applied sciences

MDPI

Review

Personalized Video Summarization: A Comprehensive Survey of Methods and Datasets

Michail Peronikolis † **and Costas Panagiotakis** *,†

Department of Management Science and Technology, Hellenic Mediterranean University, P.O. Box 128, 72100 Agios Nikolaos, Crete, Greece; peronikolis@hmu.gr

* Correspondence: cpanag@hmu.gr

† These authors contributed equally to this work.

Abstract: In recent years, the scientific and technological developments have led to an explosion of available videos on the web, increasing the necessity of fast and effective video analysis and summarization. Video summarization methods aim to generate a synopsis by selecting the most informative parts of the video content. The user's personal preferences, often involved in the expected results, should be taken into account in the video summaries. In this paper, we provide the first comprehensive survey on personalized video summarization relevant to the techniques and datasets used. In this context, we classify and review personalized video summary techniques based on the type of personalized summary, on the criteria, on the video domain, on the source of information, on the time of summarization, and on the machine learning technique. Depending on the type of methodology used by the personalized video summarization techniques for the summary production process, we classify the techniques into five major categories, which are feature-based video summarization, keyframe selection, shot selection-based approach, video summarization using trajectory analysis, and personalized video summarization using clustering. We also compare personalized video summarization methods and present 37 datasets used to evaluate personalized video summarization methods. Finally, we analyze opportunities and challenges in the field and suggest innovative research lines.

Keywords: video summarization; recommender systems; video segmentation; personalized video summary

Citation: Peronikolis, M.; Panagiotakis, C. Personalized Video Summarization: A Comprehensive Survey of Methods and Datasets. *Appl. Sci.* **2024**, *14*, 4400. https://doi.org/10.3390/app14114400

Academic Editor: Samuel Cheng

Received: 16 April 2024
Revised: 17 May 2024
Accepted: 19 May 2024
Published: 22 May 2024

1. Introduction

In recent years, the rapid advancement of technology has led to the integration of camcorders into many devices. As the number of camcorders increases, so does the number of recorded videos [1]. This results in a huge increase in videos that are uploaded daily on the Internet [2]. Daily activities or special moments are recorded, resulting in large amounts of video [3]. Human access to multimedia creation, mainly through mobile devices and the tendency to share them through social networks [3], causes an explosion of videos available on the Web. Searching for specific video content and categorization is generally time-consuming. The traditional representation of video files as a sequence of numerous consecutive frames, each of which corresponds to a constant time interval, while adequate for viewing a file in a movie mode, presents a number of limitations for the new emerging multimedia services such as content-based search, retrieval, navigation, and video browsing [4]. The need for rational time management led to the development of automatic video content summarization [2] and indexing/clustering [1] techniques to facilitate access, content search, and automatic categorization (tagging/labeling), as well as action recognition [5] and common action detection in videos [6]. The number of papers per year that contain in their title the phrase "video summarization" according to Google Scholar is depicted in Figure 1.

Many studies have been devoted to developing and designing tools that have the ability to create videos with a duration shorter than the original video, reflecting the most

important visual and semantic content [4,7]. As users grow in the base, so does the diversity between them. A summary which is uniform for all users may not suit everyone's needs. Each user can consider different important sections according to his interests, needs, and the time he will spend. Therefore, the focus should be on the personalized summary of the general video summary [8]. User preferences, which are often involved in the expected results, should be taken into account in video summaries [9]. Therefore, it is important to modify the video summary to suit the user's interests and preferences, thus creating a personalized video summary, while retaining important semantic content from the original video [10].

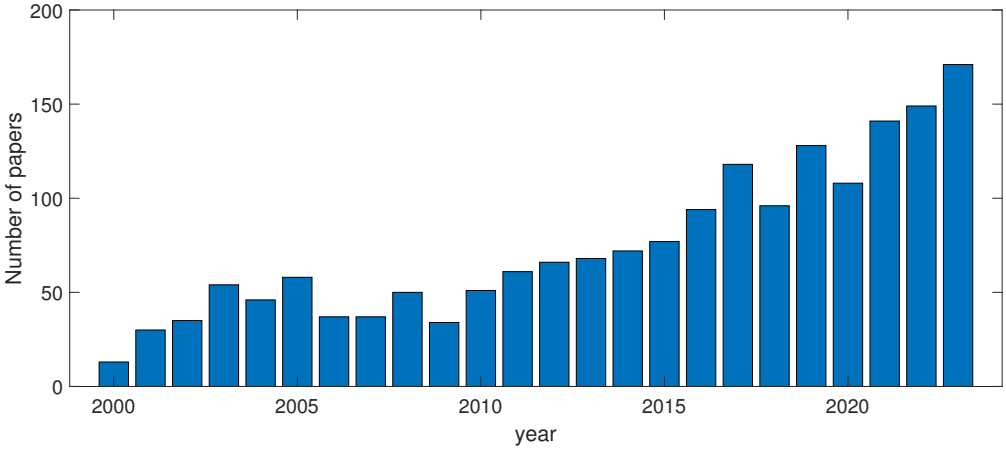

Figure 1. Thenumber of papers per year containing in their title the phrase "video summarization" according to google scholar.

Video summaries that reflect the understanding of individual users about the content of the video should be personalized in a way that is based on individual needs and intuitive. Consequently, the personalized video summary is tailored to the individual user's understanding of the video content or understanding of the content of the video. In contrast, video summaries that are not personalized are not customized in any way to the understanding of the individual user. In several ways, summaries can be personalized. For example, personalizing video summaries can involve pre-filtering content via user profiles before showing it to the user, creating summaries tailored to the user's behavior and movements during video capture, and customizing summaries according to the user's unique browsing and access habits [11].

Several studies on personalized video summarization have appeared in the literature. In [12], Tsenk et al. (2001) introduced personalized video summarization for mobile devices. After two years, Tseng et al. (2003) [13] used context information to generate hierarchical video summarization. Lie and Hsu (2008) [14] proposed a personalized summary video framework from the semantic extraction features of each frame. Shafeian and Bhanu (2012) [15] proposed a personalized system for video summarization and retrieval. Zhang et al. (2013) [16] proposed a personalized video summarization that is interactive and based on sketches. Panagiotakis et al. (2020) [9] provided personalized video summarization through a recommender system, where the output is personalized rankings of video segments. Hereafter, a rough classification of the current research is presented:

- Many works were carried out that produced a personalized video summary in real time. Valdés and Martínez (2010) [17] introduced an application for interactive video summarization in real time on the fly. Chen and Vleeschouwer (2010) [18] produced personalized basketball video summaries in real time from data from multiple streams.
- Several works were based on queries. Kannan et al. (2015) [19] proposed a system to create personalized movie summaries with a query interface. Given a given semantic query, Garcia (2016) [20] proposed a system that can find relevant digital memories

and perform personalized summarization. Huang and Worring (2020) [21] created a dataset based on a query–video pair.

- In a few works, information was extracted from humans using sensors. Katti et al. (2011) [22] used eye gaze and pupillary dilation to generate storyboard personalized video summaries. Qayyum et al. (2019) [23] used electroencephalography to detect viewer emotion.
- Few studies focused on egocentric personalized video summarization. Varini et al. (2015) [24] proposed a personalized egocentric video summarization framework. The most recent study by Nagar et al. (2021) [25] presented a framework of unsupervised reinforcement learning in daylong egocentric videos.
- Many studies have been conducted in which machine learning has been used in the development of a video summarization technique. Park and Cho (2011) [26] proposed the summarization of personalized live video logs from a multicamera system using machine learning. Peng et al. (2011) [27] proposed personalized video summarization by supervised machine learning using a classifier. Ul et al. (2019) [28] used the deep CNN model to recognize facial expression. Zhou et al. (2019) [29] proposed a Character-Oriented Video Summarization framework. Fei et al. (2021) [30] proposed a triplet deep-ranking model for personalized video summarization. Mujtaba et al. (2022) [31] proposed a framework for personalized video summarization using 2D CNN. öprü and Erzin (2022) [32] used Affective Visual Information for Human-Centric Video Summarization. Ul et al. (2022) [33] presented Object of Interest (OoI), a personalized video summarization framework based on the Object of Interest.

Figure 2 presents milestone works in personalized video summarization. This figure also shows the first paper published for the personalized video summary and the long time it took for the works to begin to be published en masse.

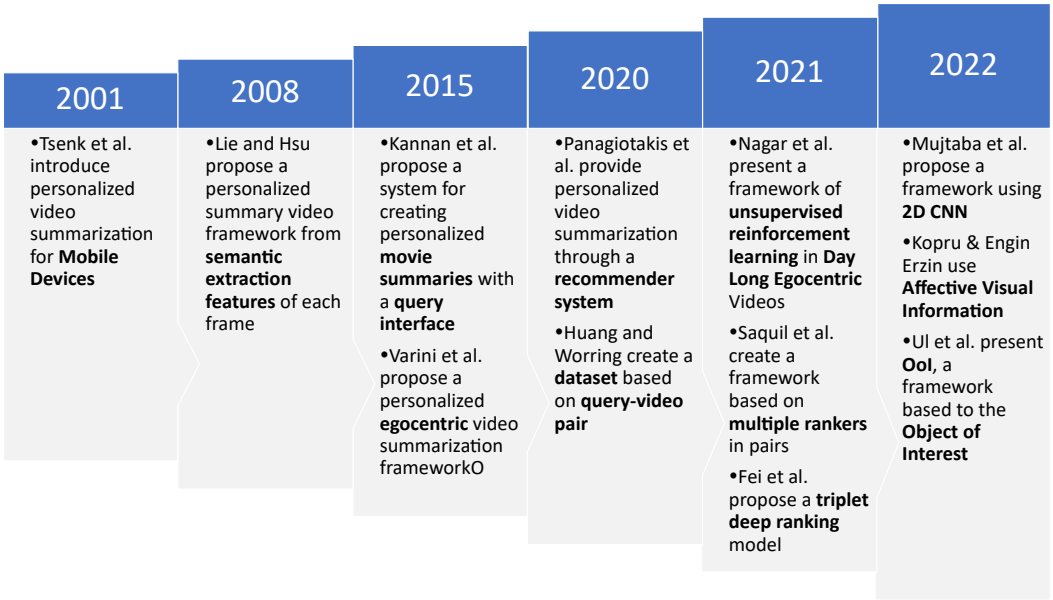

Figure 2. Milestones in personalized video summarization ([12] (2001), [14] (2008), [19,24] (2015), [9,21] (2020), [25,30,34] (2021), [31–33] (2022)).

The remainder of this paper is structured as follows. Section 2 introduces the main applications of the personalized video summary. Section 3 presents approaches to summarize personalized video summaries by separating audiovisual cues. Section 4 presents the classification of personalized video summary techniques into six categories according to the type of personalized summary, criteria, video domain, source of information, time of summarization, and machine learning technique. Section 5 classifies the techniques into five main categories according to the type of methodology used by the personalized video summarization techniques. Section 6 describes the individualized video datasets suitable

for summarization. Section 7 describes the evaluation of personalized video summarization methods. In Section 8, the quantitative comparison of personalized video summarization approaches is presented on the most prevalent datasets. In Section 9, this work is concluded by briefly describing the main results of the study conducted. In addition, opportunities and challenges in the field and suggested innovative research lines are analyzed.

2. Applications of Personalized Video Summarization

Nowadays, with the increasing affordability of video recorders, there has been an increase in the number of videos recorded. Personalized video summarization is crucial to help people manage their videos effectively by providing personalized summaries and is useful for a number of purposes. Listed below are the main applications of personalized video summarization.

2.1. Personalized Trailer of Automated Serial, Movies, Documentaries and TV Broadcasting

For the same film, there is a large discrepancy in users' preferences for summarization, so nowadays, it is necessary to create a personalized movie summarization system [19]. Personalized video summarization is useful for creating series and movie trailers to make each viewer watch the series or movies more interesting, as they focus on their interests. Examples of such works are presented below.

Darabi and Ghinea [35] proposed a framework to produce personalized video summaries based on ratings assigned by video experts in video frames, and their second method could demonstrate good results for the movie category. Kannan et al. [19] proposed a system to create movie summaries with semantic meaning for the same video, which are tailored to the user's preferences and interests and which are collected from a query interface. Fei et al. [36] proposed an event detection-based framework and expanded it to create a personalized summary of video film and soccer. Ul et al. [28] presented a personalized movie summarization scheme using deep CNN-Assisted Recognition of facial expressions. Mujtaba et al. [31] proposed a lightweight client-driven personalized video summarization framework that aimed to create subsets for long videos, such as documentaries and movies.

2.2. Personalized Sport Highlights

Sports footage is produced using manual methods; editing requires experience and is performed using any video editing tool [37]. The personalized video summary is useful for creating sports highlights so that each viewer can view the snapshots that interest them from each sport activity. Examples of such works are presented below.

Chen and Vleeschouwer [18] proposed a flexible framework to produce personalized basketball video summaries in real time, integrating general production principles, narrative user preferences on story patterns, and contextual information. Hannon et al. [38] generated video highlights by summarizing full-length matches. Chen et al. [39] proposed a complete autonomous framework production process for personalized basketball video summaries from multisensor data from multiple cameras. Olsen and Moon [40] described how the user interaction of previous viewers could be used to create a personalized video summarization, computing more interesting plays in a game. Chen et al. [41,42] produced a hybrid personalized video summarization framework by summarizing a broadcast soccer video that combines content truncation and adaptive fast forwarding. Kao et al. proposed a personal video summarization system that integrates the global positioning system and the radio frequency identification information for marathon activities [43]. Sukhwani and Kothari [44] presented an approach that can be applied by multiple matches to create complex summaries or the summary of a match. Tejero-de-Pablos et al. [45] proposed a framework that uses neural networks to generate a personalized video summarization that can be applied to any sport in which games have a sequence of actions. To select the highlights of the original video, they used the actions of the players as a contract. To evaluate the method, this project uses the case of Kendo.

2.3. Personalized Indoor Video Summary

Creating a personalized home video summary is very important, as each user can view events that interest them according to their preferences without having to watch the entire video. Niu et al. [46] presented an interactive mobile-based real-time home video summarization application that reflects user preferences. Park and Cho [26] presented a method by which the video summarization of life logs from multiple sources is performed in an office environment. Peng et al. [27] presented a system to automatically generate personalized home video summaries based on user behavior.

2.4. Personalized Video Search Engines

The Internet contains a large amount of video content. Consider a scenario in which a user is looking for videos on a specific topic. Search engines present the user with numerous video results [47]. Various search engines are used based on the user's interest to contain video topics in the form of a short video clip format [37]. SH-DPP was proposed by Sharghi et al. [48] to generate query-focused video summaries to be useful to search engines, such as displaying video clips.

2.5. Personalized Egocentric Video Summarization

Plenty of egocentric cameras have led to the daily production of thousands of egocentric videos. A large proportion of these self-centered videos are many hours long, so it is necessary to produce a personalized video summarization so that each viewer can watch the snapshots that interest them. Watching egocentric videos is difficult from start to finish due to extreme camera shakes and their unnecessary nature. Summary tools are effectively required by these videos for consumption. But traditional summary techniques developed for static surveillance films or videos and sports videos cannot be adapted to egocentric videos. However, it is a limitation to focus on important people and objects using specialized summary techniques developed for egocentric videos [25]. To summarize egocentric videos that are a day long, Nagar et al. [25] proposed an unsupervised learning framework to generate personalized summaries in terms of both content and length.

3. Approaches for Summarizing Personalized Videos

To achieve a concise and condensed presentation of the content of a video stream, video summaries have several audiovisual cues [11]. The goal of each personalized video summary is to keep the audiovisual elements that they incorporate intact. The audiovisual cues used for personalized video summaries are shown in Figure 3 and are classified as follows:

- *Keyframe cues* are the most representative and important frames or still images drawn from a video stream in the sequence of time [49]. For example, the personalized video summarization from Köprü and Erzin [32] is configured as a keyframe option that maximizes a scoring function based on sentiment characteristics.
- *Video segment cues* are a dynamic extension of keyframe cues, which are video segment cues [11]. These are the most important part of a video stream, and video summaries produced with dynamic elements in mind manage to preserve both video and audio [49]. These video summaries have the ability to preserve the sound element and movement of the video, making them more attractive to the user. A major drawback is that the user takes longer to understand the content of these videos [11]. Sharghi et al. [48] proposed a framework that checks the relevance of a shot to its importance in the context of the video and in the user's query for inclusion in the summary.
- *Graphical cues* complement other cues using syntax and visual cues [49]. Using syntax as a substitute for other conditions and visual cues shows an extra layer of detail. Users perceive an overview of the content of a video summary in more detail due to embedded widgets that other methods cannot achieve [11]. This is illustrated by Tseng and Smith, who presented a method according to which the annotation tool learns the semantic idea, either from other sources or from the same video sequence [13].

- *Textual cues* use text captions or textual descriptors to summarize content [49]. For example, the multilayered Probabilistic Latent Semantic Analysis (PLSA) model [50] presents a personalized video summarization framework based on contemporary feedback in time offered by multiple users.
- *Social cues* are connections of the user image through the social network. Yin et al. [51] proposed an automatic video summarization framework in which user interests are extracted from their image collections on the social network.

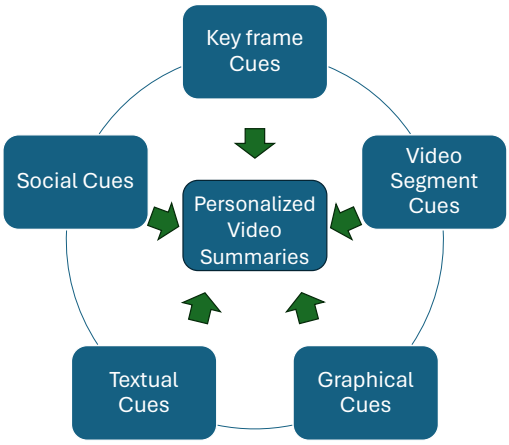

Figure 3. Audiovisual cues used for personalized video summaries.

4. Classification of Personalized Video Summarization Techniques

Many techniques have emerged to create personalized video summaries with the aim of keeping the exact content of the original video intact. Based on the characteristics and properties, the techniques are classified into the following categories that are depicted in Figure 4.

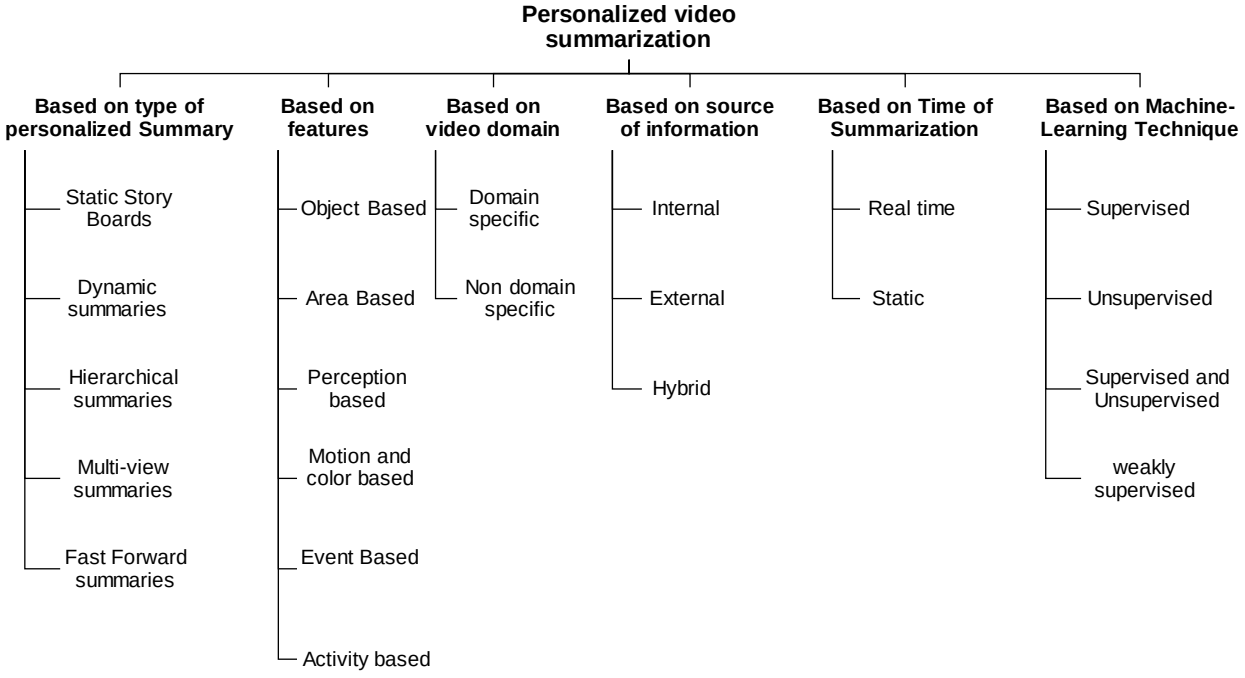

Figure 4. Categories of personalized video summarization techniques.

4.1. Type of Personalized Summary

Video summarization is a technique used to generate a concise overview of a video, which can consist of a sequence of still images (keyframes) or moving images (video skims) [52]. The personalized video summarization process involves creating a personalized summary of a specific video. The desired types of personalized summaries following this process also serve as a foundation for categorizing techniques used in personalized video summarization. Different possible outputs can be considered when approaching the problem of personalized summary, such as the following:

- *Static storyboards* are also called static summaries. Some personalized video summaries consist of a set of static images that extract highlights, and their representation is made into a photo album [15,16,21–23,29,33,35,43,44,48,52–59]. The generation of these summaries is performed by extracting the keyframes according to the user's preferences and the summary criteria.
- *Dynamic summaries* are also called dynamic video skims. Personalized video summarization is performed by selecting a subset of consecutive frames from all frames of the original video that includes the most relevant subshots that represent the original video according to the user's preferences as well as the summary criteria [12,14,17–20,25,27,28,30,31,36,38,39,45,46,48,50,55,60–70].
- A *Hierarchical summary* is a multilevel and scalable summary. It consists of a few abstractive layers, where the lowest layer contains the largest number of keyframes and more details, while the highest layer contains the smallest number of keyframes. Hierarchical video summaries provide the user with a few levels of summary, which provides the advantage of making it easier for users to determine what is appropriate [49]. Based on the preferences of each user, the video summary presents an appropriate overview of their original video. Sachan and Keshaveni [8], to accurately identify the related concept with a frame, proposed a hierarchical mechanism to classify the images. The role of the system that performs the hierarchical classification is the deep categorization of a framework against a classification that is defined. Tseng and Smith [13] suggested a summary algorithm, where server metadata descriptions, contextual information, user preferences, and user interface statements are used to create hierarchical summary video production.
- *Multiview summaries* are summaries created from videos recorded simultaneously by multiple cameras. When watching sports videos, these summaries are useful, as the video is recorded by many cameras. In the multiview summary, challenges can arise that are often due to the overlapping and redundancy of the contents, as well as the lighting conditions and visual fields from the different views. As a result, for static summary output, the basic frame can be extracted with difficulty, and for the video, to shoot border detection. Therefore, in this scenario, conventional techniques for video skimming and extracting keyframes from videos recorded by a camera cannot be applied directly [49]. In personalized video summarization, the multiview video summary is determined based on the preferences of each user and the summary criteria [18,26,39,71].
- *Fast forward summaries:* When a user watches a video that is not informative or interesting, they will often play it fast or move it forward quickly [72]. Therefore, in the personalized video summary, the user wants to fast forward the video segments that are not of interest. In [51], Yin et al., in order to inform the context, users play in fast forward mode the less important parts. Chen et al. [41] proposed an adaptive fast forward personalized video summarization framework that performs clip-level fast forwarding, choosing from discrete options the playback speeds, which include as a special case the cropping of content at a playback speed that is infinite. In personalized video summary, fast forwarding is used in sports videos, where each viewer wants to watch the main phases of the match that are of interest to them according to their preferences, while also having a brief overview of the remaining phases that are not as interesting or important to them. Chen and Vleeschouwer [42] proposed a

personalized video summarization framework for soccer video broadcasts with adaptive fast forwarding, where efficient resource allocation selection selects the optimal combination of candidate summaries.

Figure 5 shows the distribution of the papers reported in Table 1 according to the type of feature of the personalized summary. From this distribution, it is depicted that personalized summaries with a percentage of around 59% are of video skimming type. Next in the ranking are the papers whose personalized summaries are of the storyboard type, with a percentage of around 27%. At a percentage of around 8%, there are works whose personalized summaries are the multiview type. The number of works whose personalized summaries are of the hierarchical type is half that of multiview type since their percentage is around 4%. The smallest number of works is those whose personalized summaries are of the fast forward type, at only around 2%.

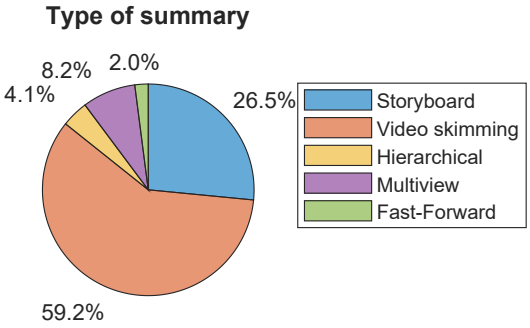

Figure 5. Distribution of the papers reported in Table 1 according to the type of personalized summary technique.

4.2. Features

Each video includes several features that are useful in creating a personalized video summary with the correct representation of the content of the original video. Users focus on one or more video features, such as identifying events, activities, objects, etc., resulting in the user adopting specific summary techniques based on these characteristics selected according to their preferences. The descriptions of the techniques based on these features are presented below.

4.2.1. Object Based

Object-based techniques focus on objects that are specific and present in the video. Techniques are useful in detecting one or more objects, such as a table, a dog, a person, etc. Some object-based summaries may include graphics or text to simplify the process of selecting video segments or keys easier, or to represent the objects contained in the video summary based on the detection of those objects [11]. Video summarization can be performed by collecting all frames with the desired object from the video. If the video does not include some type of desired object, then this method, although executed, will not be effective [37].

In the literature, there are several studies in which the creation of a personalized video summary is based on the user's query, which is taken into consideration [12,13,19,30,48,52,55,64,65,68–70,73]. Sachan and Keshaveni [8] proposed a personalized video summarization approach using classification-based entity identification. They succeeded in creating personalized summaries of desired lengths, thanks to the compression, recognition, and special classification of the macros present in the video. After the completion of video segmentation, the method of Otani et al. [74] selects video segments based on their objects. Each video segment is represented by object tags and their meaning is included. The Object of Interest (OoI), defined by Gunawardena et al. [59], is the user's interest features. They suggested an algorithm that summarizes a video that focuses on a given OoI with the key option oriented to the user's interest. Ul et al. [33] proposed a framework in which, using

deep learning technology, frames with OoIs are detected, which are combined to produce a video digest as output. This framework can detect one or more objects presented in the video. Nagar et al. [25] proposed a framework for egocentric video summarization that uses unsupervised reinforcement learning and 3D convolutional neural networks (CNNs) that incorporate user preferences such as places, faces, and user choice in an interactive way to exclude or include that type of content. Zhang et al. [16] proposed a method to generate personalized video summaries based on sketch selection that relies on the generation of a graph from keyframes. The features of the sketches are represented by the generated graph. The proposed algorithm interactively selects those frames related to the same person so that there is direct user interaction with the video content. Depending on the different requirements of each user, the interaction and design are carried out with gestures.

The body-based method presented by Tejero-de-Pablos et al. [45], which focuses on the characteristics of the body joint based on these characteristics of human appearance, has little variability, and human movement is well represented. The method based on a neural network uses the actions of the players by combining holistic features and body joints so that the most important points can be extracted from the original video. Two types of action-related features are extracted and then classified as uninteresting or interesting. The personalized video summaries presented in [14,17,26,44,46,53,54,58,60,62,66,67] are also object-based.

4.2.2. Area Based

Area-based techniques are useful for detecting one or more areas, such as indoors or outdoors, as well as detecting specific scenes in that area, such as mountains, nature, skyline, etc. Lie and Hsu [14] proposed a video summary system that takes into account user preferences based on events, objects, and area. Through a user-friendly interface, the user can adjust the frame number or time limit. Kannan et al. [67] proposed a personalized video summarization system. The proposed system includes 25 detectors for semantic concepts based on objects, events, activities, and areas. Each captured video is assigned shot relevance scores according to the set of semantic concepts. Based on the time period and shots that are most relevant to the user's preferences, the video summary is generated. Tseng et al. [60] proposed a system consisting of three layers of server-middleware-client to produce a video summary according to user preferences based on objects, events, and scenes in the area. Many studies [53,54,62,66,68] have proposed that personalized video summarization is also based on area.

4.2.3. Perception Based

Perception-based video summaries refer to the ways, represented by high-level concepts, in which the content of a video can be perceived by the user. Some relevant examples include the association of the content of the video with some level of importance by the user, the size of the excitement that the user can have by watching the video, the amount that the user can perceive from the distraction videos, or the kind of emotions or the intensity that the user may perceive from the content. Other research areas apply theories such as semiotic theory, human perception theory, and user-attention theories as a means of interpreting the semantics of video content to achieve high-level abstraction. Therefore, unlike video summaries based on events and objects in which tangible events and objects present in the video content are identified, user perception-based summaries focus on extracting how the user perceives or can perceive the content of the video [11]. Hereafter, we present perception-based video summary methods.

Miniakhmetova and Zymbler [75] proposed a framework to generate a personalized video summarization that takes into account the user's evaluations of previously watched videos. The rating can be "like", "neutral", or "dislike". Scenes that affect the user of the video the most are collected, and their sequence forms the personalized video summary. A behavior pattern classifier presented by Varini et al. [62] is based on 3D convolutional neural networks (3D CNNs) and uses visual evaluation features and 3D motion. The selection

of elements is based on the degree of narrative, visual, and semantic perspective along with the user's preferences, as well as the user's attention behavior. Dong et al. [53,54] proposed a video summarization system based on the intentions of the user according to a set of predefined idioms and expected duration. Ul et al. [28] proposed a framework for summarizing movies based on people's emotional moments through facial expression recognition (FER). The emotions on which it is based are disgust, surprise, happy, sad, neutral, anger, and fear. Köprü and Erzin [32] proposed a human-centered video summary framework based on information derived from emotions and extracted from visual data. Initially, with the use of repetitive convolutional neural networks, emotional information is extracted, and then the emotional mechanisms of attention and information are expanded to enrich the video summary. Katti et al. [22] presented a semi-automated gaze-based method of the eye for emotional video analysis. The behavioral signal received from the user and entered is pupil dilation (PD) to assess user engagement and arousal. In addition to discovering regions of interest (ROIs) and emotional video segments, the method includes fusion with content-based features and gaze analysis. The Interest Meter (IM) was proposed by Peng et al. [27] to measure user interest through spontaneous reactions. By using a fuzzy fusion scheme, emotion and attention features are combined; this results in viewing behaviors being converted into quantitative interest scores, and a video summary is produced by combining those parts of the video that are considered interesting. Yoshitaka and Sawada [72] proposed a framework for video summary based on the observer's behavior while monitoring the content of the video. The observer's behavior is detected on the basis of the operation of the video remote control and eye movement. Olsen and Moon [40] proposed the DOI function to obtain a set of features of each video based on the interactive behavior of the viewers.

4.2.4. Motion and Color Based

Producing a video summary is difficult when it is based on motion and especially when the camera is involved [37]. Sukhwani and Kothari [44] proposed a framework for creating a personalized video summary in which the use of colors identifies football clips as event segments. To isolate the football events from the video, they modeled models of the player's activity and movement. The actions of the footballers are described using the dense characteristics descriptions in the trajectory descriptions. For the immediate identification of the players in the moving frames, they used the deep learning method for the identification of the player, and for the modeling of the football field, the Gaussian mixture model (GMM) method to achieve background removal. AVCutty was used in the work of Darabi et al. [35] to detect the boundaries of the scene and thus detect when a change in scene occurs through the motion and color features of the frames. The study by Varini et al. [62] is also motion-based.

4.2.5. Event Based

To detect abnormal and normal events presented in videos, event-based approaches are useful. There are many examples such as terrorism, mobile hijacking, robbery scenes, the recognition and monitoring of sudden changes in the environment, etc., in which the observation of anomalous/suspicious features is performed using detection models. To produce the video digest, the frames with abnormal scenes are joined using a summarization algorithm [37]. Hereafter, we present such event-based approaches.

Valdés and Martínez [17] described a method for creating video skims using an algorithm to dynamically build a skimming tree. They presented an implementation to obtain features that are different through online analysis. Park and Cho [26] presented a study in which a single sequence of events is generated from multiple sequences and the production of a personalized video summary is performed using fuzzy TSC rules. Chen et al. [39] described a study in which metadata are acquired based on the detection of events and objects in video. Taking metadata into account divides the video into segments so that each segment covers a self-contained period of a basketball game. Fei et al. [36]

proposed two methods for event detection. The first is detection by combining a support vector machine (CNNs-SVM) with a convolutional neural network, and the second is detection using an optimized summary network (SumNet). Lei et al. [61] proposed a method to produce a personalized video summarization without supervision based on interesting events. The unsupervised Multigraph Fusion (MGF) method was proposed by Ji et al. [71] to find events that are automatically relevant to the query. The LTC-SUM method was proposed by Mujtaba et al. [31] based on the design of a 2D CNN model to detect individual events through thumbnails. This model solves privacy and computation problems on end-user devices that are resource constrained. At the same time, due to this model, the efficiency of storage and communication is improved, as the computational complexity is reduced to a significant extent. PASSEV was developed by Hannon et al. [38] to create personalized real-time video summaries by detecting events using data from the Web. Chung et al. [50] proposed a PLSA-based model for video summarization based on user preferences and events. Chen and Vleeschouwer proposed a video segmentation method based on clock events [18]. More specifically, it is based on the rule that exists in basketball that each team has 24 s to attack by making a shot. Many events, such as fouls, shots, interceptions, and others, are immediately monitored with the restart, end, and start of the clock. Many studies [13,14,16,19,25,46,53,54,58,60,62,66,67] have proposed that personalized video summarization is also based on events.

4.2.6. Activity Based

Activity-based techniques focus on specific activities present in the video, such as lunch at the office, sunset at the beach, drinks after work, friends playing tennis, bowling, cooking, etc. Ghinea et al. [66] proposed a summarization system that can be used with any kind of video in any domain. From each keyframe, 25 semantic concepts are detected for categories of visual scenes that cover people, areas, settings, objects, events, and activities that are sufficient to represent a keyframe in a semantic space. From each video, relevance scores are assigned for a set of semantic concepts. The score shows the relevance between a particular semantic concept and a shot. Garcia [20] proposed a system that accepts from the user an image, text, or video query, then retrieves from the memories small subshots stored in the database, and produces the summary according to the user's activity preferences. The following studies [17,19,58,67] are also based on activities.

The distribution of the papers reported in Table 1 is shown in Figure 6 according to the type of features of the personalized summary. According to the distribution, most tasks perform a personalized object-based summarization at a percentage of around 39%. With a difference of 10% and a percentage of around 29%, those tasks follow in which the personalized video summary is event-based. Next, we have the tasks whose personalized summary is an area based on a percentage of around 10%. In fourth place are personalized perception-based and activity-based summaries with the same number of tasks and with a percentage of around 9%. The smallest number of works is the one with a percentage of around 4%, whose abstract is based on motion and color.

Table 1. Classification of personalized summarization techniques.

Paper	Type of Summary					Features						Domain		Time		Method		
	Storyboard	Video Skimming	Hierarchical	Multiview	Fast Forward	Object Based	Area Based	Perception Based	Motion and Color Based	Event Based	Activity Based	Domain Specific	Non-Domain Specific	Real Time	Static	Supervised	Weakly Supervised	Unsupervised
Tsenk et al., 2001 [12]		x				x							x	x				
Tseng and Smith, 2003 [13]			x			x				x			x	x		x		
Tseng et al., 2004 [60]		x				x	x			x			x	x		x		
Lie and Hsu, 2008 [14]		x				x	x			x			x		x	x		
Chen and Vleeschouwer, 2010 [18]		x		x						x		x		x				
Valdés and Martínez, 2010 [17]		x				x				x	x		x	x				
Chen et al., 2011 [39]		x		x		x				x	x	x		x				
Hannon et al., 2011 [38]		x								x		x		x				
Katti et al., 2011 [22]	x							x					x		x			
Park and Cho, 2011 [26]				x		x				x		x			x			x
Peng et al., 2011 [27]		x						x					x		x	x		
Yoshitaka and Sawada, 2012 [72]					x			x					x		x			
Kannan et al., 2013 [67]		x				x	x			x	x		x		x			
Niu et al., 2013 [46]	x					x				x			x	x		x		
Zhang et al., 2013 [16]	x					x				x		x			x			
Chung et al., 2014 [50]		x								x		x		x				x
Darabi et al., 2014 [35]	x								x				x		x			
Ghinea et al., 2014 [66]		x				x	x			x	x		x		x			
Kannan et al., 2015 [19]		x				x				x	x	x			x	x		
Garcia, 2016 [20]		x									x		x		x			
Sharghi et al., 2016 [48]	x	x				x							x	x		x		
del et al., 2017 [68]		x				x	x						x		x	x		
Otani et al., 2017 [74]		x				x							x		x			x
Sachan and Keshaveni, 2017 [8]			x			x						x			x			
Sukhwani and Kothari, 2017 [44]	x					x			x			x			x	x		
Varini et al., 2017 [62]		x				x	x	x	x	x		x			x	x		
Fei et al., 2018 [36]		x								x		x			x	x		
Tejero-de-Pablos et al., 2018 [45]		x				x						x			x	x		
Zhang et al., 2018 [64]		x				x							x		x	x		
Dong et al., 2019 [53]	x					x	x	x		x		x			x	x		
Dong et al., 2019 [54]	x					x	x	x		x		x			x	x		
Gunawardena et al., 2019 [59]	x					x							x		x			x
Jiang and Han, 2019 [69]		x				x							x		x	x		

Table 1. *Cont.*

Paper	Type of Summary						Features					Domain		Time		Method		
	Storyboard	Video Skimming	Hierarchical	Multiview	Fast Forward	Object Based	Area Based	Perception Based	Motion and Color Based	Event Based	Activity Based	Domain Specific	Non-Domain Specific	Real Time	Static	Supervised	Weakly Supervised	Unsupervised
Ji et al., 2019 [71]				x						x			x		x			x
Lei et al., 2019 [61]		x								x			x		x			x
Ul et al., 2019 [28]		x						x				x			x	x		
Zhang et al., 2019 [65]		x				x							x		x	x		
Baghel et al., 2020 [52]	x					x							x		x	x		
Xiao et al., 2020 [55]	x	x				x							x		x	x		
Fei et al., 2021 [30]		x				x							x		x	x		
Nagar et al., 2021 [25]		x				x				x			x		x			x
Narasimhan et al., 2021 [58]	x									x	x		x		x	x		x
Mujtaba et al., 2022 [31]		x								x			x	x		x		
Ul et al., 2022 [33]	x					x							x		x	x		
Cizmeciler et al., 2022 [70]		x				x							x		x		x	

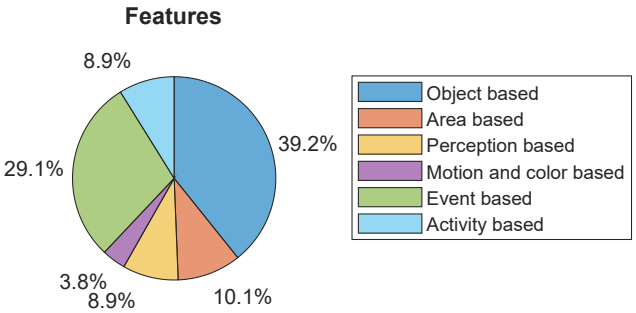

Figure 6. Distribution of the papersreported in Table 1 according to the type of features of the personalized summary.

4.3. Video Domain

The techniques can be divided into two categories: those whose analysis is domain specific and those whose analysis is not domain specific.

- In techniques that refer to *domain-specific* analysis, video summarization is performed in that domain. Common content areas include home videos, news, music, and sports. When performing an analysis of video content, blur levels are reduced when focusing on a specific domain [11]. The video summary must be unique to the domain. To produce a good summary in different domains, the criteria vary dramatically [54].
- In contrast to domain-specific techniques, *non-domain-specific* techniques perform video summarization for any domain, so there is no restriction on the choice of video to produce the summary. The system proposed by Kannan et al. [67] can generate a video summarization without relying on any specific domain. The summaries presented in [12–14,17,20,22,25,27,30,31,33,35,46,48,52,55,56,58–61,64–66,68–72,74] are not domain specific. The types of domains found in the literature are personal

videos [53,54], movies/serial clips [19,28,29,36,50], sports videos [18,36,38–45], cultural heritage [62], zoo videos [8], and office videos [26].

Table 2 provides a classification of works by domain type. Figure 7 presents the distribution of papers in the domain of personalized summary. In Figure 7, it can be seen that the large number of works produce personalized non-domain-specific summaries at a percentage of around 67%, in contrast to the works that produce personalized domain-specific summaries at a percentage of around 33%.

Table 2. Classification of domains.

Domain	Papers
Cultural heritage	[62]
Movies/serial clips	[19,28,29,36,50]
Office videos	[26]
Personal videos	[53,54]
Sports video	[18,36,38–45]
Zoo videos	[8]

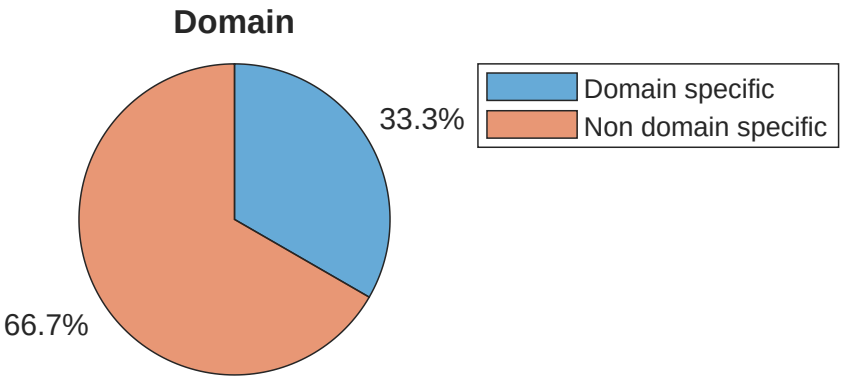

Figure 7. Distribution of the papers reported in Table 1 in the domain of personalized summary.

4.4. Source of Information

The video content life cycle includes three stages. The first stage is the capture stage, during which the recording or capturing of the video takes place. The second stage is the production stage; during this stage, a message or a story is conveyed through the format the video has been converted into after being edited. The third stage is the viewing stage, during which the video is shown to a targeted audience. Through these stages, the video content is desemanticized, and then the various audiovisual elements are extracted [11]. Based on the source of information, personalized video summarization techniques can be divided into three categories: internal personalized summarization techniques, external personalized summarization techniques, and hybrid personalized summarization techniques.

4.4.1. Internal Personalized Summarization Techniques

Internal personalized summarization techniques are the techniques that analyze and use information that is internal to the content of the video stream. During the second stage of the video life cycle and, more specifically, for the video content produced, internal summarization techniques apply. Through these techniques, low-level text, audio, and image features are automatically parsed from the video stream into abstract semantics suitable for video summarization [11].

- *Image features* may have changes in the motion, shape, texture, and color of objects that come from the video image stream. These changes can be used to perform video segmentation in shots by identifying fades or hot boundaries. Fades are identified

by slow changes in the characteristics of an image. Hot boundaries are identified by changes in an image's features in a sharp way, such as clipping. Specific objects can be detected, and an improvement in the depth of summarization can also be achieved for videos with a known structure by the analysis of image features. Sports videos are suitable for event detection because of their rigid structure. At the same time, event and object detection can also be achieved in other content areas that present a rigid structure, such as in news-related videos, as the start includes an overview of headlines, then a series of references is displayed, and finally the anchor face is the return [11].

- *Audio features* are related to the video stream and appear in the audio stream in different ways. Audio features include music, speech, silence, and sounds. Audio features can help identify segments that are candidates for inclusion in a video summary, and improving the depth of the summary can be achieved using domain-specific knowledge [11].
- *Text features* in the form of text captions or subtitles are displayed in the video. Instead of being a separate stream, text captions are "burned" or integrated into the video's image stream. Text may contain detailed information related to the content of the video stream and thus be an important source of information [11]. For example, in a football match broadcast live, captions showing the names of the teams, the score between them, the percentage of possession, the shots on target at that moment, etc., should appear during the match. As with video and audio features, events can also be identified from text. Otani et al. [74] proposed a text-based method. According to the text-based method, video blog posts use supporting texts that are used in the video summary at an earlier time. First, the video is segmented and then, according to the relevance of each segment to the input text, its priority is assigned to the summary video. Then, a subset of segments that have content similar to the content of the input text is selected. Therefore, based on the input text, a different video summary is produced.

4.4.2. External Personalized Summarization Techniques

External personalized summarization techniques analyze and use information that is external to the content of the video stream in the form of metadata. The life cycle of the video at each stage of its information is analyzed using external summarization techniques [11]. An external source of information is contextual information. Contextual information does not come from the video stream or the user and is additional [11]. The method presented by Katti et al. [22] is based on gaze analysis by combining the features of the video content to discover the regions of interest (ROIs) and emotional segments. First, eye tracking is performed to record pupil dilation and eye movement information. Then, after each peak pupil dilation stimulation, a determination of the first fixation is made, and the corresponding video frames are marked as keyframes. Linking keyframes creates a storyboard sequence.

4.4.3. Hybrid Personalized Summarization Techniques

Hybrid personalized summarization techniques analyze and use information that is both external and internal to the content of the video stream. From the life cycle of the video, at each stage, its information is analyzed by external summarization techniques. Any combination of outer and inner summarization techniques can form hybrid summarization techniques. Each approach tries to capitalize on its strengths while minimizing its weaknesses, to make video summaries as effective as possible [11]. Combining text metadata with the capabilities of image-level video frames can help improve summary performance [47]. Furthermore, for non-domain-specific techniques, hybrid approaches have proven useful [11].

4.5. Based on Time of Summarization

The techniques can be divided into two categories depending on whether the personalized summary is conducted live or on a pre-recorded video, respectively. The first category is real time, and the second category is static. Both are presented below.

- In *real-time* techniques, the production of the personalized summary takes place during the playback of the video stream. Due to the fact that the output should be delivered very quickly, it is a difficult process to produce in real time. In real-time systems, an output that is delayed is incorrect [47]. The Sequential and Hierarchical Determinant Point Process (SH-DPP) is a probabilistic model developed by Sharghi et al. [48] to be able to produce extractive summaries from streaming video or long-form video. The personalized video summaries presented in [12,13,17,18,24,31,38,39,46,50,60] are in real time.
- In *static* techniques, the production of the personalized summary takes place on a recorded video. Most studies are static in time [8,14,16,19,20,22,25–28,33,35,36,44,45,52–56,58,59,61,62,64–72,74].

Figure 8 presents the distribution of the papers, reported in Table 1, between the time of the personalized summary. This distribution dominates the number of papers whose personalized summaries are static, at a percentage of around 76%, as opposed to the number of papers whose summaries are in real time, at a percentage of around 24%.

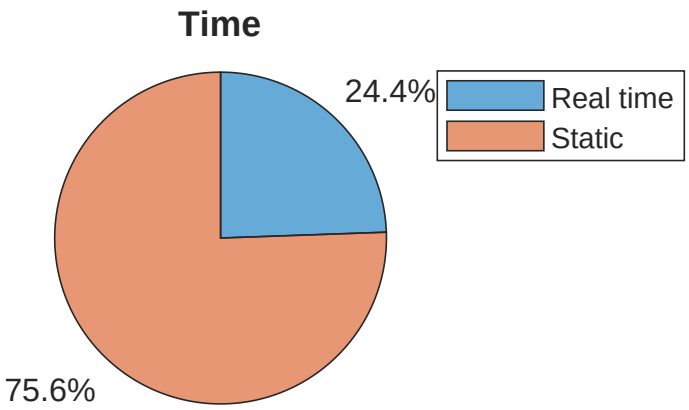

Figure 8. Distribution of the papers reported in Table 1 between the time of the personalized summary.

4.6. Based on Machine Learning

Techniques related to machine learning are developed to identify objects, areas, events, etc. Various machine learning techniques apply algorithms to generate personalized video summaries. Based on the algorithm applied, the techniques are divided into supervised, weakly supervised, and unsupervised categories. The methods in the above categories use convolutional neural networks (CNNs).

In large-scale video and image recognition, CNNs have been very successful. The learning of high-level features in a progressive manner and the obtaining of the original image with the best representation are supported by CNNs [36]. Features for the recognition of holistic action that are more generalized and reliable than hand-made characteristics can be extracted from CNN. Therefore, CNN has overcome traditional methods [45]. The FasterRCNN model was proposed by Nagar et al. [25] for face detection. Varini et al. [62] proposed a 3D CNN with 10 layers. The layers are trained in frame features and visual motion evaluation. The way for unsupervised and supervised summarization techniques has been paved by the success of deep neural networks (DNNs) in learning video representations and complex frames [25]. In neural networks, memory networks are used to flexibly model the attentional scheme. In addition, to deal with visual answers to questions and answers to questions, memory networks are used [5].

Through deep learning based on artificial neural networks, computers learn to process data as a human would. By using many data and through training models, they learn the

characteristics that are their own, with room for optimization. On the basis of the algorithm applied, the techniques are divided into the following categories.

4.6.1. Supervised

In supervised techniques, a model is first trained using labeled data, and then the video summary is produced based on that model. Hereafter, we present supervised approaches. Deep learning for player detection was proposed by Sukhwani and Kothari [44]. To address the multiscale matching problem in the person search, Zhou et al. [29] used the Cross-Level Semantic Alignment (CLSA) model. From the identity features, the most discriminative representations are learned using the end-to-end CLSA model. The DeepSORT algorithm was used by Namitha et al. [73] to track objects. Huang and Worring [21] proposed a deep learning method to generate query-based video summarization for a visual text embedding space. The method is end-to-end and consists of a video digest generator, a video digest controller, and a video digest output module. OoI detection was performed using YOLOv3 by Ul et al. [33]. A single neural network was applied to the full video. The frames are divided into regions, delineated, and the neural network predicts probabilities. A deep architecture was trained by Choi et al. [57] to perform efficient learning of the semantic embeddings of video frames. The learning is conducted through the progressive exploitation of the data from the captions of the images. According to the implemented algorithm, the semantically relevant segments of the original video stream are selected according to the context of the sentence or text description provided by the user. Studies [5,7,13,14,19,24,27,32,33,36,44,48,60,62] used the supervised technique to produce a summary.

Active Video Summary (AVS) was suggested by Garcia del Molino et al. [68], which constantly asks the user questions to obtain the user's preferences about the video, updating the summary online. From each extracted frame, object recognition is performed using a neural network. Using a line search algorithm, the summary is generated. Fei et al. [30] introduced a video summarization framework that employs a deep ranking model to analyze large-scale collections of Flickr images. This approach identifies user interests to autonomously create significant video summaries. This framework does not simply use hand-crafted features to calculate the similarity between a set of web images and a frame.

Tejero-de-Pablos et al. [45] proposed a method that uses two separate neural networks to transform the joint positions of the human body and the RGB frames of the video stream as input types. To identify the highlights, the two streams are merged into a single representation. The network is trained using the UGSV dataset from the lower to the upper layers. Depending on the type of summary, there are many ways that the personalized summary process can be modeled.

- *Keyframe selection:* The goal is to identify the most inclusive and varied content (or frames) in a video for brief summaries. Keyframes are used to represent significant information included in the video [52]. The keyframe-based approach chooses a limited set of image sequences from the original video to provide an approximate visual representation [31]. Baghel et al. [52] proposed a method in which user preference is entered as an image query. The method is based on object recognition with automatic keyframe recognition. From the input video, the important frame is selected, so that the output video is produced from these frames. Based on the similarity score between the input video and the query image, a keyframe is selected. A selection table is created from the keyframe that is decided to be selected. A threshold is applied to the selection score. If the frame has a selection score greater than the threshold value, then this frame is a keyframe; otherwise, the frame is not considered a keyframe and is discarded.
- *Keyshot selection:* The keyshots comprise standard continuous video segments extracted from full-length video, each of which is shorter than the original video. Keyshot-driven video summarization techniques are used to generate excerpts from short videos (such as user-created TikToks and news) or long videos (such as full-length movies and

soccer games) [31]. Mujtaba et al. [31] presented the LTC-SUM method to produce personalized keyshot summaries that minimize the distance between the semantic information of the side and the selected video frame. Using a supervised encoder–decoder network, the importance of the frame sequence is measured. Zhang et al. [65] proposed a mapping network (MapNet) to express the degree of association of a shot with a given query, to create a visual information mapping in the query space. Using deep reinforcement learning (SummNet), they proposed to build a summarization network to integrate diversity, representativeness, and relevance to produce personalized video summaries. Jiang and Han proposed a scoring mechanism [69]. In the hierarchical structure, the mechanism is based on the scene layer and the shooting layer and receives the output. Each shot is scored through this mechanism, and as basic shots, the shots are selected as high-rated shots.

- *Event-based selection:* The process of personalized summarization detects events from a video based on the user's preferences. In the above method [31], to identify thumbnail events that are not specific domains, a two-dimensional convolutional neural network (2D CNN) model was implemented.
- *Learning shot-level features:* It involves learning advanced semantic information from a video segment. The Convolutional Hierarchical Attention Network (CHAN) method was proposed by Xiao et al. [55]. After dividing the video into segments, visual features are extracted using the pre-trained network. To perform shot-level feature learning, visual features are sent to the feature encoding network. To perform learning on a high-level semantic information video segment, they proposed a local self-attention module. To manage the semantic relationship between the given query and all segments, they used a global attention module that responds to the query. To reduce the length of the shot sequence and the dimension of the visual feature, they used a fully convolutional block.
- *Transfer learning:* Transfer learning involves adjusting the information gained from one area (source domain) to address challenges in a separate, yet connected area (target domain). The concept of transfer learning is rooted in the idea that when tackling a problem, we generally rely on the knowledge and experience we have gained from addressing similar issues in the past [76]. Ul et al. [28] proposed a framework using transfer learning to perform facial expression recognition (FER). More specifically, they presented the learning process that, to be completed, includes two steps. In the first step, a CNN model is trained for face recognition. In the second step, transfer learning is performed for the FER of the same model.
- *Adversarial learning*. This is a technique employed in the field of machine learning to trick or confuse a model by introducing harmful input, which can be used to carry out an attack or cause a malfunction in a machine learning system. A competitive three-player network was proposed by Zhang et al. [64]. The content of the video, as well as the representation of the user query, is learned from the generator. The parser receives three pairs of digests based on the query so that the parser can distinguish the real digest from a random one and a generated one. To train the classifier and the generator, a lossy input of three players is performed. Training avoids the generation of random summaries which are trivial, as the summary results are better learned by the generator.
- *Vision-language:* A vision language model is an artificial intelligence model that integrates natural language and computer vision processing abilities to comprehend and produce textual descriptions of images, thus connecting visual information with natural language explanations. Plummer et al. [77] used a two-branch network to learn the integration model of the vision language. Of the two branches of the network, one receives the text features and the other the visual features. The triple loss based on the margin trains the network by combining a neighborhood-preserving term and two-way ranking terms.

- *Hierarchical self-attentive network:* The hierarchical self-attentive network (HSAN) is able to understand the consistent connection between video content and its associated semantic data at both the frame and segment levels. This ability enables the generation of a comprehensive video summary [78]. A hierarchical self-attentive network was presented by Xiao et al. [78]. First, the original video is divided into segments, and then, using a pre-trained deep convolutional network, the visual feature is extracted from each frame. To record the semantic relationship at the section level and at the context level, a global and a local self-care module are proposed. To learn the relationship between visual content and caption, the self-attention results are sent to a caption generator, which is enhanced. An importance score is generated for each frame or segment to produce the video summary.
- *Weight learning:* The weight learning approach was proposed by Dong et al. [54], in which using maximum-margin learning, it can automatically learn the weights of different objects. Learning can occur for processing styles that are not the same or different types of product, as these videos contain annotations that are highly relevant to the domain expert's decisions. For different processing styles or product categories, there may be different weightings of audio annotations built directly with domain-specific processing decisions. For efficient user exploration of the design space, there may be default storage of these weights.
- *Gaussian mixture model:* A Gaussian mixture model is a clustering method used to estimate the likelihood that a specific data point is part of a cluster. In the user preference learning algorithm proposed by Niu et al. [46], the most representative are initially selected as temporary keyframes from the extracted frames. To indicate a scene change, temporary frames are displayed to the user. If the user is not satisfied with the selected temporary keyframes, they can interact by manually selecting the keyframes. A Gaussian mixture model (GMM) is modeled by learning user preferences. The parameters of the GMM are automatically updated based on the user's manual selection of keyframes. Production of the personalized summary is performed in real time as the personalized frames update the selected frames from the temporary base. Personalized keyframes represent user preferences and taste.

4.6.2. Unsupervised

In unsupervised techniques, clusters of frames are first created based on the quality of their content, and then the video summary is created by concatenating the keyframes of each cluster in chronological order. An unsupervised method, called FrameRank, was introduced by Lei et al. [61]. They constructed a graph where frame similarity is measured at the edges and video frames correspond to vertices. To measure the relative importance of each segment and each video frame, they applied a graph-ranking technique. Depending on the type of summary, there are many ways in which the personalized summary process can be modeled, which are described hereafter.

- *Contrastive learning:* Using a pretext, self-supervised pretraining of a model can be performed, which is the approach to contrastive learning. According to contrastive learning, the model learns to repel representations intended to be far away, called negative representations, and to attract them from positive representations intended to be close to discriminate between different objects [56].
- *Reinforcement learning:* A framework for creating an unsupervised personalized video summary that supports the integration of diverse pre-existing preferences, along with dynamic user interaction for selecting or omitting specific content types, was proposed by Nagar et al. [25]. Initially, the egocentric video captures spatio-temporal features using 3D convolutional neural networks (3D CNNs). Then, the video is split into non-overlapping frame shots, and the features are extracted. Subsequently, the features are imported by the reinforcement learning agent, which employs a bi-directional long- and short-term memory network (BiLSTM). Using forward and backward flow, BiLSTM serves to encapsulate future and past information from each subshot.

- *Event-based keyframe selection (EKP):* EKP was developed by Ji et al. [71] so that keyframes can be presented in groups. The separation of groups is based on specific facts that are relevant to the query. The Multigraph Fusion (MGF) method is implemented to automatically find events that are relevant to the query. The keyframes in the different event categories are then separated from the correspondence between the videos and the keyframes. Through the two-level structure, the summarization is represented. Event descriptions are the first layer, and keyframes are the second layer.
- *Fuzzy rule based:* To represent human knowledge, which includes fuzziness and uncertainty, a method is a fuzzy system. From the theory of fuzzy sets, the fuzzy system is a representative and important application [26]. Park and Cho [26] used a system based on fuzzy TSK rules. This system was used to evaluate video event shots. Also, in this rule-based system, consistency is a function of the variables used as input and not a linguistic variable, and therefore, the time-consuming decomposition process can be avoided. The summaries in [50,51,56,59,74] use an unsupervised technique to produce a summary.

4.6.3. Supervised and Unsupervised

Narasimhan et al. [58] presented a model that can be trained with and without supervision and that belongs to the supervised and unsupervised categories. The supervised setting uses reconstruction, diversity, and classification as loss functions, whereas the unsupervised setting uses reconstruction and diversity as loss functions.

4.6.4. Weakly Supervised

Weakly supervised video summary methods use less expensive labels without using basic truth data. The labels they use are imperfect compared to tags that are complete in human annotations. However, they can lead to an effective training of summary models [79]. Compared to the supervised video summary approach, a weakly supervised summary approach needs a smaller set of training to carry out the video summary. Cizmeciler et al. [70] suggested a personalized video summary approach with weak supervision. Weak supervision is carried out as semantic maps of reliability. Through predictions from pre-trained classifiers of actions/characteristics, semantic maps are obtained.

The distribution of the papers reported in Table 1, according to the type of personalized summary method, is shown in Figure 9. It is evident from the distribution that in the largest number of papers, the proposed method of producing a personalized video summary is supervised, at a percentage of around 73%. Second, the percentage of unsupervised personalized video summarization methods is around 24%. In the last place, and with a percentage of around 3%, is the weakly supervised methods.

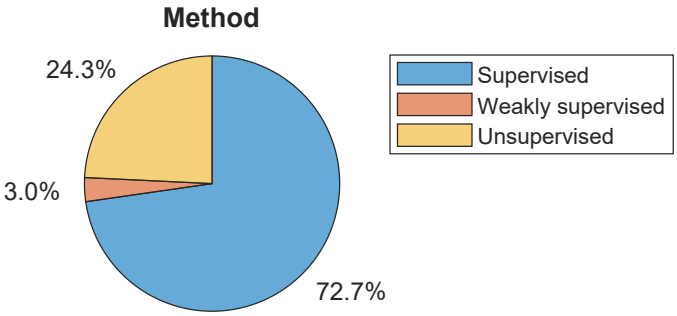

Figure 9. Distribution of the papers reported in Table 1 among types of personalized summary method.

5. Types of Methodology

According to the type of methodology used by personalized video summary techniques for the summary production process, the techniques are classified into five broad categories.

5.1. Feature-Based Video Summarization

Based on speech transcription, audiovisual material, object, dynamic content, gestures, color, and movement, categorization of summarization techniques can be carried out [49]. Gunawardena et al. [59] proposed the Video Shot Sampling Module (VSSM) to initially divide the video into shots consisting of similar visual content to conquer and divide each shot separately. The proposed method is based on hierarchical clustering but also on changing the color in the histogram to perform boundary detection in the video shot. In the framework developed by Lie and Hsu [14], an analysis of the colors and intensities of the flame is performed in each frame to detect the explosion frames.

5.2. Keyframe Selection

The primary versions of the video summarization systems fundamentally employed frame categorization methods that relied on the labeled data inputted into them. Features drawn from video frames are used as input for an algorithm that categorizes the frames as being of interest or not being of interest. Only frames of interest are incorporated into the summary. The description and detection of important criteria is a challenge and is used to select the appropriate keyframes. The techniques for keyframe selection are presented below.

- *Object-based keyframe selection:* Summaries that are based on objects specifically concentrate on certain objects within a video. These are scrutinized and used as benchmarks to identify the video keyframes. The techniques in the following studies [13,16,26,46,48,52,58] are object-based keyframe selection.
- *Event-based keyframe selection:* Event-oriented summaries pay special attention to particular events in a video, which are examined and used as benchmarks for the summarization process. The keyframes of the video are identified based on the selected event. The personalized video summary technique in [71] is the selection of event-based keyframes.
- *Perception-based keyframe selection:* The detection of the important criterion of user perception is a factor in keyframe selection. In studies [22,32], the keyframes are detected based on the user's perception.

5.3. Keyshot Selection

A video shot is a segment of video that is captured in continuous time through a camera. The keyframes are selected after the video has been split into shots. Another challenge arises when one shot is selected by the keyframes [49]. Identifying and defining crucial factors in this method is a hurdle and is used to choose the appropriate keyshots. Subsequently, methods for keyshot selection are introduced.

- *Activity-based shot selection:* The activity-based summaries focus particularly on specific activities of a video, which are analyzed and considered criteria for detecting keyshots of the video. In [20,66], the selection of the keyshot is based on activity.
- *Object-based shot selection:* The summaries that are object based place special emphasis on distinct objects within a video, which are examined and used as benchmarks for the summarization process. Based on the selected object, the keyshots of the video are detected. In the following studies [12,19,60,62,66,67,69,70], the keyshot selection technique is based on object detection.
- *Event-based shot selection:* Video keyshots are identified based on a particular event that the user finds interesting. For the personalized summaries in [31,62], the selection of keyshots are based on events.
- *Area-based shot selection:* Area-based summaries focus particularly on specific regions within a video, which are scrutinized and employed as reference points for the summarization procedure. Depending on the chosen region, the video keyshots are identified. For the personalized summaries presented in [62,66], the selection of keyshots is based on the area.

5.4. Video Summarization Using Trajectory Analysis

Chen et al. [39] introduced a method that identifies 2D candidates from each separate view, subsequently calculating 3D ball positions through triangulation and confirmation of these 2D candidates. After identifying viable 3D ball candidates, examining their trajectories aids in distinguishing true positives from false positives, as the ball is expected to adhere to a ballistic trajectory, unlike most false detections (typically related to body parts). Chen et al. [41] proposed a method in which the video surveillance trajectory is extracted to help identify group interactions.

Dense Trajectories

In order to describe in space and time the actions of soccer players, Sukhwani and Kothari used dense feature trajectory descriptors [44]. Trajectories are extracted from several spatial scales. Samples are taken from the area of a player, which is in the form of a grid, and monitored separately at each scale. The motion of the camera is corrected for by calculating features, which is a default property. Feature points that do not belong to player areas in event representations are suppressed and cut to avoid inconsistencies. Noise is introduced through the non-player points, as they are not representations of actual player actions. To record the player's movements, the HOF is calculated along the entire length of the dense trajectories. Depending on the angle, each flow vector subtends with the horizontal axis, and depending on the magnitude of the vector, it is weighted. Using full orientation, quantization is performed in eight bins of HOF orientations.

5.5. Personalized Video Summarization Using Clustering

The text method proposed by Otani et al. [74] first extracts the nouns from the input text and then the videos are clustered into groups based on each event after being segmented. The priority of each section is calculated according to the clusters that have been created. Segments that are high priority have a higher chance of being included in the video summary. The video summary is then generated by selecting the subset of segments that is optimal according to the calculated priority calculation.

Based on clustering, the categories into which the personalized video summary can be divided are hierarchical clustering, aggregation hierarchical clustering, hierarchical context clustering, k-means clustering, dense-neighbor-based clustering, concept clustering and affinity propagation.

- *Hierarchical clustering:* Hierarchical clustering, or hierarchical cluster analysis, is a method that aggregates similar items into collections known as clusters. The final result is a series of clusters, each unique from the others, and the elements within each cluster share a high degree of similarity. Yin et al. [51] proposed a hierarchy to encode visual features. The hierarchy will be used as a dictionary to encode visual features in an improved way. Initially, a cluster of leaf nodes is created hierarchically based on their paired similarities. The same process is then retrospectively performed to cluster the images into subgroups. The result of this process is the creation of a semantic tree from the root to the leaves.
- *K-means clustering:* The Non-negative Matrix Factorization (NMF) method was proposed by Liu et al. [63] to produce supervised personalized video summarization. The video is segmented into small clips, and each clip is described by a word from the bag-of-words model. The NMF method is used for action segmentation and clustering. Using the k-means algorithm, the dictionary is learned. The unsupervised method proposed by Ji et al. [71] uses a k-means algorithm to cluster words containing similar meanings. In the method proposed by Cizmeciler et al. [70], clustering is performed using the k-means algorithm. From the frames, the extraction of visual characteristics is carried out using CNN, which is pre-trained. Each shot is represented by the average of visual characteristics. With the k-means method, the shots are clustered into a set of clusters. The centers are then initialized using the Euclidean distance metric by taking random samples from the data. When the shots closest to each center of the cluster are

selected, the summary is created. To cluster photos, Darabi and Ghinea [7] used the K-means algorithm in their supervised framework to categorize them based on RGB color histograms.

- *Dense-neighbor based clustering:* Lei et al. [61] proposed an unsupervised framework to produce a personalized video summary. With a clustering method, the video is divided into separate segments. The method is based on dense neighbors. They clustered the video frames into discrete segments that are semantically consistent frames using a clustering center clustering algorithm.
- *Concept clustering:* Sets of video frames that are different but semantically close to each other are grouped together. Clusters of concepts are formed from this grouping. For the representation of macro-optical concepts, which are created from similar types of micro-optical concepts of video units that are different, these clusters are intended. Using the cumulative clustering technique, sets of frames are clustered. The summation technique is top–down and is based on the pairwise similarity function (SimFS) operating on two sets of frames [8].
- *Affinity propagation:* It is a clustering approach that has a higher fitness value than other approaches. AP is based on factor graphs and has been used as a tool to create storyboard summaries and cluster video frames [15].

6. Personalized Video Datasets

Datasets are needed to test, train, and benchmark various personalized video summarization techniques. Figure 10 depicts personalized video datasets (x-axis) and the number of corresponding articles (y-axis) that report results on each of them. The following datasets have been created to evaluate the summarization of personalized videos.

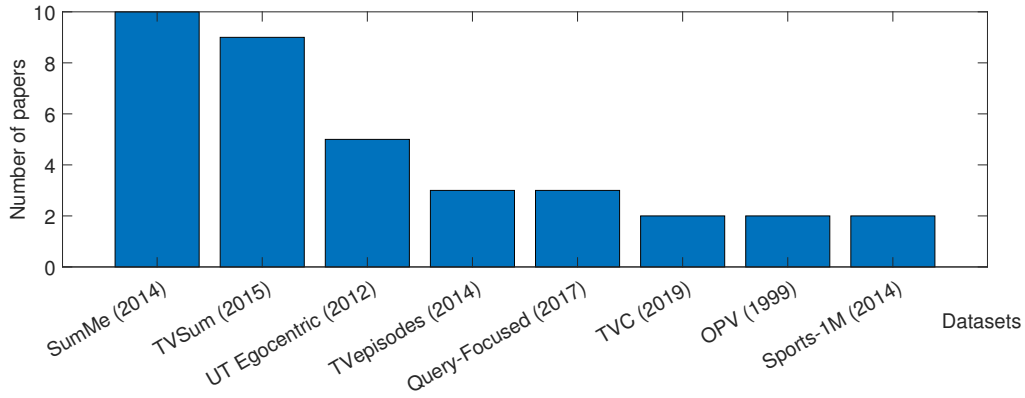

Figure 10. Personalized video datasets (x-axis) and the number of corresponding papers (y-axis) that report results on each of them (SumMe [80], TVSum [81], UT Egocentric [82], TV episodes [83], Query-Focused [64], TVC [53], OVP [84], Sports-1M [85]).

1. *Virtual Surveillance Dataset for Video Summary (VSSum):* According to the permutation and combination method, Zhang et al. [86] created a dataset in 2022. The dataset consists of virtual surveillance videos and includes 1000 videos that are simulated in virtual scenarios. Each video has a resolution of 1920 × 1080 in 30 fps and a duration of 5 min.
2. *The Open Video Project (OVP):* This dataset was created in 1999, including 50 videos recorded at a resolution of 352 × 240 pixels at 30 fps, which are in MPEG-1 format and last from 1 to 4 min. The types of content are educational, lecture, historical, ephemeral, and documentary [87]. This repository contains videos longer than 40 h, mostly in MPEG-1 format, pulled from American agencies such as NASA and the National Archives. A wide range of video types and features are reflected in the repository content, such as home, entertainment, and news videos [84].
3. *TVC:* In 2019, Dong et al. [53] created the TVC dataset which consists of 618 TV commercial videos.

4. *Art City:* The largest unconstrained dataset of heritage tour visits, including egocentric videos sharing narrative landmark information, behavior classification, and geolocation information, was published by Varini et al. [62] in 2017. The videos have recorded cultural visits to six art cities by tourists and contain geolocation information.
5. *Self-compiled:* Based on the UT Egocentric (UTE) dataset [82], a new dataset was compiled after collecting dense concept annotations on each video capture by Sharghi et al. [5] in 2017. The new dataset is designed for query-focused summarization to provide different summarizations depending on the query and provide automatic and efficient evaluation metrics.
6. *Common Film Action:* This repository includes 475 video clips from 32 Hollywood films. The video clip that contains it is from eight categories of actions that have been annotated [88].
7. *Soccer video:* This dataset contains 50 soccer video clips whose annotation was manually collected from the UEFA Champions League and the FIFA World Cup 2010 championship [36].
8. *UGSV Kendo:* This dataset was created in 2016. It contains 10 RGB-D videos containing 12 combats totaling 90 min. Videos were recorded using a Microsoft Kinect V2 sensor [89]. Eighteen additional self-recorded RGB-D Kendo videos were added to the dataset. The videos have a total duration of 246 min and a frame of around 20 fps [45].
9. *Hollywood-2:* It includes 2859 videos with a resolution of 400–300 × 300–200 pixels from 69 films, some of which are educational and others test. For testing, it is 884 actions and for training, it is 823 actions. For testing, it is 582 scenes, and for training, it is 570 scenes [90].
10. *Sports-1M:* It was published in 2014. This dataset includes 1 million YouTube sports videos that have been annotated in 487 classes. Each class includes 1000–3000 videos, and five percent of the videos have been annotated with at least one class. On average, each video is 5 min and 36 s long [85].
11. *Weizman dataset:* This dataset includes 90 video sequences which have a resolution of 180 × 144 pixels and are de-interlaced at 50 fps. The videos show for each of the 9 people the 10 physical actions they perform [63].
12. *UGSum52:* This dataset was created in 2019. It contains 52 videos, one to nine minutes long, that have been recorded in various ways such as egocentric, dynamic, and static views. Each video contains 25 summaries of people. All videos are minimally edited [61].
13. *UT Egocentric (UTE):* It was published in 2012. Using the portable Looxcie camera, 10 videos have been recorded with a resolution of 320 × 480 pixels and 15 fps each. Each video has a duration of 3 to 5 h, and the total duration of all videos is 37 h. Videos capture various physical daily activities. It is based on the most important people and the most important objects. It is used to produce first-person view-based video summaries. Object annotation is included in the video data [82].
14. *SumMe:* This dataset was created in 2014. It includes 25 videos that are slightly processed or raw. The content of the videos is related to sports, events, and holidays. The duration of each video varies from 1 to 6 min. The dataset was annotated for video segments with human scores [80].
15. *TVSum:* It was published in 2015. This dataset includes a collection of 50 YouTube videos. Videos are divided into 10 categories. Each category contains 5 videos, and each video represents a specific genre such as documentaries, news, and egocentrics. The duration of each video is 2 to 10 min. Through crowd sourcing, video scores are annotated at the level of shots [81].
16. *YouTube Action:* This dataset [91] was published in 2009. This dataset includes, from 11 categories of realistic action, 1581 videos [15].
17. *UTEgo:* It was created in 2012. Using the Looxcie Wearable Camera, the 10 videos were collected. Each video has a resolution of 320 × 480 pixels at 15 fps that lasts 3 to 5 h. Various activities are recorded in the videos, such as cooking, driving, lecture

tracking, shopping, and food. As ground truth, the annotations taken in Mturk were used. Each video has annotated frames that average 680 [82].

18. *CSumm:* This dataset was published in 2017. Using a Google glass, 10 videos were captured with a 720 × 1280 pixel resolution at 29 fps. Each video is annotated, lasts between 15 and 30 min, and is non-restrictive. Some of the activities included in the videos are having dinner, enjoying nature, watching sports, exercising, etc. A wide range of motions and viewpoints are included in the videos. The best segments are manually annotated with a 10 s time limit [68].
19. *Query-Focused dataset (QFVS):* It was published in 2017. The repository includes 4 videos with different scenarios of everyday life that are uncontrollable. Each video lasts between 3 and 5 h. For each user's query, a dictionary of 48 concepts related to everyday life is provided in order to make a video summary based on queries [64]. Users annotate in each video the absence/presence of concepts [5].
20. *Animation:* From 32 cartoons, the dataset includes 148 videos. Each video includes 3000 bullet screen comments and lasts about 24 min [29].
21. *Movie:* The dataset includes 19 movies, 2179 bounding boxes, and 96 characters. The sample tests consist of 452 bounding boxes and 20 characters [29].
22. *Recola:* This dataset is used to identify emotions using multiple models [92]. The Recola dataset was used by Köprü and Erzin to evaluate and educate Cer-Net [32].
23. *TV episodes:* This dataset was published in 2014. It includes ground truth summaries and text annotations for TV episodes. The number of episodes is 4 and the duration of each episode is 45 min. The total duration of the annotations is 40 h of data. In 11 videos, the 40 h separation takes place [83].
24. *Large-scale Flickr images:* This dataset was created in 2021. It includes 420,000 athletic images that were shared by photographers. Images are divided into seven categories, and 30,000 queried photos are taken in response to each category. Pictures of "Interest" are the photos that appear after each question. To train the improved deep model, images of "no interest" and "interest" are used [30].
25. *ActivityNet* [93]*:* It was published in 2015. This dataset includes 20,000 videos and 100k descriptions. The total duration of all videos is 849 h. Each video has an average duration of 180 s [78].
26. *Query–video pair based:* This dataset was published in 2020. The repository includes 190 YouTube videos with annotations for each video. All videos are sampled with one frame per second (FPS). Annotations are frame-based. Amazon Mechanical Turk (AMT) was used to produce frame annotations with the labels of the question scores that are relevant and text-based [21].
27. *MVS1K:* It was published in 2019. This dataset includes 10 queries of around 1000 videos. Videos are crawled by YouTube, and web images that are relevant, annotations that are manual, and video tags. As queries have been selected from Wikipedia News, the list includes 10 hot events from 2011 to 2016. For each query, about 100 videos were collected from YouTube. Each video lasts between 0 and 4 min. Annotation was applied to videos of human judgments [71].
28. *TRECVID 2001:* This dataset was created in 2001. It includes TV and documentaries videos. Videos are short, that is, less than 2 min, or long, that is, longer than 15 min [59].
29. *INRIA Holiday:* This dataset includes mostly high-resolution personal vacation photos. A very wide variety of scene types are included in the dataset. Each individual scene is represented by one of the 500 groups of images contained in the dataset [94].
30. *ImageNet:* This dataset was published in 2009. It includes 5247 categories with 3.2 million annotated images in total [95].
31. *UCF101:* This dataset was published in 2012. It contains more than 13K clips, which are downloaded from YouTube. The videos are of type AVI, have a resolution of 320 × 240 and a frame rate of 25 fps. In the dataset, videos are divided into 5 types and include 101 action categories [96].

32. *Relevance and Diversity Dataset (RAD):* This dataset was created in 2017. It is a query-specific dataset. It includes 200 annotated videos with query-specific and diversity-relevant labels. Given a different question, the retrieval of each video was performed. The top queries were pulled between 2008 and 2016 from YouTube with seed queries in 22 different categories. Each video is between 2 to 3 min long and is sampled at one frame per second [97].
33. *FineGym:* It was published in 2020. This dataset includes 156 high-resolution videos (720p and 1080p) from a YouTube gymnasium of 10 min each. The annotations provided in FINEGYM are related to the fine-grained recognition of human action in gymnastics [98].
34. *Disney:* This dataset was published in 2012. Videos were collected by more than 25 people using a GoPro camera mounted on the head. The videos have resolutions of 1280×720 and were sampled at 30 FPS. The dataset includes 8 topics and the total duration of the videos is more than 42 h. At 15 fps, the images were extracted, and in total more than 2 million images were exported. Throughout the video, the end and start times of the intervals corresponding to the types of social interaction were manually labeled [99].
35. *HUJI:* This dataset [100] was created in 2016. This repository contains 44 self-centered videos of daily activities, performed both externally and internally, by 3 people. Each video is less than 30 min long [25].
36. *New YouTube:* Includes 25 videos in 5 categories, such as food, urban placement, natural scene, animals, and landmark [51].
37. *YouTube Highlights:* This dataset was published in 2014. It includes YouTube videos in six different fields, such as skiing, surfing, parkour, dogs, gymnastics, and skating. Each section contains about 100 videos of varying lengths. The total duration of all videos is equal to 1430 min. The dataset has been labeled using Amazon Mechanical Turk for evaluation and analysis purposes [101].

Table 3 provides some datasets with their characteristics that were used for personalized video summarization.

Table 3. Video datasets used for the personalized summarization task.

Dataset Name	Year	Short Description	Num. Vids	Length per Video (Min.)
OVP [84]	1999	Home, entertainment and news videos	50	[1, 4]
Sports-1M [85]	2014	YouTube sports videos	10^6	5.6 (average)
UTE [82]	2012	Egocentric videos capture Physical daily activities	10	[180, 300]
SumMe [80]	2014	Sport, event and holiday videos	25	[1, 6]
TVSum [81]	2015	Documentaries, news and egocentric videos	50	[2, 10]
FineGym [98]	2020	YouTube gymnasium videos	156	10

7. Evaluation

To evaluate the performance of the video summary produced and the performance of the summary system, the video summary technique is evaluated. It is not easy to evaluate both the quantitative (objective) and qualitative (subjective) performance of the video summary because of the lack of standard measurements. The evaluation process becomes even more difficult due to the lack of a sufficient number of videos and annotations in conjunction with the subjectivity involved in qualitative evaluation [49].

7.1. Characteristics of Well-Crafted Summary

Summarizing videos is a highly subjective endeavor since a summary created manually reflects the individual preferences of the annotator or evaluator. However, it is possible to identify specific characteristics of a well-crafted summary that can be considered essential

requirements [56]. The following characteristics can provide valuable information for the development of a successful video summarization model.

- *Representativeness:* The summary should effectively convey the main points of the original video. Although some information may be lost in the process of summarizing, it is essential to create a concise summary that clearly distinguishes the original video from others. Each video consists of various sub-videos, each presenting a distinct storyline. Therefore, the ideal summary is a blend of sub-videos that closely represents all narratives within the original video while remaining distinct from sub-videos in other videos [56].
- *Sparsity:* Original videos could come from various sources, including video news, a video blog, long online streams, or a television program. In all instances, the preferred synopsis should be brief, typically lasting only a few seconds, as that is the typical duration a viewer will invest before determining whether to engage further or move on. Although, for concise videos, the ideal summary might be approximately 15% of the total duration, this proportion may decrease considerably for longer videos [56].
- *Diversity:* Another essential characteristic of a video summary is the diversity present in its frames. While it may be feasible to generate a summary for certain videos and datasets that predominantly consists of similar frames and clips, it is still preferable to have a summary that offers a broader range of visual content. Although a viewer can grasp the essence of the video by examining a single frame, it is advantageous to have a summary that encompasses a diverse set of visual information [56].

7.2. Datasets

The following datasets prevail in the personalized video summary bibliography: UT Egocentric (UTE) [82], SumMe [80], TVSum [81], Open Video Project (OVP) [84], Query-Focused dataset [5] and TV episodes dataset [83]. To evaluate the performance of personalized video summarization, Table 4 provides the datasets used by each method.

Table 4. Datasets used by each method to evaluate the performance of personalized video summarization.

Method	**Dataset**																	
	SumMe	**TVSum**	**UGSum**	**TVC**	**UT Egocentric**	**TV Episodes**	**OVP**	**Youtube Action**	**VSUMM**	**Sports-1M**	**Hollywood-2**	**Query-Focused**	**ActivityNet**	**FineGym**	**Flickr**	**UGSum52**	**Disney**	**MVS1K**	**HUJI**
Shafeian and Bhanu, 2012 [15]								x											
Sharghi et al., 2016 [48]					x	x													
Yin et al., 2016 [51]	x														x				
Plummer et al., 2017 [77]					x	x													
Choi et al., 2018 [57]					x	x													
Mirzasoleiman et al., 2018 [102]							x	x											
Tejero-de-Pablos et al., 2018 [45]										x	x								
Zhang et al., 2018 [64]												x							
Dong et al., 2019 [53]				x															
Dong et al., 2019 [54]	x			x															
FrameRank, 2019 [61]	x	x	x		x											x			
Gunawardena et al., 2019 [59]	x																		

Table 4. *Cont.*

Method	Dataset																		
	SumMe	TVSum	UGSum	TVC	UT Egocentric	TV Episodes	OVP	Youtube Action	VSUMM	Sports-1M	Hollywood-2	Query-Focused	ActivityNet	FineGym	Flickr	UGSum52	Disney	MVS1K	HUJI
Jiang and Han, 2019 [69]												x							
Ji et al., 2019 [71]		x																x	
Qayyum et al., 2019 [23]							x												
Xiao et al., 2020 [78]	x	x											x						
Nagar et al., 2021 [25]	x	x			x					x							x		x
Narasimhan et al., 2021 [58]	x	x										x							
Saquil et al., 2021 [34]	x	x												x					
Köprü and Erzin, 2022 [32]		x																	
Mujtaba et al., 2022 [31]	x																		
Ul et al., 2022 [33]		x							x										
Sosnovik et al., 2023 [56]	x	x						x											

7.3. Evaluation Metrics and Parameters

Ground truth-based metrics evaluate the performance of video summary techniques. Users who have watched the videos in the recording phase of the experiment prepare the ground truth, as each method is for a personal video summary.

7.3.1. Objective Metrics

The evaluation of produced summaries, specifically the evaluation of the effectiveness of a video summarizer, poses a significant hurdle. This stems primarily from the absence of a singular qualitative and quantitative measure for assessment. The frequently used quantitative (objective) evaluation metrics are outlined below.

Using different sets of quantitative evaluation metrics, the summaries produced by different methods are compared with the ground truth. A set of evaluation metrics proposed in [87] is the user summary error rate comparison (CUS_E) and the user summary accuracy comparison (CUS_A), defined as follows:

$$CUS_A = \frac{n_{mAS}}{n_{US}} \tag{1}$$

$$CUS_E = \frac{n_{m'AS}}{n_{US}} \tag{2}$$

The fundamental explanations of the variables utilized in the equations above are provided below:

- n_{mAS}: Denotes the number of matching keyframes in the automated summary (AS).
- $n_{m'AS}$: Denotes the number of non-matching keyframes in the automated summary.
- n_{US}: Denotes the number of keyframes in the user manual summary (US).

The values of CUS_A vary between 0 (representing the worst scenario, where none of the AS keyframes align with the keyframes from the US, or vice versa) and 1 (indicating the best scenario, where all keyframes from the US align with those from AS). It is essential to understand that a CUS_A value of 1 does not necessarily imply that all AS and US keyframes are matched. For instance, if $n_{US} < n_{AS}$ (where n_{AS} is the number of AS keyframes) and $CUS_A = 1$, it indicates that some AS keyframes did not find a match. Concerning CUS_E, the

values range from 0 (the optimal scenario, where all AS keyframes match US keyframes) to n_{AS}/n_{US} (the least favorable scenario, where none of the AS keyframes align with US keyframes, or vice versa). This means that the metrics CUS_A and CUS_E are complementary, with the highest summary quality achieved when $CUS_A = 1$ and $CUS_E = 0$, indicating a precise match between all AS and US keyframes [87].

A set of frequently used evaluation metrics for the video summarization problem is recall, precision, and F-score [23]. Recall is defined as the quotient of retrieved and relevant frames divided by all relevant frames. Precision is defined as the quotient of relevant and retrieved frames divided by all retrieved frames. F-score is the most commonly used metric and is defined as the harmonic mean of precision and recall. Popular datasets such as TVSum and SumMe include reference summaries (human-annotated summaries). Evaluation can be performed by comparing some reference summaries from a dataset with the personalized video summary generated.

Recall, precision, and F-score are defined as follows:

$$REC = \frac{TP}{TP + FN} \tag{3}$$

$$PR = \frac{TP}{TP + FP} \tag{4}$$

$$F - score = 2 \cdot \frac{PR \cdot REC}{PR + REC} \tag{5}$$

The fundamental explanations of the variables utilized in the equations above are provided below:

- *TP (true positive):* Denotes the number of keyframes extracted by the summarization method that exist in the ground truth.
- *FP (false positive):* Denotes the number of keyframes extracted by the summarization method that do not exist in the ground truth.
- *FN (false negative):* Denotes the number of keyframes existing in the ground truth that are not extracted by the summarization method.

Precision and recall metrics vary between 0 and 1, with 0 indicating the lowest performance and 1 indicating the highest. These two metrics are interdependent and cannot be considered separately. An optimal automated video summary is characterized by a high F-score, indicating elevated values for both precision and recall [23].

Another set of evaluation metrics proposed in [59] is the True Positive (PTP) and False Positive (PFP) percentages. The percentages of True Positive and False Positive were determined in comparison to the generated keyframe(s). A threshold of 0.4 was applied to interpret the predictions. If the calculated value exceeds 0.4, it is classified as a keyframe. Modifying the threshold value can lead to different results. True positive and false positive are defined as follows:

$$PTP = 100 \cdot \frac{TP}{TP + FN} \tag{6}$$

$$PFP = 100 \cdot \frac{FP}{FP + TN} \tag{7}$$

TN (true negative) denotes the number of keyframes not extracted by the summarization method that do not exist in the ground truth.

Accuracy (*ACC*) is widely recognized as a fundamental evaluation criterion in the field of machine learning. It serves as a numerical indicator of the proportion of accurate forecasts relative to all forecasts generated. Although this evaluation is particularly clear cut in scenarios involving binary and multiclass classification, it is crucial to grasp the subtleties and constraints associated with it. Accuracy is defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Other metrics were also used, such as the root mean square error (*RMSE*), mean absolute error (*MAE*), average accuracy rate (*AAR*), Kappa statistics, mean Average Precision (*MAP*) [101], mean average distance (*mAD*), Relaxed F-score (*RFS*) [103], Average Human Rating (*AHR*), Kendall's coefficient [104], Spearman's coefficient [105] and the ROUGE-SU metric [106].

7.3.2. Subjective Metrics

There is no universally accepted measure to evaluate the quality of a summary produced by qualitative or subjective means. Subjective evaluation is based on subjective judgments and is performed using questionnaires in which users answer the questions in the form of subjective metrics such as ranking, cover, understanding, representativeness, diversity, information, etc. User feedback is typically collected to collect scores that are subsequently analyzed for evaluation purposes [49].

7.4. Evaluation Measures and Protocols

In the following, we present the evaluation measures and protocols of personalized video methods.

7.4.1. Evaluating Personalized Video Storyboards

Early personalized video summarization techniques using representative frames created a static personalized video summary. To evaluate the quality of the summaries, the first attempts used more human judges, which are subjective criteria. Judgment with human judges was based on criteria such as the percentage of interesting events in the videos produced [22], the provision of questionnaires to users to evaluate feedback [16,46], or the relevance to the video content of each keyframe [16].

Evaluations relying on subjective criteria have certain drawbacks, including the limitation that the outcomes of the evaluation may not be applicable to future comparisons and the fact that subjective judgments by human evaluators typically consume more time compared to objective evaluations. Consequently, recent research endeavors have shifted towards utilizing ground truth summaries and objective metrics for evaluating summaries, aiming to circumvent the limitations associated with subjective evaluation. Recent studies evaluate the effectiveness of their keyframe-based summaries using objective metrics and comparing them with ground truth summaries [79].

In this context, ref. [23] estimates the performance of the proposed methodology using the RMSE, MAE, AAR, and Kappa statistics. Instead of the *RMSE*, *MAE*, *AAR*, and Kappa statistics, in [23], the evaluation is based on the known comparison of the user summary error rate (*CUSE*) and a comparison of the user summary accuracy (*CUSA*), and recall, precision, and the F-score. In addition to recall, precision, and F-score in [33], accuracy was used to evaluate the quality of the summary produced. Recall, precision, and F-score were also used in [59] to evaluate the Pipe-1 processing pipeline by investigating the performance of the units used to select the minimum number of keyframes by which the maximum number of objects was represented. In [59], the Pipe-2 processing pipeline was evaluated by measuring the ability to recognize frames included in the OoI using True Positive and False Positive percentages. Each video was tested with a specific OoI, and calculations were performed accordingly.

In [58], a different method was proposed that evaluates the summary generated according to keyframes that received high scores, comparing the F-score of their method with the F-score of other approaches in the SumMe, TVSum, and query-focused video summarization datasets (QFVS). A five-fold cross-validation with all splits having an average F-score was used in [56] to quantitatively evaluate the quality of the summaries produced. The evaluation of the method's ability to produce videos in various manners is conducted using keyframes by calculating the harmonic mean F-score in [54].

In [57], an evaluation was conducted using MAP to measure the degree of similarity of ground truth video summaries with predicted video summaries. In [57], also the mAD

metric was proposed to measure the average normalized time distance between the closest portion of the predicted summary and the summary reporting portions of the ground truth. In [32], two new metrics were defined to evaluate the emotional video summary. In the first measurement, differences in the F-score were normalized with the baseline relative to videos whose number of impressions of the face was higher. The percentage of recall of frames displayed on faces in the exported summary was defined as the second metric.

7.4.2. Evaluating Personalized Video Skims

Personalized video summarization techniques using representative video fragments create a dynamic personalized video summary. To accomplish this, they choose the most characteristic video segments and combine them in a specific order to create a shorter video. The evaluation methods used in dynamic video summary studies evaluate the quality of video summaries by comparing them with human preferences, primarily based on objective criteria. In the first work that evaluated the quality of the personalized video summary, subjective tests were performed using human judges who had adopted the MOS (Mean Opinion Score) test for the score [14]. MOS test is a subjective metric in which each test evaluates the produced video twice with a score, and the final score is the average of two scores.

The temporal segmentation methodology [61] in which the segments were semantically consistent and appropriate was evaluated by calculating the F-score. The effectiveness of the two-stream method [45] in extracting highlights was evaluated on the basis of the F-score. A highlight could be made up of successive video segments. Hence, while the absence of a segment may not have a substantial effect on the F-score, it does influence the flow of the video, which in turn impacts the clarity and user experience of the summary. In this context, a measure is established to assess the completeness of a highlighted text snippet by calculating the overlap proportion between the highlighted snippet and its corresponding true representation. The matching of extracted and true highlights is not straightforward and was performed using a greedy algorithm that aims to maximize the total completeness of all highlights [45].

The performance of the real-time method, called STREAMING LOCAL SEARCH [102], was measured by recall, precision, and F-score on the YouTube and Open Video Project datasets under a segment size equal to ten. A different evaluation method was introduced in [5]. This technique gauges the effectiveness of the model by determining the semantic similarity between various shots, rather than simply assessing temporal overlap or depending on basic visual features. Initially, the conceptual similarity between each pair of video shots is calculated based on the Intersection-over-Union (IoU) of their associated concepts. Following that, the conceptual similarity is used as the edge weights in the bipartite graph between two summaries, replacing basic visual features that lack semantic information, to identify the graph's maximum weight matching. Ultimately, precision, recall, and F-score can be determined based on the number of matching pairs. This approach was also used in [78]. Nagar et al. [25] used four evaluation measures to demonstrate the effectiveness of their proposed framework. For the TVSum and SumMe datasets, they created a video summary, which is a percentage of around 15% of the length of the original video, and to measure the quality, from several ground truth summaries, they reported the average traditional F-score. RFS was the metric that they used in the first assessment measure. Based on the natural language description, they evaluated the video summaries in the second evaluation measure. The third rating was called the AHR, as it was performed with the confidence rating of 10 participants based on the enjoyment and completeness of the video summary. In the fourth evaluation measure, the summary score is based on the number of unique recorded events and the present jerks.

Rather than traditional methods, ref. [34] employed different assessment measures that juxtapose the significance scores of the reference and the projected summary. These measures are rank correlation coefficients, specifically Kendall's τ coefficient, which was used to assess the technique and contrast it with cutting-edge methods, and Spearman's ρ

coefficient. According to [34], segment-level GT summaries were used to train this model, and frame-level reference summaries were used to test this model and the baseline methods. When a model is trained on segment-level features, the importance score predicted for a frame is identical to the importance score predicted for the segment that includes that frame. The video correlation coefficient resulting from a video reference summary and a predicted summary is the average of the correlation coefficients between the predicted summary and each reference summary.

Sharghi et al. [48] introduced a new metric known as hitting recall to assess system summaries from a query-focused point of view. They conducted an evaluation of the system-generated video summary, which was based on dense text annotations. Specifically, text representations were assigned to video summaries, and then these were compared using the precision, recall, and F-score provided by the ROUGE-SU metric. An alternative method was also used in [77] to evaluate the approach suggested in the text domain in the TV episodes dataset and the UT Egocentric dataset. The F-score and the recall for each dataset were reported using the ROUGE-SU score. When a video summary was generated using their method during the testing, a corresponding summary was created by stringing together the text annotations of the sections of the original text that constitute the summary. The recall-based ROUGE metric was then used to compare the generated summary with three reference summaries and three summaries provided by humans.

7.4.3. Evaluating Personalized Fast Forward

Initial efforts to assess the quality of the generated fast forward personalized video summaries relied on user studies, employing human evaluators to judge the outcome of the summaries, a process that inherently involves subjective criteria. The first fast forward personalized video summarization method in [72] used human judges to assess the quality of the summary through a comparison of three different types of summaries. A summary was prepared using periodic sampling with a constant segment length and interval, a different kind of summary was created using the proposed method based on the behavior of another individual, and a summary was made using the proposed method based on the subject's behavioral data. In the study by [42], two subjective evaluations were performed. The first evaluation focused on determining the appropriate playback speeds. A group of 25 participants was requested to identify their maximum bearable playback speed, their comfortable playback speeds, and their most comfortable playback speed. This was performed while four sets of video samples, each with different playback speeds, were shown while the players browsed the broadcast soccer videos. The second objective evaluation aimed to collect the general perception of the audience through a comparative analysis of the summaries produced. A group of 23 participants was requested to express their preference among three summaries created using the various methods mentioned above (in no particular order). Their judgment was based on the completeness of the summaries, the ease of understanding, and the efficient use of time.

An alternative evaluation method was proposed by Chen et al. [41] for the purpose of investigating certain behaviors of the personalized summarization technique. They compared, with several objective criteria they defined, the personalized summarization method. The first criterion L1 was established as the normalized information density of its frames, with the aim of scrutinizing adaptive fast forwarding for browsing semantic events. A higher L1 value indicates that the method has more semantic significance in relation to the annotated events, as it allocates slower playback speeds to clips that possess both increased event importance and scene activities. The comfort summary should be replayed at a sufficiently slow pace, with the speeds smoothly transitioning to avoid annoying flickering. The comfort assessment is based on two factors: the mean playback speed L2 and the variability in playback speeds from one frame to the next L3. L2 and L3 were conceptualized with the goal of adaptive fast forwarding to produce a summary that is visually pleasing. The elevated level of fluctuation L3 signifies frequent and drastic changes in the playback speed within the summary. By reducing the playback speed L2 through

the omission of less crucial content, the fluctuation level L3 is significantly lowered as a result of clip-based summarization. In the end, they defined L4 as the standardized concentration of data associated with a particular object in the summary, with the intent to facilitate adaptive fast forwarding for the arrangement of narrative stories. When an object is identified, its corresponding clips will have increased weights, leading to a higher L4 value.

The evaluation in [51] is based on widely recognized precision, recall, and F-score metrics. Experiments were conducted on 20 videos from the publicly available SumMe dataset. The videos were categorized into three groups based on the characteristics of the camera: static, moving, and egocentric. A set of study participants was tasked with creating video summaries that encapsulate most of the significant content, and each video in the dataset was summarized by 15 to 18 unique individuals. Segments surrounding the highest-ranked frames were chosen, and a summary was generated, whose length was determined to be 15% of the original video. Furthermore, the results of assigning arbitrary scores to characterize the dataset were reported. The pairwise F-score was utilized as a metric. Precision and recall were calculated per frame for each human-generated ground truth. Lastly, the F-scores across the ground truth chosen by various individuals were averaged to serve as the final evaluation measure.

7.4.4. Evaluating Personalized Multiview

Initial attempts were made to evaluate the quality of personalized summaries created from videos recorded simultaneously with multiple cameras, employing human evaluators to judge the outcome of the summaries, a process that inherently involves subjective criteria. In [39], separate subjective evaluations were performed on viewpoint selection and summarization. The generated results were evaluated on the basis of their overall impact and visual/storytelling elements, confirming the effectiveness of each respective method. Consequently, subjective assessments were performed to validate the pertinence of their personalized video production idea and the efficacy of their suggested execution to realize this personalized production.

A different evaluation approach was suggested in [26], where initially, the summaries of the results were contrasted using the proposed method alongside the summaries of three users. Upon inputting the DOI values, the users chose 12 distinct shots of the 29 available. Precision and recall metrics were calculated for each user. Given that certain events appeared more than twice in the scenario, all but one shot, containing a single event, were disregarded by the users. On the assumption that shots featuring the same event and individual are identical, the results were promising, with mean precision and recall values of 0.94. The count of shots chosen by the users was compared with those selected using the proposed method. A subjective test was carried out to assess the efficacy of the method. Ten university students were asked two questions. The responses to these questions ranged from -3 to +3, with the symbols '$-$' and '+' representing 'No' and 'Yes', respectively. This test facilitated the comparison of three view transition techniques.

The effectiveness of the QUASC method [71] was evaluated based on precision, recall, and F-score. The intrinsic quality of the QUASC technique was gauged by juxtaposing the automatically produced keyframes with the manually annotated ground truth from their proprietary MVS1K dataset and the title-based video summarization (TVSum) dataset. Specifically, the Euclidean distance was initially computed between each produced keyframe and each ground truth keyframe individually. If the normalized distance fell below a predetermined experimental threshold of 0.6, the two keyframe varieties would be deemed a match; subsequently, they would be omitted in the next comparison cycle. In this work, a subjective user study was also conducted as a further evaluation among 6 participants with 4 females and 2 males. Each user was familiar with the video content to be summarized and was required to evaluate the summarizations generated by the three approaches for the ten query-based video sets Q1–Q10. Participants were required to assign each summarization a score between 1 (poor) and 10 (good), indicating whether the

summarization satisfies the three properties with high visual quality. This is in accordance with the approval of the study ethics to ensure that the participants can score fairly.

8. Quantitative Comparisons

In this section, we present quantitative comparisons of personalized video summarization approaches on the most prevalent datasets in the literature, which are TVSum and SumMe. The evaluation of unsupervised personalized video summarization methods using the F-score in the SumMe dataset is presented in Table 5. Table 6 shows the performance of the personalized unsupervised video summarization methods in the TVSum dataset evaluated using the F-score. Based on the F-score results in Tables 5 and 6, the following observations are worth mentioning.

Table 5. Comparison of unsupervised methods on SumMe dataset.

Method	F-Score
FrameRank (KL divergence based) [61]	0.453
Actor–Critic [25]	0.464
CLIP-It [58]	0.525
CSUM-MSVA [56]	0.582

Table 6. Comparison of unsupervised methods on TVSum dataset.

Method	F-Score
FrameRank (KL divergence based) [61]	0.601
QUASC [71]	0.54
Actor–Critic [25]	0.583
CLIP-It [58]	0.63
CSUM-MSVA [56]	0.639

- The FrameRank (KL divergence based) approach performs higher on the TVSum dataset than it does on the SumMe dataset. This unbalanced performance shows that this technique is more suited to the TVSum dataset. On the contrary, the CSUM-MSVA and CLIP-It approaches have balanced performance, as they show high performance on both datasets.
- A very good choice for unsupervised personalized video summarization is to use the contrastive learning framework which includes diversity and representatives, as this framework is based on the CSUM-MSVA method, which is the most effective and top performing method in both datasets.
- The advanced CLIP-It method (CLIP-Image + Video Caption + Transformer) provides a high F-score on both datasets. From the comparison of this method with other methods (GoogleNet + bi-LSTM, ResNet + bi-LSTM, CLIP-Image + bi-LSTM, CLIP-Image + Video Caption + bi-LSTM, GoogleNet + transformer, ResNet + transformer, CLIP-Image + transformer) in [58], the benefits of this method are realized, as it has better results in all three settings. According to the CLIP-It method, the fusion of language and image embedding is performed using pre-trained networks through learned language, which is guided by multiple attention heads. All frames are tracked together using the Frame Score Transform, which predicts frame relevance scores. Using the Knapsack algorithm, frame scores are converted into high-scoring shot scores. Therefore, it is one of the top methods for unsupervised personalized video summarization.

  Figure 11 underscores the importance of incorporating language, particularly through dense video captions, for creating generic video summaries with a qualitative example.

The video features a woman explaining the process of making a chicken sandwich. The ground truth summary presents scores derived by averaging user annotations and highlights the keyframes that scored highly. Subsequently, the figure displays outcomes from the baseline CLIP-Image+Transformer, which relies solely on visual cues without linguistic input. The scoring indicates that frames highly rated in the ground truth also score well in the baseline; however, numerous irrelevant frames also score highly. Consequently, the model selects keyframes where the individual is either speaking or consuming the sandwich (marked in red), which are not representative of the crucial steps in the video. Introducing language through generated captions mitigates this issue. The final row depicts captions created by the bi-modal transformer (BMT) [107]. These results, from the complete CLIP-It model, show predicted scores aligning more closely with the ground truth, with the highest scoring keyframes matching those of the ground truth.

Figure 11. Comparison of the reference summary with the outcomes obtained from the CLIP-Image + Transformer and the complete CLIP-It model (CLIP-Image + Video Caption + Transformer). The content provided is a video demonstrating a cooking procedure. Without captions, the model gives high scores to some irrelevant frames, like those showing the woman talking or eating, which negatively impacts the precision. By including captions, the cross-attention mechanism guarantees that frames featuring significant actions and objects receive elevated scores [58].

- The multimodal Multigraph Fusion (MGF) method, introduced in the QUASC approach to automatically identify events that are related to the query based on the tag information in the video, such as descriptions and titles, does not allow one to create a good summary. As a result, the QUASC approach is not competitive with any other approach in the TVSum dataset.
- The Actor–Critic method allows the creation of a satisfactory personalized video summary, which is shown by the F-score in the two datasets. To capture spatio-temporal features, 3D convolutional neural networks (CNNs) are used. In addition, a bi-directional short-term memory network (BiLSTM) is used to learn the extracted features. However, the Actor–Critic method is not competitive with the pioneering CSUM-MSVA and CLIP-It method.

Table 7 shows the performance of the supervised personalized video summarization methods evaluated using the F-score in the SumMe dataset. Table 8 refers to the performance of the personalized supervised video summarization methods in the TVSum dataset

evaluated by the F-score. Based on the F-score results in Tables 7 and 8, the following observations are worth mentioning.

Table 7. Comparison of supervised methods on SumMe dataset.

Method	F-Score
Dong et al. [54]	0.432
Xiao et al. [78]	0.435
CLIP-It [58]	0.542

Table 8. Comparison of supervised methods on TVSum dataset.

Method	F-Score
Xiao et al. [78]	0.573
CLIP-It [58]	0.663

- The use of structural and optical information labeling in the method of Dong et al. [54] does not appear to allow the production of a good personalized supervised summary, as it provides satisfactory results that are competitive only with the method of Xiao et al. [78] in the SumMe dataset.
- Using the reinforced attentive description generator and the query-aware scoring module(QSM) does not seem to help, as Xiao et al.'s [78] method is not competitive compared to the CLIP-It method on both datasets.
- Balanced performance seems to be given by the [78] method of Xiao et al. and by the CLIP-It method, as neither seems to be adapted to a single dataset.
- The CLIP-It method is a very good choice for creating supervised personalized video summarization, as no other method has competitive F-score values on both datasets.

Based on the F-score results in Tables 5–8, we draw the following conclusions: The CLIP-It method is a good choice for creating an unsupervised personalized video summary and even better for creating a supervised personalized video summary. In each dataset, better results are provided in the F-score for the supervised setting than for the unsupervised setting. The superiority of the F-score in the supervised setting is due to the use of three loss functions (reconstruction, diversity, and classification) in the model training, in contrast to the unsupervised setting that uses only two loss functions (reconstruction and diversity). Finally, contrastive learning outperforms multimodal language-guided transformation in unsupervised video summarization. Therefore, a better choice seems to be the CSUM-MSVA method for unsupervised personalized video summarization, while the CLIP-It method enables a good summarization.

9. Conclusions and Future Directions

In this survey, we presented the evolution of the proposed methodologies for personalized video summarization since the first published method [12] in 2001. We present the usability, the applications, and the classification of the personalized video summarization techniques according to their characteristics. To the best of our knowledge, this is the first survey of personalized video summarization methods and datasets. In the following, we list the main conclusions of this survey.

- The personalized video summary techniques can be classified based on the type of personalized summary, on the criteria, on the video domain, on the source of information, on the time of summarization, and on the machine learning technique.
- Depending on the type of methodology, the techniques can be classified into five major categories, which are feature-based video summarization, keyframe selection, shot selection-based approach, video summarization using trajectory analysis, and personalized video summarization using clustering.

- RNNs and CNNs, when integrated into neural networks, can greatly contribute to the production of high-quality personalized video summaries.
- The advancement of current techniques and the invention of new methods in unsupervised personalized video summarization call for increased research efforts, given that supervised summarization methods frequently require extensive datasets with human annotations.
- Machine learning methods tend to outperform conventional techniques, given their enhanced ability to extract efficient features.

In conclusion, it is seen without a doubt that personalized video summarization techniques are very useful in many applications, such as those we mentioned in Section 2 of this paper. Concerning the future directions of personalized video summarization techniques, the goals of future works on personalized video summarization and given the current technology should be the following:

- For a personalized summary to be more effective, it is crucial to incorporate user preferences extensively. Merely extracting images from the Web is not enough, but there is a need to pull out additional data like user preferences, status, tags, shares, etc. To accomplish this data extraction, the use of advanced joint vision language models along with the application of multimodal information is a necessity in the future.
- The advancement of convolution-oriented deep learning techniques for video summarization in image classification, aimed at conceptual cluster labeling, enhances the quality of summarization while maintaining the precision of image classification. Event detection precision can be significantly improved by developing more intricate and expansive structures with the use of deep neural networks.
- Investigating the potential of architectures using CNNs to improve run-time performance, as query-based summarization methods suffer from an accuracy versus efficiency trade-off.
- Exploring the capacity of query-driven methods for summarization to include the extent to which a user retains an image, entropy, and user focus grounded on visual consistency.
- Further research in summarization techniques that utilize the open problem of audio from a clip to identify objects, subjects, or events, for instance, in sports footage, locating scoring occurrences, or identifying a suspect in surveillance footage. The user's speech can also be considered to gather additional information to enhance the video summary.
- The development of personalized unsupervised video summarization approaches that incorporate visual language into the multitask learning framework to accurately map the relationship between query and visual content.
- The absence of a uniform approach to evaluate produced summaries using standard datasets is a concern, as each study may employ its own evaluation technique and dataset. This situation raises questions about the precision of the outcomes from objective comparisons across numerous studies, complicating the comparison of personalized summarization methods. Consequently, to ensure the credibility of the results of objective comparisons and to facilitate the comparison of personalized summary methods, the establishment of shared datasets and evaluation techniques is imperative.

Author Contributions: The authors contributed equally to this work. Conceptualization, M.P. and C.P.; methodology, M.P. and C.P.; formal analysis, M.P.; investigation, M.P. and C.P.; resources, M.P.; writing—original draft preparation, M.P. and C.P.; writing—review and editing, M.P. and C.P.; visualization, M.P. and C.P.; supervision, C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pritch, Y.; Rav-Acha, A.; Peleg, S. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1971–1984. [CrossRef] [PubMed]
2. Rochan, M.; Wang, Y. Video summarization by learning from unpaired data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7902–7911.
3. Gygli, M.; Grabner, H.; Van Gool, L. Video summarization by learning submodular mixtures of objectives. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3090–3098.
4. Panagiotakis, C.; Doulamis, A.; Tziritas, G. Equivalent key frames selection based on iso-content principles. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *19*, 447–451. [CrossRef]
5. Sharghi, A.; Laurel, J.S.; Gong, B. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–27 July 2017; pp. 4788–4797.
6. Panagiotakis, C.; Papoutsakis, K.; Argyros, A. A graph-based approach for detecting common actions in motion capture data and videos. *Pattern Recognit.* **2018**, *79*, 1–11. [CrossRef]
7. Darabi, K.; Ghinea, G. Personalized video summarization using sift. In Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, 13–17 April 2015; pp. 1252–1256.
8. Sachan, P.R.; Keshaveni, N. Utilizing Image Classification based Semantic Recognition for Personalized Video Summarization. *Int. J. Electron. Eng. Res.* **2017**, *9*, 15–27.
9. Panagiotakis, C.; Papadakis, H.; Fragopoulou, P. Personalized video summarization based exclusively on user preferences. In Proceedings of the European Conference on Information Retrieval, Lisbon, Portugal, 14–17 April 2020; pp. 305–311.
10. Darabi, K.; Ghinea, G. Personalized video summarization based on group scoring. In Proceedings of the 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), Xi'an, China, 9–13 July 2014; pp. 310–314.
11. Money, A.G.; Agius, H. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Commun. Image Represent.* **2008**, *19*, 121–143. [CrossRef]
12. Tseng, B.L.; Lin, C.Y.; Smith, J.R. Video summarization and personalization for pervasive mobile devices. In Proceedings of the Storage and Retrieval for Media Databases 2002, San Jose, CA, USA, 19–25 January 2002; Volume 4676, pp. 359–370.
13. Tseng, B.L.; Smith, J.R. Hierarchical video summarization based on context clustering. In Proceedings of the Internet Multimedia Management Systems IV, Orlando, FL, USA, 7–11 September 2003; Volume 5242, pp. 14–25.
14. Lie, W.N.; Hsu, K.C. Video summarization based on semantic feature analysis and user preference. In Proceedings of the 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (sutc 2008), Taichung, Taiwan, 11–13 June 2008; pp. 486–491.
15. Shafeian, H.; Bhanu, B. Integrated personalized video summarization and retrieval. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 996–999.
16. Zhang, Y.; Ma, C.; Zhang, J.; Zhang, D.; Liu, Y. An interactive personalized video summarization based on sketches. In Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, Hong Kong, China, 17–19 November 2013; pp. 249–258.
17. Valdes, V.; Martínez, J.M. Introducing risplayer: Real-time interactive generation of personalized video summaries. In Proceedings of the 2010 ACM Workshop on Social, Adaptive and Personalized Multimedia Interaction and Access, Firenze, Italy, 29 October 2010; pp. 9–14.
18. Chen, F.; De Vleeschouwer, C. Automatic production of personalized basketball video summaries from multi-sensored data. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 565–568.
19. Kannan, R.; Ghinea, G.; Swaminathan, S. What do you wish to see? A summarization system for movies based on user preferences. *Inf. Process. Manag.* **2015**, *51*, 286–305. [CrossRef]
20. Garcia del Molino, A. First person view video summarization subject to the user needs. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1440–1444.
21. Huang, J.H.; Worring, M. Query-controllable video summarization. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 242–250.
22. Katti, H.; Yadati, K.; Kankanhalli, M.; Tat-Seng, C. Affective video summarization and story board generation using pupillary dilation and eye gaze. In Proceedings of the 2011 IEEE International Symposium on Multimedia, Dana Point, CA, USA, 5–7 December 2011; pp. 319–326.
23. Qayyum, H.; Majid, M.; ul Haq, E.; Anwar, S.M. Generation of personalized video summaries by detecting viewer's emotion using electroencephalography. *J. Vis. Commun. Image Represent.* **2019**, *65*, 102672. [CrossRef]
24. Varini, P.; Serra, G.; Cucchiara, R. Personalized egocentric video summarization for cultural experience. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 539–542.
25. Nagar, P.; Rathore, A.; Jawahar, C.; Arora, C. Generating Personalized Summaries of Day Long Egocentric Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *45*, 6832–6845. [CrossRef]
26. Park, H.S.; Cho, S.B. A personalized summarization of video life-logs from an indoor multi-camera system using a fuzzy rule-based system with domain knowledge. *Inf. Syst.* **2011**, *36*, 1124–1134. [CrossRef]

27. Peng, W.T.; Chu, W.T.; Chang, C.H.; Chou, C.N.; Huang, W.J.; Chang, W.Y.; Hung, Y.P. Editing by viewing: Automatic home video summarization by viewing behavior analysis. *IEEE Trans. Multimed.* **2011**, *13*, 539–550. [CrossRef]
28. Ul Haq, I.; Ullah, A.; Muhammad, K.; Lee, M.Y.; Baik, S.W. Personalized movie summarization using deep cnn-assisted facial expression recognition. *Complexity* **2019**, *2019*, 3581419. [CrossRef]
29. Zhou, P.; Xu, T.; Yin, Z.; Liu, D.; Chen, E.; Lv, G.; Li, C. Character-oriented video summarization with visual and textual cues. *IEEE Trans. Multimed.* **2019**, *22*, 2684–2697. [CrossRef]
30. Fei, M.; Jiang, W.; Mao, W. Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization. *Expert Syst. Appl.* **2021**, *166*, 114036. [CrossRef]
31. Mujtaba, G.; Malik, A.; Ryu, E.S. LTC-SUM: Lightweight Client-driven Personalized Video Summarization Framework Using 2D CNN. *arXiv* **2022**, arXiv:2201.09049.
32. Köprü, B.; Erzin, E. Use of Affective Visual Information for Summarization of Human-Centric Videos. *IEEE Trans. Affect. Comput.* **2022**, *14*, 3135–3148. [CrossRef]
33. Ul Haq, H.B.; Asif, M.; Ahmad, M.B.; Ashraf, R.; Mahmood, T. An Effective Video Summarization Framework Based on the Object of Interest Using Deep Learning. *Math. Probl. Eng.* **2022**, *2022*, 7453744. [CrossRef]
34. Saquil, Y.; Chen, D.; He, Y.; Li, C.; Yang, Y.L. Multiple Pairwise Ranking Networks for Personalized Video Summarization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1718–1727.
35. Darabi, K.; Ghinea, G. Personalized video summarization by highest quality frames. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 14–18 July 2014; pp. 1–6.
36. Fei, M.; Jiang, W.; Mao, W. Creating personalized video summaries via semantic event detection. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 14931–14942. [CrossRef]
37. Haq, H.B.U.; Asif, M.; Ahmad, M.B. Video summarization techniques: A review. *Int. J. Sci. Technol. Res.* **2020**, *9*, 146–153.
38. Hannon, J.; McCarthy, K.; Lynch, J.; Smyth, B. Personalized and automatic social summarization of events in video. In Proceedings of the 16th International Conference on Intelligent User Interfaces, Palo Alto, CA, USA, 13–16 February 2011; pp. 335–338.
39. Chen, F.; Delannay, D.; De Vleeschouwer, C. An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study. *IEEE Trans. Multimed.* **2011**, *13*, 1381–1394. [CrossRef]
40. Olsen, D.R.; Moon, B. Video summarization based on user interaction. In Proceedings of the 9th European Conference on Interactive TV and Video, Lisbon, Portugal, 29 June 29–1 July 2011; pp. 115–122.
41. Chen, F.; De Vleeschouwer, C.; Cavallaro, A. Resource allocation for personalized video summarization. *IEEE Trans. Multimed.* **2013**, *16*, 455–469. [CrossRef]
42. Chen, F.; De Vleeschouwer, C. Personalized summarization of broadcasted soccer videos with adaptive fast-forwarding. In Proceedings of the International Conference on Intelligent Technologies for Interactive Entertainment, Mons, Belgium, 3–5 July 2013; pp. 1–11.
43. Kao, C.C.; Lo, C.W.; Lu, K.H. A personal video summarization system by integrating RFID and GPS information for marathon activities. In Proceedings of the 2015 IEEE 5th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany, 6–9 September 2015; pp. 347–350.
44. Sukhwani, M.; Kothari, R. A parameterized approach to personalized variable length summarization of soccer matches. *arXiv* **2017**, arXiv:1706.09193.
45. Tejero-de Pablos, A.; Nakashima, Y.; Sato, T.; Yokoya, N.; Linna, M.; Rahtu, E. Summarization of user-generated sports video by using deep action recognition features. *IEEE Trans. Multimed.* **2018**, *20*, 2000–2011. [CrossRef]
46. Niu, J.; Huo, D.; Wang, K.; Tong, C. Real-time generation of personalized home video summaries on mobile devices. *Neurocomputing* **2013**, *120*, 404–414. [CrossRef]
47. Basavarajaiah, M.; Sharma, P. Survey of compressed domain video summarization techniques. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–29. [CrossRef]
48. Sharghi, A.; Gong, B.; Shah, M. Query-focused extractive video summarization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 3–19.
49. Tiwari, V.; Bhatnagar, C. A survey of recent work on video summarization: Approaches and techniques. *Multimed. Tools Appl.* **2021**, *80*, 27187–27221. [CrossRef]
50. Chung, C.T.; Hsiung, H.K.; Wei, C.K.; Lee, L.S. Personalized video summarization based on Multi-Layered Probabilistic Latent Semantic Analysis with shared topics. In Proceedings of the The 9th International Symposium on Chinese Spoken Language Processing, Singapore, 12–14 September 2014; pp. 173–177.
51. Yin, Y.; Thapliya, R.; Zimmermann, R. Encoded semantic tree for automatic user profiling applied to personalized video summarization. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 181–192. [CrossRef]
52. Baghel, N.; Raikwar, S.C.; Bhatnagar, C. Image Conditioned Keyframe-Based Video Summarization Using Object Detection. *arXiv* **2020**, arXiv:2009.05269.
53. Dong, Y.; Liu, C.; Shen, Z.; Han, Y.; Gao, Z.; Wang, P.; Zhang, C.; Ren, P.; Xie, X. Personalized Video Summarization with Idiom Adaptation. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1041–1043.

54. Dong, Y.; Liu, C.; Shen, Z.; Gao, Z.; Wang, P.; Zhang, C.; Ren, P.; Xie, X.; Yu, H.; Huang, Q. Domain Specific and Idiom Adaptive Video Summarization. In Proceedings of the ACM Multimedia Asia, Beijing, China, 15–18 December 2019; pp. 1–6.
55. Xiao, S.; Zhao, Z.; Zhang, Z.; Yan, X.; Yang, M. Convolutional hierarchical attention network for query-focused video summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12426–12433.
56. Sosnovik, I.; Moskalev, A.; Kaandorp, C.; Smeulders, A. Learning to Summarize Videos by Contrasting Clips. *arXiv* **2023**, arXiv:2301.05213.
57. Choi, J.; Oh, T.H.; Kweon, I.S. Contextually customized video summaries via natural language. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1718–1726.
58. Narasimhan, M.; Rohrbach, A.; Darrell, T. CLIP-It! language-guided video summarization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 13988–14000.
59. Gunawardena, P.; Sudarshana, H.; Amila, O.; Nawaratne, R.; Alahakoon, D.; Perera, A.S.; Chitraranjan, C. Interest-oriented video summarization with keyframe extraction. In Proceedings of the 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2–5 September 2019; Volume 250; pp. 1–8.
60. Tseng, B.L.; Lin, C.Y.; Smith, J.R. Video personalization and summarization system for usage environment. *J. Vis. Commun. Image Represent.* **2004**, *15*, 370–392. [CrossRef]
61. Lei, Z.; Zhang, C.; Zhang, Q.; Qiu, G. FrameRank: A text processing approach to video summarization. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 368–373.
62. Varini, P.; Serra, G.; Cucchiara, R. Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Trans. Multimed.* **2017**, *19*, 2832–2845. [CrossRef]
63. Liu, H.; Sun, F.; Zhang, X.; Fang, B. Interactive video summarization with human intentions. *Multimed. Tools Appl.* **2019**, *78*, 1737–1755. [CrossRef]
64. Zhang, Y.; Kampffmeyer, M.; Liang, X.; Tan, M.; Xing, E.P. Query-conditioned three-player adversarial network for video summarization. *arXiv* **2018**, arXiv:1807.06677.
65. Zhang, Y.; Kampffmeyer, M.; Zhao, X.; Tan, M. Deep reinforcement learning for query-conditioned video summarization. *Appl. Sci.* **2019**, *9*, 750. [CrossRef]
66. Ghinea, G.; Kannan, R.; Swaminathan, S.; Kannaiyan, S. A novel user-centered design for personalized video summarization. In Proceedings of the 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, China, 14–18 July 2014; pp. 1–6.
67. Kannan, R.; Ghinea, G.; Swaminathan, S.; Kannaiyan, S. Improving video summarization based on user preferences. In Proceedings of the 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Jodhpur, India, 18–21 December 2013; pp. 1–4.
68. del Molino, A.G.; Boix, X.; Lim, J.H.; Tan, A.H. Active video summarization: Customized summaries via on-line interaction with the user. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, Francisco, CA, USA, 4–9 February 2017.
69. Jiang, P.; Han, Y. Hierarchical variational network for user-diversified & query-focused video summarization. In Proceedings of the 2019 on International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 202–206.
70. Cizmeciler, K.; Erdem, E.; Erdem, A. Leveraging semantic saliency maps for query-specific video summarization. *Multimed. Tools Appl.* **2022**, *81*, 17457–17482. [CrossRef]
71. Ji, Z.; Ma, Y.; Pang, Y.; Li, X. Query-aware sparse coding for web multi-video summarization. *Inf. Sci.* **2019**, *478*, 152–166. [CrossRef]
72. Yoshitaka, A.; Sawada, K. Personalized video summarization based on behavior of viewer. In Proceedings of the 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, Sorrento, Italy, 25–29 November 2012; pp. 661–667.
73. Namitha, K.; Narayanan, A.; Geetha, M. Interactive visualization-based surveillance video synopsis. *Appl. Intell.* **2022**, *52*, 3954–3975. [CrossRef]
74. Otani, M.; Nakashima, Y.; Sato, T.; Yokoya, N. Video summarization using textual descriptions for authoring video blogs. *Multimed. Tools Appl.* **2017**, *76*, 12097–12115. [CrossRef]
75. Miniakhmetova, M.; Zymbler, M. An approach to personalized video summarization based on user preferences analysis. In Proceedings of the 2015 9th International Conference on Application of Information and Communication Technologies (AICT), Rostov on Don, Russia, 14–16 October 2015; pp. 153–155.
76. Seera, M.; Lim, C.P. Transfer learning using the online fuzzy min–max neural network. *Neural Comput. Appl.* **2014**, *25*, 469–480. [CrossRef]
77. Plummer, B.A.; Brown, M.; Lazebnik, S. Enhancing video summarization via vision-language embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5781–5789.
78. Xiao, S.; Zhao, Z.; Zhang, Z.; Guan, Z.; Cai, D. Query-biased self-attentive network for query-focused video summarization. *IEEE Trans. Image Process.* **2020**, *29*, 5889–5899. [CrossRef]
79. Apostolidis, E.; Adamantidou, E.; Metsai, A.I.; Mezaris, V.; Patras, I. Video summarization using deep neural networks: A survey. *Proc. IEEE* **2021**, *109*, 1838–1863. [CrossRef]

80. Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. Creating summaries from user videos. In *Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part VII 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 505–520.
81. Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A. Tvsum: Summarizing web videos using titles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5179–5187.
82. Lee, Y.J.; Ghosh, J.; Grauman, K. Discovering important people and objects for egocentric video summarization. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1346–1353.
83. Yeung, S.; Fathi, A.; Fei-Fei, L. Videoset: Video summary evaluation through text. *arXiv* **2014**, arXiv:1406.5824.
84. Geisler, G.; Marchionini, G. The open video project: Research-oriented digital video repository. In Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, TX, USA, 2–7 June 2000; pp. 258–259.
85. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
86. Zhang, Y.; Xie, Y.; Zhang, Y.; Dai, Y.; Ren, F. VSSum: A Virtual Surveillance Dataset for Video Summary. In Proceedings of the 5th International Conference on Control and Computer Vision, Xiamen, China, 19–21 August 2022; pp. 113–119.
87. De Avila, S.E.F.; Lopes, A.P.B.; da Luz, A., Jr.; de Albuquerque Araújo, A. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.* **2011**, *32*, 56–68. [CrossRef]
88. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
89. Tejero-de Pablos, A.; Nakashima, Y.; Sato, T.; Yokoya, N. Human action recognition-based video summarization for RGB-D personal sports video. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
90. Gautam, S.; Kaur, P.; Gangadharappa, M. An Overview of Human Activity Recognition from Recordings. In Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 12–13 October 2018; pp. 921–928.
91. Nitta, N.; Takahashi, Y.; Babaguchi, N. Automatic personalized video abstraction for sports videos using metadata. *Multimed. Tools Appl.* **2009**, *41*, 1–25. [CrossRef]
92. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8.
93. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
94. Jegou, H.; Douze, M.; Schmid, C. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008*; Proceedings, Part I 10; Springer: Berlin/Heidelberg, Germany, 2008; pp. 304–317.
95. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
96. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
97. Vasudevan, A.B.; Gygli, M.; Volokitin, A.; Van Gool, L. Query-adaptive video summarization via quality-aware relevance estimation. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 582–590.
98. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2616–2625.
99. Fathi, A.; Hodgins, J.K.; Rehg, J.M. Social interactions: A first-person perspective. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1226–1233.
100. Poleg, Y.; Ephrat, A.; Peleg, S.; Arora, C. Compact cnn for indexing egocentric videos. In Proceedings of the 2016 IEEE winter conference on applications of computer vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9.
101. Sun, M.; Farhadi, A.; Seitz, S. Ranking domain-specific highlights by analyzing edited videos. In *Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part I 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 787–802.
102. Mirzasoleiman, B.; Jegelka, S.; Krause, A. Streaming non-monotone submodular maximization: Personalized video summarization on the fly. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
103. Garcia del Molino, A.; Lim, J.H.; Tan, A.H. Predicting visual context for unsupervised event segmentation in continuous photo-streams. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 10–17.
104. Kendall, M.G. The treatment of ties in ranking problems. *Biometrika* **1945**, *33*, 239–251. [CrossRef]

105. Zwillinger, D.; Kokoska, S. *CRC Standard Probability and Statistics Tables and Formulae*; CRC Press: Boca Raton, FL, USA, 1999.
106. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
107. Iashin, V.; Rahtu, E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv* **2020**, arXiv:2005.08271.