retrieval a reality, enabling queries with complex semantic intent (e.g., "a person joyfully reuniting with a dog at an airport").

4.4 Implications and future trajectory

The convergence of transformer architectures, self-supervised pre-training, and large-scale multimodal learning has redefined the benchmarks for CBVR. The current state-of-the-art is characterized by models that are pre-trained on massive, often noisy, web-collected video-text datasets (e.g., HowTo100M, WebVid). The immediate research challenges are no longer solely about accuracy but also about efficiency (deploying billion-parameter models), temporal grounding (localizing moments within long videos), and robustness to real-world distribution shifts. Furthermore, the intersection of retrieval with generative AI (e.g., using diffusion models for query augmentation or video synthesis) presents a fertile new research frontier. Modern CBVR systems are thus evolving from closed, feature-centric systems to open, scalable platforms for multimodal video understanding.

5. Applications of CBVR

Video has become an essential source of information in almost every field of life, and multimedia is widely used by people in their daily activities. The content of videos is particularly valuable for video search and browsing. While static images rely primarily on low-level features, videos utilize both static features and high-level features such as objects, motion, and activities. Content-based video search and retrieval (CBVR) enables faster and more efficient identification of relevant video content. CBVR has numerous practical applications, with several major use cases implemented worldwide. Video retrieval can be considered from the perspective of different user types, such as novice/non-technical users and professional/trained users. Although retrieval systems can benefit both groups, their requirements differ significantly and must be taken into account during system development. For example, intelligence agencies and television broadcast companies maintain extensive video and photo archives, and it is essential to design the system with the intended users in mind. Interfaces for video retrieval may include text queries, example objects, date information, sketch-based queries, and other semantic parameters. Novice users may struggle to retrieve videos using such complex queries; therefore, retrieval systems for non-technical users should feature simple, intuitive interfaces that allow users to interact easily and access relevant videos. In contrast, professional users who are familiar with searching multimodal video or image data in large organizational archives or web content can leverage advanced interfaces to query multimedia based on complex structures and semantic concepts. In these systems, videos are automatically indexed using various

techniques and then made available for browsing or retrieval, enabling professional users to efficiently access the desired content.

5.1 Archival search and browsing

A primary application of content-based retrieval is the efficient and effective search and browsing of large-scale video archives. Major news agencies and television broadcasters, for instance, maintain extensive archives that can be leveraged by external users for diverse needs. While traditional methods have relied on indexing to manage this complex data, often limiting constructs to textual tags, object labels, or sketches to facilitate user queries, content-based approaches enable more intuitive and direct searching of the visual material itself.

5.2 Multimedia retrieval in e-learning

Large digital media libraries significantly enhance the education sector by allowing efficient retrieval of multimedia content based on its actual video or audio substance. Educational institutions maintain such archives locally or via the internet to serve students and educators. Digital university libraries are a prime example, providing centralized access to a wide array of multimedia resources including text, images, audio, and video. Platforms like YouTube also serve as vast repositories of educational material, where content-based retrieval techniques can help users find specific instructional videos based on visual or topical content rather than metadata alone.

5.3 Automated behavior analysis for security

Video surveillance represents a critical application domain for content-based video retrieval. Organizations implement security systems that archive footage from cameras placed in strategic locations for monitoring activities, assessing behavior, and observing areas of interest. In a retail setting, for example, cameras monitor both interior and exterior store spaces. While traditional systems record continuous video for later review by human operators, advanced automated systems utilize content-based retrieval to analyze footage in real time. These systems can identify specific behaviors, such as potential theft, by extracting and analyzing key frames where suspicious activity is detected, thereby triggering automatic alerts without constant human oversight.

5.4 Content discovery in entertainment platforms

The entertainment industry is a major beneficiary of CBVR technology. Online video platforms index and store vast libraries of content based on structural elements like shots, scenes, and programs, along with associated descriptive information. This content-level indexing allows users to search with great flexibility; for instance, a user might search for a specific movie scene, a type of shot, or an entire program, navigating the entertainment archive through its visual and narrative components rather than just titles or genres.

5.5 Telemedicine and remote clinical support

CBVR facilitates remote instruction by enabling real-time analysis of video content. In telemedicine, for example, a doctor can remotely view a live video feed from a clinical setting. A retrieval system can extract key frames capturing a patient's critical behavior or condition. The physician can then review this visual information and provide immediate, informed instructions to on-site medical staff, enhancing the quality and timeliness of remote care.

5.6 Video summarization for efficient browsing

To improve accessibility, video content is processed using indexing and summarization techniques that generate a concise overview or abstract of a complete program. Video summarization and annotation methods condense a video by identifying and presenting its most important scenes in a shortened format, such as a trailer. This abstract version must accurately represent the essence of the full video. Users can browse these summaries, created by extracting and integrating representative key frames from each major video shot, to quickly evaluate content before deciding to view or download the complete version, streamlining the discovery and access process.

6. Discussion

The proliferation of digital video content, driven by platforms like YouTube and the demands of domains from entertainment to surveillance, has made Content Based Video Retrieval (CBVR) not merely an academic pursuit but a critical technological need. The central challenge remains unchanged: bridging the semantic gap between the user's high-level information need and the low-level visual data stored in databases. Over the past decades, the research community, supported by initiatives like the TRECVID benchmark, has made significant strides, evolving from hand-crafted feature engineering to data-driven deep learning paradigms. This review has systematically charted this evolution. The review examines the TRECVID 2002 feature extraction evaluation [5], in which 13 runs submitted by eight groups, such as CMU, IBM Research, Microsoft Research Asia, MediaMill, and CLIPS-LIMSI, are considered. The assessment is carried out by comparing the performance of each group's retrieval system on a shared dataset. Each system returns at most the top 1000 shots per feature. Table 4 shows the participants and the features they employ in their retrieval systems.

Table 4: Features Used for CBVR System

Name	Features Used	No. of features
CMU_r1	Outdoor, Face	2
CMU_r2	People, Text	2
CLIPS-LIT_GEOD	Face, Speech, Monolog	3
CLIPS-LIT-LIMSU	Speech, Monolog	2
DCUFE2002	Speech	1
Eurecom1	Outdoor, Cityscape, Text	3
Fudan_FE_Sys1	Indoor, People, Landscape, Sound	4
Fudan_FE_Sys2	Indoor, Cityscape, Sound	3
IBM-1	Outdoor, Indoor, Face, Cityscape, Landscape, Text, Speech, Monolog	8
IBM-2	Outdoor, Face, Text	3
MediaMill1	Outdoor, Speech	2
MediaMill2	Monolog, Speech	2
MSRA	Outdoor, Indoor, Face, Cityscape, Text, Sound	6

Numerous retrieval systems have been developed by the research community. The second column of Table 4 enumerates the features utilized by each system. These systems are typically designed for specific purposes. For instance, a system that employs speech and monologue features is likely intended for audio content retrieval or for locating videos featuring a specific individual. These retrieval systems generally employ combination-based queries to retrieve multimedia content. Their performance is evaluated using the TRECVID benchmark, with precision serving as the primary metric. The Average Precision (AP) computes the average precision values at the ranks where relevant items are retrieved, providing a single score that rewards both high precision and high recall within the top results. AP is computed for each feature set, and graphical charts are used to illustrate the performance of the retrieval systems developed by various

organizations.

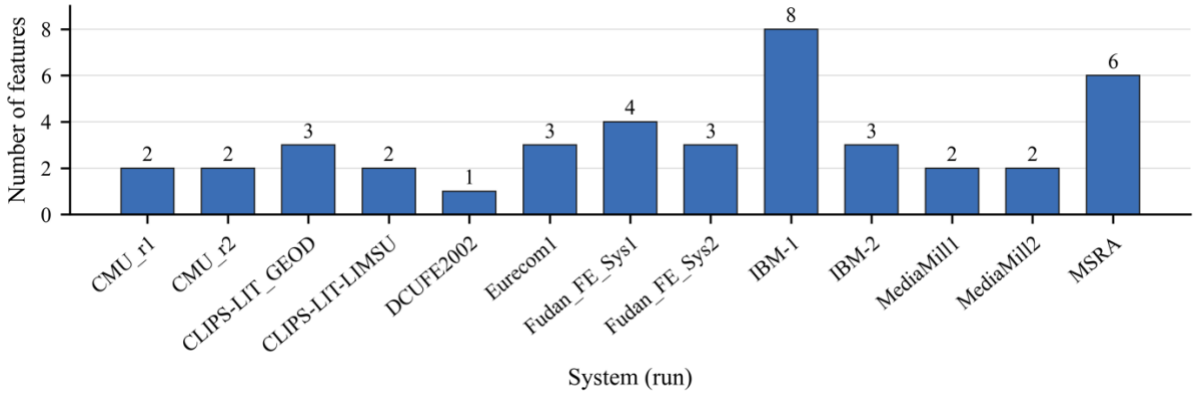

Figure 5: Number of features used by each system

Figure 5 presents a comparative analysis of the number of semantic features utilized by each system included in our benchmark. The results show a clear variation in feature complexity across systems. Six of the 13 systems rely on a minimal set of one or two features, reflecting lightweight designs aimed at optimizing specific semantic categories such as Speech, Face, or Monolog. Mid-range systems, including Eurecom1 and the Fudan submissions, integrate three to four features, suggesting a more balanced multimodal strategy. In contrast, IBM-1 and MSRA exhibit substantially richer feature representations, using eight and six features respectively. These systems leverage diverse combinations spanning indoor–outdoor cues, scene context, human presence, and audio signals, which likely contribute to more robust content interpretation. Overall, the results indicate that systems with broader feature sets tend to adopt more comprehensive multimodal modeling approaches.

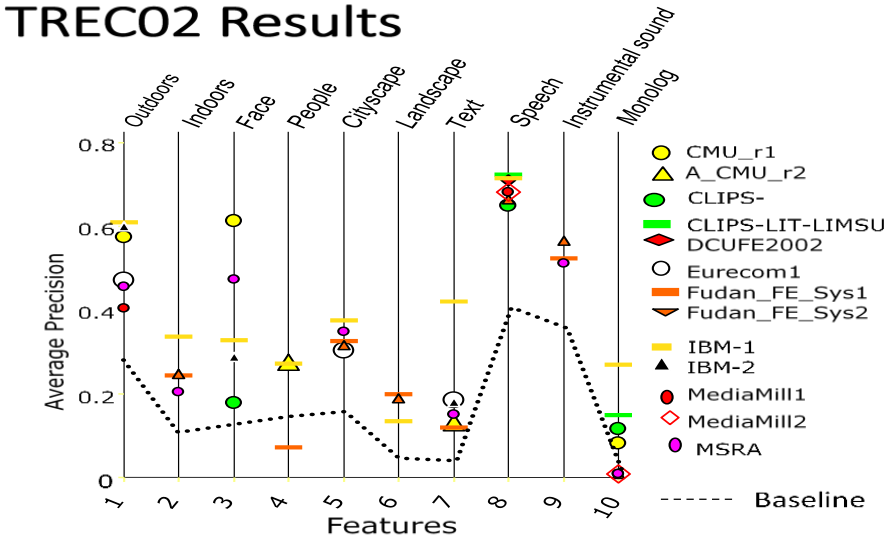

Figure 6: TRECVID 2002 feature extraction results (adapted from [5])

The TREC02 feature extraction results demonstrate substantial variation in system performance across different semantic categories. As shown in Figure 6, audio-related features such as Speech and Instrumental Sound consistently achieve the highest average precision values, with several systems exceeding the 0.60 mark on Speech. This indicates that these acoustic cues are more reliably captured and discriminated by early multimedia retrieval algorithms. In contrast, visually complex categories, including Cityscape and Landscape, and the Monolog feature exhibit the lowest precision, often remaining close to the baseline. Visual categories such as Outdoors, Indoors, Face, and People achieve moderate performance, reflecting

the limitations of early 2000s computer vision techniques in handling high intra-class variation and scene complexity. Overall, the results highlight a clear performance gap between speech and sound detection and most visual feature detection, emphasizing the relative maturity of audio-based models during the TREC02 evaluation period.

The following discussion synthesizes the key trends emerging from this body of work, articulates the persistent open challenges that constrain current systems, and proposes concrete avenues for future research.

6.1 Synthesis of major trends

The trajectory of CBVR research reveals several dominant, interconnected trends:

6.1.1 From Hand-Crafted to Learned Representations

The field has decisively shifted from designing static features (color, texture, motion histograms) towards architectures (CNNs, Transformers) that learn optimal, hierarchical representations directly from data. This has been the primary driver in narrowing the semantic gap.

6.1.2 From Single Modality to Multimodal Fusion

Early systems have relied solely on visual content. The trend now is firmly towards multimodal understanding, integrating audio, text (speech transcripts, subtitles, user queries), and even temporal metadata to create a richer, more query-aligned representation of video content.

6.1.3 From Shallow to Deep Temporal Modeling

Modeling time has progressed from simple keyframe sequencing and statistical motion features to sophisticated long-range temporal reasoning using 3D CNNs, LSTMs, and most recently, spatio-temporal vision transformers (e.g., ViViT, TimeSformer), which capture complex action dynamics and narrative structure.

7. Conclusion

CBVR has emerged as a critical approach for efficiently managing and accessing the exponentially growing volume of multimedia data. The increasing complexity of modern video content necessitates robust indexing and retrieval mechanisms capable of handling large-scale datasets in real time. Since indexing forms the foundation of video retrieval, the effectiveness of these systems relies heavily on the quality of visual and semantic features extracted from key frames, shots, and scenes. This survey has presented the fundamental concepts of video indexing and retrieval, highlighting the techniques used to extract visual features that closely align with human perception. Applications of visual content–based video retrieval have been discussed to emphasize its practical significance. Furthermore, the TRECVID evaluation framework has been reviewed, demonstrating how feature selection and system design influence retrieval efficiency, accuracy, and reliability in large video archives. Future research should focus on advanced feature extraction techniques, integration of semantic understanding, and adaptive retrieval frameworks to further enhance system performance. Additionally, exploring multimodal approaches that combine visual, audio, and textual cues may offer significant improvements in retrieval relevance and robustness. Overall, continued development in these areas will advance the effectiveness of visual CBVR systems in real-world applications.

Funding: No specific funding has been received for this research.

Data availability: No new data have been generated in this review article.

Conflicts of interest: The authors declare no conflict of interest.

Authors contributions: Conceptualization: AKB, UI; methodology: ZI, UI; validation: AKB, UI, ZI; writing, original draft preparation: ZI, UI, AKB; writing, review and editing: AKB, UI, ZI, UI; visualization, supervision and project administration: AKB, UI. The authors have approved the final version.

AI usage disclosure: The AI tool Grammarly has been used only for language polishing and grammatical correction. All scientific and technical content has been created and validated by the authors.

Ethics approval: Not applicable

References

[1] Sun, M. (2024). "An intelligent retrieval method for audio and video content: deep learning technology based on artificial intelligence". *IEEE Access*, 12, 123430-123446. https://doi.org/10.1109/ACCESS.2024.3450920

[2] Singh, A. (2024). "Content-based image and video retrieval based on hybrid feature extraction techniques". *International Journal of Intelligent Systems and Applications in Engineering*, 12(4), 3323. https://doi.org/10.17762/ijisae.v12i4.6832

[3] Tian, M., Li, G., Qi, Y., et al. (2024). "Rethink video retrieval representation for video captioning". *Pattern Recognition*, 156, 110744. https://doi.org/10.1016/j.patcog.2024.110744

[4] Liu, Z., and Song, R. (2025). "Survey of dense video captioning: techniques, resources, and future perspectives". *Applied Sciences*, 15(9), 4990. https://doi.org/10.3390/app15094990

[5] Smeaton, A. F., and Over, P. (2003). "The TREC-2002 video track report". In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, NIST Special Publication 500-251. NIST. https://doi.org/10.6028/NIST.SP.500-251.video-overview

[6] Hu, W., Xie, D., Fu, Z., et al. (2007). "Semantic-based surveillance video retrieval". *IEEE Transactions on Image Processing*, 16(4), 1168-1181. https://doi.org/10.1109/TIP.2006.891352

[7] Sivic, J., and Zisserman, A. (2006). "Video Google: efficient visual search of videos". In *Toward Category-Level Object Recognition, Lecture Notes in Computer Science*, 4170 (pp. 127-144). Springer. https://doi.org/10.1007/11957959_7

[8] Aytar, Y., Shah, M., and Luo, J. (2008). "Utilizing semantic word similarity measures for video retrieval". In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE. https://doi.org/10.1109/CVPR.2008.4587822

[9] Kennedy, L. S., Natsev, A. P., and Chang, S. F. (2005). "Automatic discovery of query-class-dependent models for multimodal search". In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 882-891). ACM. https://doi.org/10.1145/1101149.1101339

[10] Yan, R., and Hauptmann, A. G. (2006). "Probabilistic latent query analysis for combining multiple retrieval sources". In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 324-331). ACM. https://doi.org/10.1145/1148170.1148228

[11] Browne, P., and Smeaton, A. F. (2005). "Video retrieval using dialogue, keyframe similarity and video objects". In *IEEE International Conference on Image Processing 2005* (pp. III-1208). IEEE. https://doi.org/10.1109/ICIP.2005.1530615

[12] Wu, Y., Zhuang, Y., and Pan, Y. (2000). "Content-based video similarity model". In *Proceedings of the Eighth ACM International Conference on Multimedia* (pp. 465-467). ACM. https://doi.org/10.1145/354384.376380

[13] Snoek, C. G. M., Huurnink, B., Hollink, L., et al. (2007). "Adding semantics to detectors for video retrieval". *IEEE Transactions on Multimedia*, 9(5), 975-986. https://doi.org/10.1109/TMM.2007.900156

[14] Le, T. L. (2011). "Relevance feedback for surveillance video retrieval at object level". In *Future Information Technology, Communications in Computer and Information Science*, 185 (pp. 175-182). Springer. https://doi.org/10.1007/978-3-642-22309-9_21

[15] Schoeffmann, K., and Leopold, M. (2025). "AI-based video content understanding for automatic and interactive multimedia retrieval". In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 3750-3758). IEEE. https://doi.org/10.1109/CVPRW67362.2025.00360

[16] Joachims, T. (2002). "Optimizing search engines using clickthrough data". In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133-142). ACM. https://doi.org/10.1145/775047.775067

[17] Ghosh, H., Poornachander, P., Mallik, A., et al. (2007). "Learning ontology for personalized video retrieval". In *Workshop on Multimedia Information Retrieval on the Many Faces of Multimedia Semantics* (pp. 39-46). ACM. https://doi.org/10.1145/1290067.1290075

[18] Khan, M. A., Javed, K., Khan, S. A., et al. (2024). "Human action recognition using fusion of multiview and deep features: an application to video surveillance". *Multimedia Tools and Applications*, 83(5), 14885-14911. https://doi.org/10.1007/s11042-020-08806-9

[19] Sivic, J., Everingham, M., and Zisserman, A. (2005). "Person spotting: video shot retrieval for face sets". In *Image and Video Retrieval (CIVR 2005), Lecture Notes in Computer Science*, 3568 (pp. 226-236). Springer. https://doi.org/10.1007/11526346_26

[20] Luan, H., Neo, S. Y., Goh, H. K., et al. (2007). "Segregated feedback with performance-based adaptive sampling for interactive news video retrieval". In *Proceedings of the 15th ACM International Conference on Multimedia* (pp. 293-296). ACM. https://doi.org/10.1145/1291233.1291295

[21] Nallappan, M., and Velswamy, R. (2024). "Exploring deep learning-based content-based video retrieval with Hierarchical Navigable Small World index and ResNet-50 features for anomaly detection". *Expert Systems with Applications*, 247, 123197. https://doi.org/10.1016/j.eswa.2024.123197

[22] Ragedhaksha, Darshini, Shahil, et al. (2023). "Deep learning-based real-world object detection and improved anomaly detection for surveillance videos". *Materials Today: Proceedings*, 80, 2911-2916. https://doi.org/10.1016/j.matpr.2021.07.064

[23] Deekshita, P., Bonu, V., Ramyasri, A., et al. (2025). "A hybrid CBIR framework using vision transformers and genetic algorithm for enhanced image retrieval". *Journal of Applied Science and Technology Trends*, 6(2), 277-289. https://doi.org/10.38094/jastt62304

[24] Peronikolis, M., and Panagiotakis, C. (2024). "Personalized video summarization: a comprehensive survey of methods and datasets". *Applied Sciences*, 14(11), 4400. https://doi.org/10.3390/app14114400

[25] Gia, B. T., Khanh, T. B. C., Thanh, T. L. T., et al. (2025). "NII-UIT at VBS2025: Multimodal video retrieval with LLM integration and dynamic temporal search". In *MultiMedia Modeling (MMM 2025), Lecture Notes in Computer Science* (pp. 318-325). Springer. https://doi.org/10.1007/978-981-96-2074-6_38

[26] Guironnet, M., Pellerin, D., Guyader, N., et al. (2007). "Video summarization based on camera motion and a subjective evaluation method". *EURASIP Journal on Image and Video Processing*, 2007, 060245. https://doi.org/10.1155/2007/60245

[27] Xie, X., and Wu, F. (2008). "Automatic video summarization by affinity propagation clustering and semantic content mining". In *2008 International Symposium on Electronic Commerce and Security* (pp. 203-208). IEEE. https://doi.org/10.1109/ISECS.2008.118

[28] Xiao, R., Wang, Y., Pan, H., et al. (2008). "Automatic video summarization by spatio-temporal analysis and non-trivial repeating pattern detection". In *2008 Congress on Image and Signal Processing* (pp. 555-559). IEEE. https://doi.org/10.1109/CISP.2008.601

[29] Gong, Y. (2003). "Summarizing audiovisual contents of a video program". *EURASIP Journal on Advances in Signal Processing*, 2003(2). https://doi.org/10.1155/S1110865703211082

[30] Wang, F., and Ngo, C. W. (2007). "Rushes video summarization by object and event understanding". In *Proceedings of the International Workshop on TRECVID Video Summarization* (pp. 25-29). ACM. https://doi.org/10.1145/1290031.1290035

[31] Hu, W., Xie, N., Li, L., et al. (2011). "A survey on visual content-based video indexing and retrieval". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6), 797-819. https://doi.org/10.1109/TSMCC.2011.2109710

[32] Park, J., Kim, J., Seok, J., et al. (2024). "Contrasting multi-modal similarity framework for video scene segmentation". *IEEE Access*, 12, 32408-32419. https://doi.org/10.1109/ACCESS.2024.3370676

[33] Sameer, R. A. (2025). "The performance of different video segmentation or video shot boundary detection techniques: a survey". *AL-Rafidain Journal of Computer Sciences and Mathematics*, 19(1), 134-143. https://doi.org/10.33899/csmj.2025.156866.1165

[34] Kar, T., Kanungo, P., Mohanty, S. N., et al. (2024). "Video shot-boundary detection: issues, challenges and solutions". *Artificial Intelligence Review*, 57(4). https://doi.org/10.1007/s10462-024-10742-1

[35] Chavate, S., Mishra, R., Singh, S. K., et al. (2023). "Hybridized neural network based approaches used for video shot boundary detection". *Indian Journal of Science and Technology*, 16(33), 2670-2680. https://doi.org/10.17485/ijst/v16i33.1671

[36] Sunuwar, J., and Borah, S. (2024). "A comparative analysis on major key-frame extraction techniques". *Multimedia Tools and Applications*, 83(30), 73865-73910. https://doi.org/10.1007/s11042-024-18380-z

[37] Tseng, S. M., Yeh, Z. T., Wu, C. Y., et al. (2023). "Video scene detection using transformer encoding linker network (TELNet)". *Sensors*, 23(16), 7050. https://doi.org/10.3390/s23167050

[38] Arnab, A., Dehghani, M., Heigold, G., et al. (2021). "ViViT: A video vision transformer". In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6816-6826). https://doi.org/10.1109/ICCV48922.2021.00676

[39] Bertasius, G., Wang, H., and Torresani, L. (2021). "Is space-time attention all you need for video understanding?". In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 139 (pp. 813-824). PMLR. https://doi.org/10.48550/arXiv.2102.05095

[40] Liu, Z., Ning, J., Cao, Y., et al. (2022). "Video Swin Transformer". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3192-3201). https://doi.org/10.1109/CVPR52688.2022.00320

[41] Qian, R., Meng, T., Gong, B., et al. (2021). "Spatiotemporal contrastive video representation learning". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6960-6970). https://doi.org/10.1109/CVPR46437.2021.00689

[42] Tong, Z., Song, Y., Wang, J., et al. (2022). "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training". *Advances in Neural Information Processing Systems*, 35, 10078-10093. https://doi.org/10.52202/068431-0732

[43] Bain, M., Nagrani, A., Varol, G., et al. (2021). "Frozen in time: A joint video and image encoder for end-to-end retrieval". In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1708-1718). https://doi.org/10.1109/ICCV48922.2021.00175

[44] Luo, H., Ji, L., Zhong, M., et al. (2022). "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning". *Neurocomputing*, 508, 293-304. https://doi.org/10.1016/j.neucom.2022.07.028