systems

MDPI

Review

Video Synopsis Algorithms and Framework: A Survey and Comparative Evaluation

Palash Yuvraj Ingle and Young-Gab Kim *

Department of Computer and Information Security, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: alwaysgabi@sejong.ac.kr

Abstract: With the increase in video surveillance data, techniques such as video synopsis are being used to construct small videos for analysis, thereby saving storage resources. The video synopsis framework applies in real-time environments, allowing for the creation of synopsis between multiple and single-view cameras; the same framework encompasses optimization, extraction, and object detection algorithms. Contemporary state-of-the-art synopsis frameworks are suitable only for particular scenarios. This paper aims to review the traditional state-of-the-art video synopsis techniques and understand the different methods incorporated in the methodology. A comprehensive review provides analysis of varying video synopsis frameworks and their components, along with insightful evidence for classifying these techniques. We primarily investigate studies based on single-view and multiview cameras, providing a synopsis and taxonomy based on their characteristics, then identifying and briefly discussing the most commonly used datasets and evaluation metrics. At each stage of the synopsis framework, we present new trends and open challenges based on the obtained insights. Finally, we evaluate the different components such as object detection, tracking, optimization, and stitching techniques on a publicly available dataset and identify the lacuna among the different algorithms based on experimental results.

Keywords: object detection; object tracking; optimization; synopsis stitching; video surveillance; video synopsis

Citation: Ingle, P.Y.; Kim, Y.-G. Video Synopsis Algorithms and Framework: A Survey and Comparative Evaluation. *Systems* **2023**, *11*, 108. https://doi.org/10.3390/systems11020108

Academic Editors: Tetiana Hovorushchenko, Ivan Izonin and Hakan Kutucu

Received: 26 January 2023
Revised: 13 February 2023
Accepted: 14 February 2023
Published: 17 February 2023

1. Introduction

With technological advancement and internet connectivity, a massive amount of multimedia data are trafficked today via thew World Wide Web. Complex frameworks have been proposed to deal with the analysis and management of these data. Such frameworks represent the amalgamation of different techniques that can ensure data quality and security. Today, video surveillance technology is an intelligent information technology that monitors public space. Recently, there has been massive development and demand for smart video surveillance technology to record daily life activities. With the speedy development of artificial intelligence (AI) technologies, various subsidiary methods are being incorporated into many worldwide applications. Most of the current surveillance technologies are highly dependent on an AI techniques to ensure better efficiency and effectiveness. With each passing day, the amount of generated video content doubles, leading to scarcity in terms of storage requirement. The International Data Corporation (IDC) has issued a statistical report showing that global data throughput is expected to increase significantly, to approximately 175 zettabytes by 2025 [1], with video surveillance data providing the largest contribution. In a traditional video surveillance system [2], a single human operator is responsible for analyzing the video content, which is a tedious, time-consuming, and error-prone job [3]. As the operator is responsible for viewing the content of multiple video cameras simultaneously, many relevant or anomalous activities are skipped. Furthermore, the content gathered in most video surveillance scenarios is redundant, while the essential activity is very limited.

Numerous video condensation methods have been proposed to deal with the above-mentioned issues, such as video summarization and video synopsis. Video summarization deals with creating a summary of video content, thereby creating a condensed video in a timeline. Video summarization can be stated as extracting the key scenes (i.e., keyframes or custom frames) from the original footage in a time sequence [4,5]. The most prominent video condensation technology is video synopsis, first presented in 2006. In video synopsis, the video is abstracted in both the time and space domains. Thus, video synopsis is defined as the time and space domain shifting of extracted foreground objects on a common background. Video synopsis is more effective than video summarization approaches, as it provides a more detailed and smaller video for analysis. Video synopsis projects more than one object in the same space at a given interval of time, preserving the essential activity in the original video. Activity that occurs at different time sequences in the video is shifted in the time domain, and these shifted activities are projected simultaneously in the space domain. Therefore, the object of interest is accessed quickly as the synopsis dramatically decreases the video's length and storage space. Video synopsis has long been an attractive research area in the field of computer vision. Several diverse publicly available datasets incorporating different challenging scenarios have been published, such as KTH [6], WEIZMAN [7], CAVIAR [8], PETS 2009 [9], Hall Monitor [10], Daytime [11], and F-building [12]. New methods have been proposed to adhere to outstanding synopsis challenges such as object detection, energy optimization, tube generation, and stitching.

In the past few decades, video surveillance data processing has seen a tremendous amount of research, mainly in video synopsis and summarization. The increase in surveillance camera connectivity via cheap internet and corresponding advanced technologies such as artificial intelligence, cloud storage, machine learning (ML), and deep learning (DL) have been beneficial to the overall growth in this research field due to the need to deal with the massive amount of surveillance data. Using these methods, more complex video synopsis frameworks have been proposed for generating condensed video for analysis. Several previous survey papers have been published on video synopsis [13–15].

Our motivation for this literature review is to analyze the different methodologies and discover insights based on experimental evaluation. We tried to answer the following questions. *How is the proposed study different from existing survey papers?* All of the existing survey papers highlight a comprehensive review of the synopsis method and its usage; however, they have not evaluated the performance of these methods on a standard dataset in order to clearly define the strengths and weaknesses of different studies. *What are the different techniques on which the performance of a synopsis method is dependent?* The synopsis framework is composed of different steps, such as detection, optimization, collision, and stitching; the generated synopsis is directly affected by these methods. *What are the different frameworks used in video synopsis, and how do different studies evaluate these frameworks through common evaluation metrics and datasets?* We present the most common synopsis frameworks, then discuss and evaluate each component using popular datasets and evaluation metrics. We then classify and evaluate existing studies based on their application usage and research impact.

In this article, we are interested in following the various trends and challenges involved on different synopsis scenarios. As a result, this survey provides researchers with a detailed performance review of all the video synopsis framework components and their respective strengths and weaknesses. Furthermore, we conduct quantitative and qualitative analyses of studies from the initial years of research until 2022. The main contributions presented in this study can be summarized as follows:

- Based on video synopsis usage scenarios, we put forward three different synopsis frameworks, then present a taxonomy of video synopsis techniques along with their respective steps.
- We clearly define the lacuna and complexity of the existing studies based on a comprehensive comparison of various current techniques (i.e., object detection, object

tracking, stitching algorithms), then perform an evaluation through experimentation on publicly available datasets.

- This is the first survey paper to study video synopsis in the context of distinguishing different performance methodologies. Compared with the existing reviews, in the article we focus on determining the most effective video synopsis methods, rather than on describing all types of methods.

The remainder of this paper is organized as follows: Section 2 provides a detailed classification of different synopsis techniques; Section 3 explains the existing synopsis frameworks and their methods; Section 4 provides a brief experimental analysis and comparison of these various methods along with a description of the dataset and evaluation metrics; Section 5 focuses on the new trends and their challenges; finally, Section 6 concludes the paper.

2. Classification of Video Synopsis Techniques

In general, video synopsis techniques have a number of standardized properties in common, which can be quantified as follows: (a) the video synopsis should contain the maximum activity with the least redundancy; (b) the chronological order and spatial consistency of objects in space and time must be preserved; (c) in the resultant synopsis video, there must be minimal collision; and (d) the synopsis video must be smooth and able to permit viewing without losing the region of interest. As depicted in Figure 1, we classify the different video synopsis techniques as follows: keyframe-based, object-based, action-based, collision graph-based, and abnormal content-based.

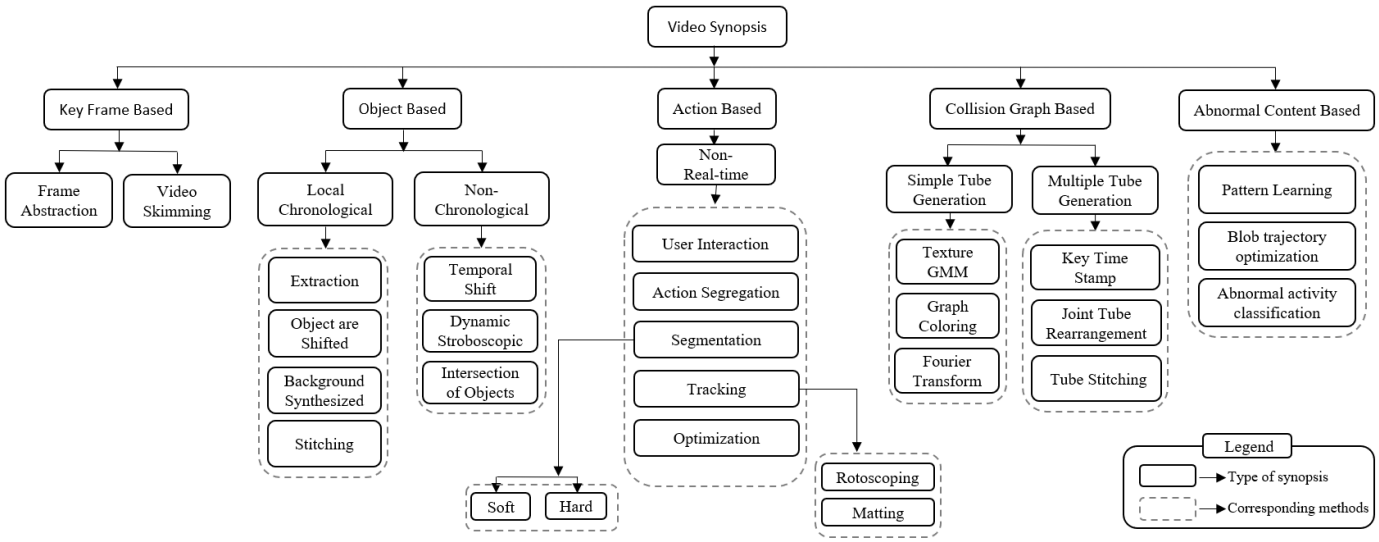

Figure 1. A taxonomy of video synopsis techniques and their properties.

2.1. Keyframe-Based Synopsis

In keyframe-based techniques, frames play an essential role in constructing a video synopsis. These techniques can be classified into two corresponding methods: frame-based approaches and video skimming. In frame-based techniques, a video is built from the necessary keyframes [16]. For example, Choudhary et al. [17] built an offline stroboscopic synopsis. A background can be constructed for stitching the extracted foreground, then aligned using a clustering tracking algorithm [18]; every single frame is used in this process. Pritch et al. [19] proposed an approach in which similar activities are clustered using a k-nearest neighbor (KNN) method. Wang et al. [20] proposed a method for storing and browsing the synopses using flags by incorporating a detail-based algorithm to map different frames. Standardized datasets such as F-Buildings, Hall-Monitor, and Daytime were used to evaluate the proposed detail-based synopsis, showing satisfactory results on a static background. In

addition, a fast-forward method [21] has been developed to minimize the loss when dropping frame activities. Dealing with each frame is a tedious task; instead of extracting each frame, Smith et al. [22] developed a video skimming method that extracted smaller essential video clips from the source video in order to construct a shorter video, ignoring the less critical video clips. The frame-based approach is simpler compared to the video skimming approach; however, this method's computational cost is very high, and there can be significant loss of activity in the resulting video, leading to footage that is unrealistic.

Table 1 summarizes the studies referred to in Section 2.1. We provide a comparison of the properties associated with single-camera and multiview camera approaches to video synopsis, on the basis of which we highlight their insightful pros and cons. We evaluated the parameters based on these classification and insights. The first parameter indicates the deployment type, the second provides the viewpoints, the third determines the summary generation type and visualization (i.e., static or dynamic) concerning the best view, The fourth and fifth dictate the corresponding lacuna and the traditional method's time complexity, respectively, and finally the sixth parameter states the application.

Table 1. Comparison of keyframe-based synopsis techniques.

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Choudhary et al., 2008 [17]	✓		✓		1	S		Stroboscopic, Background Subtraction	Unable to deal with illumination and clutter effects.	$O((NKD)^3)$	Video Indexing
Pritch et al., 2009 [19]	✓		✓		1	S	✓	K-Nearest Neighbors, Temporal Shifting	Substantial number of frames get dropped causing flickering effect.	$O(d*n*log(n))$	Non-Chronological
Wang et al., 2011 [20]	✓		✓		1	S		Simulated Annealing (SA)	Method suffers from occlusion and memory inefficient.	$O(n4)$	Video Browsing

2.2. *Object-Based Synopsis*

In object-based techniques, moving objects are extracted in spatiotemporal space. *Single view camera platforms*: Initially, Pal et al. [23–25] proposed a novel approach that draws out only essential activities from discrete-time sequences instead of selecting entire video frames. In another study, Kang et al. [26] implemented object extraction in space and time; for alignment of this object sequencing, they used a graph cut algorithm and energy optimization. This method's major problem is that unwanted seams emerge in the resulting video, leading to misinterpretation of the content. These techniques can be further classified based on whether they follow local chronology or are non-chronological. In 2006, based on this concept, Rav-Acha et al. [27] pioneered the idea and first coined the term "video synopsis". In their study, they used a low-level optimization concept based on 3D Markov Random Fields (MRF) [27]. First, they detected actions and tracked movements for storage in a queue; after alignment of these activities, they applied the concepts of background generation and stitching, whereby the objects were moved in space and time using simulated annealing (SA). In object-based synopsis, the solution is to predetermine the video's synopsis length in order to feasibly minimize energy usage problems. When rearranging objects were without maintaining the local chronology, collision costs can be very high. Thus, the constraint of this approach is that it creates a video synopsis only for a specified period. To overcome this constraint, Pritch et al. [28] constructed a synopsis for an endless video stream. In their methodology, a min-cut [28] algorithm is used to extract objects in the form of a tube, and the local chronology is distributed as the tubes are moved temporally. Though the aforementioned authors pioneered this field of

study, they experienced multiple serious limitations, such as colossal memory consumption and issues with activity density. To deal with the problem of memory consumption, non-chronological synopsis has been proposed.

Xu et al. [29] solved the optimization problem by implementing set theory to maximize visual content in synopsis videos. Their proposed method outperformed previous methods. In another study, Pritch et al. [30] employed non-chronological synopsis to combine activities from different time zones. Only moving objects were considered when generating the synopsis. In a comprehensive study, Wang et al. [31] applied region of interest (ROI) information for faster browsing of surveillance video. Based on intraframe coding, they labeled each region of interest, significantly boosting the scale browsing in video synopsis. They tested their methodology on the F-Building, Daytime, and Hall Monitor, showing slightly better performance on the latter. A similar strategy was used by Sun et al. [32]; they obtained background modeling and tracking of the event of interest, then used the maximum motion power to generate a summary. Zhu et al. [33] contributed by improving the selection criteria for generating a tube. The resulting video synopsis method obtained a higher compression ratio compared to previous methods. Random surveillance videos from the PETS dataset were used to test their methodology on both single-camera and multi-camera networks. Unlike traditional temporal shifting, Nie et al. [34] shifted the temporal and spatial axis of the activity when constructing a video synopsis. They expanded the background in order to avoid collision and fit the objects. As the object is pulled along both the temporal and spatial axes, these approach involves several other challenges, such as changes in the background that make understanding of the view more difficult and the resulting synopsis being too dense for analysis. Thus, this method is only applicable in limited-view scenarios.

Yao et al. [35] proposed an object-based video synopsis method to tackle collision problems using a multi-target tracking approach. They tested their method on an indoor video surveillance dataset, where they faced errors such as moving object detection and tracking. Olivera et al. [36] published an open-source library for constructing video synopses. Their study contributed by providing a tool for creating video summaries by automatically extracting the objects through background subtraction and segmentation. Their method is simple, and is applicable for simple video synopsis generation; however, it shows poor performance in crowded video sequences. Tian et al. [37] implemented a similar temporal shift approach, with the difference that they broke down long-term moving objects into segments. Ahmed et al. [38] were able to overcome the problem of multiple trajectories while creating the synopsis. Their study used two publicly available datasets, namely, VIRAT [39] and in-house KIST; using these datasets, their method was able to construct a meaningful synopsis. Due to unchronological tube shifting, however, this approach can cause chaotic collisions. To overcome this, Yi et al. [40] suggested spatiotemporal event rearrangement of objects. They tested this idea on a minimal video sequence. However, the results continued to showed collision events in more extensive video sequences. To address this problem, Li et al. [41] applied comprehensive video synopsis based on different scenarios. They extracted whole video clips containing large crowds, and were able to obtained synopses with fewer collisions and less overlapping. In solving the energy minimization problem, simulated annealing plays a vital role; however, it suffers from high computational costs. To solve this issue, Ghatak et al. [42] proposed a hybrid of simulated annealing and teaching–learning-based optimization (TLBO). However, their study mostly focused on improving energy optimization. They evaluated their study on the PETS, MIT, and UMN datasets, on which they achieved significantly better performance than previous studies. Their model used the traditional Kalman filter for multi-object tracking, which is a significant drawback of this study due to the high pre-processing time it requires. A method for accurately detecting and extracting objects to create a video synopsis using deep learning was implemented by Mona et al. [43]. Their study used a convolution neural network (CNN) based on You Only Look Once (YOLOv3) [44] to detect and extract the object. However, this approach suffers from high computational complexity and time con-

sumption when creating the synopsis. Nevertheless, the authors found that it significantly outperformed a genetic algorithm in comparative testing on the VIRAT dataset.

All the studies mentioned above used offline-based methodologies. To perform video synopsis in real-time, Yildiz et al. [45] proposed pixel-based analysis instead of dealing with the entire video frame. To this end, they extracted only those video clips with a high degree activity. Similarly, Vural et al. [46] applied a pipeline-based framework for constructing real-time video synopses with low memory consumption. An online approach for background selection and synopsis generation was developed by Feng et al. [47]. Huang et al. [48] proposed a method for making object detection and tracking possible in real-time using a table-driven method. In another study [49], the same authors proposed using online synopsis tables to maintain the chronology of the extracted tubes, and incorporated maximum posterior estimation to ensure the tubes' alignment. However, when analyzing the resulting synopses the video suffered from low visual quality, especially in dense activity scenarios, due to the approach they used being pixel-based. Sun et al. [50] formulated a map-based online synopsis generation technique to improve the visual quality of the generated synopses. Using a complex tree algorithm, Hsia et al. [51] implemented video retrieval, which they found to be an efficient method for constructing synopses. In both studies, there was a significant drop in the number of object frames. Fu et al. [52] considered activity relationships and optimization while proposing a real-time video synopsis framework (RTVS). In another study, Ghatak et al. [53] showcased a hybridization of the SA and JAYA algorithm to improve energy minimization. Chen et al. [54] incorporated a CNN-based methodology to detect and extract the required object and integrate it with a collision algorithm in order to handle local transparency and avoid collisions. For better visualization of obtained synopsis video, Namitha et al. [55] suggested an interactive visualization model in which the synopsis is constructed based on user requirements. Their study smartly formed user queries using both temporal and spatial attributes. All of the studies mentioned above only performed synopsis based on a single-view camera or single input video sequence. Kostadinov et al. [56] implemented an ML model to extract the background, with objects subsequently being localized and segmented based on timestamps to constructing the video synopsis. They divided the entire process into two phases, namely, analysis and generation. Li et al. [57] proposed an infrared video synopsis framework (IVSF) to construct a video synopsis from an infrared video, mainly utilizing image similarity in the space and time domain to minimize the space ratio in order to create a summary.

Multiview camera platforms: Most surveillance systems encompass multiple camera inputs. Zhu et al. [58] proposed a multi-camera joint video synopsis in which objects are selected based on the key timestamp. They performed object reidentification to maintain the chronological order of items for both one camera and multiple cameras using the key timestamp. Hoshen et al. [59] suggested a similar strategy to implement a master–slave camera approach. Preprocessing was carried out for detection at the level of the master camera, while the slave camera was responsible for extracting the tube sequence for that period. Instead of focusing on master camera processing, Mahapatra et al. [60] proposed a multiview video synopsis by combining the features of video summarization and video abstraction. They applied each camera's field of view (FoV) on a standard background surface. They set the detection priorities based on seven items (e.g., running, walking, waving, jumping). Their study was able to create a summary only for these activities. Zhang et al. [61] implemented joint object stitching and camera view stitching to provide a more compact and understandable synopsis in order to overcome issues with overlapping FoV. First, they synchronized the input video by grouping similar activities, then shifted the entire grouped activity along the time axis to obtain a multiview camera synopsis. As this scenario involves a single object being viewed by multiple cameras, the complexities involved in optimization are greatly increased. Xie et al. [62] considered locating the camera's position and the field of view, thereby helping to create an image observability model responsible for obtaining a synopsis of a geographic scene. They proposed a geospatial video synopsis framework (GSVSF) for multiple virtual viewpoints;

however, their study applies only to quite specific scenarios. Priyadharshini et al. [63] implemented a spherical video synopsis framework (SVSF) in which they considered a 360-degree FoV. Instead of creating a synopsis for all the objects, they selected only crucial items based on user requirements, which they achieved using an action recognition model. on the whole, multi-camera synopsis approaches are widely accepted for real-world surveillance systems. An insightful analysis of object-based synopsis techniques is presented in Table 2 for Section 2.2.

Table 2. Comparison of object-based synopsis techniques.

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Rav-Acha et al., 2006 [27]	✓		✓		1	S		Markov Random Field (MRF), Graph Cut	For long time video synopsis, occlusion of objects is observed.	$O(mlog3n)$	Low level Synopsis
Pritch et al., 2007 [28]	✓		✓		1	S		SA	This framework fails to compute synopsis for moving cameras.	$O(n4)$	Query based Synopsis
Yildiz et al., 2008 [28]		✓	✓		1	S		Nonlinear Image, Dynamic Programming (DP)	Computationally expensive (CE) and space complexity is more.	$O(VE)$	Real-time Synopsis
Xu et al., 2008 [29]	✓		✓		1	S		Mean Shift Algorithm	Spatial dimension has not been considered while designing the framework.	$O(Tn2)$	Video Synopsis
Pritch et al., 2008 [30]	✓		✓		1	S/D		ObjectDetection (OD), Tube Generation (TG)	This technique is not applicable to a video with dense activity.	$O(ElogV + VlogV)$	Video Indexing
Vural et al., 2009 [46]		✓	✓		1	S		Frequency Background Subtraction, DP	Information loss 3D to 2D projection, mount eye-glaze camera challenging.	$O(VE)$	Real-time Synopsis
Feng et al., 2010 [47]		✓	✓		n	S		Background Subtraction	This study is not applicable for crowded scenario and is CE.	$O((NKD)^3)$	Online Synopsis
Wang et al., 2012 [31]	✓		✓		1	S/D		Object Region Flag	Method suffers from occlusion and memory inefficient.	$O(VE)$	Scalable Browsing
Sun et al., 2011 [32]	✓		✓		1	S		Maximum Motion Power	Cannot work with motion cameras. Also, illumination and cluttering effect not minimized.	$O(N)$	Video Synopsis
Huang et al., 2012 [48]		✓	✓		1	S		Object Tracking, Table Driven Approach	Flickering effect and occlusion can be majorly observed.	$O(VE)$	Online Synopsis
Sun et al., 2012 [50]		✓	✓		1	S		Map-Based Optimization	The obtained synopsis is densely condensed creating confusion.	$O(nlogn)$	Online Synopsis
Zhu et al., 2012 [33]	✓		✓		1	S		Key observation	Problem of occlusion arises since spatial dimension is neglected.	$O(n4)$	Video Synopsis
Nie et al., 2013 [34]	✓		✓		n	S	✓	Alpha-Beta Graph Cut (ABGC)	Incapable to work with moving cameras.	$O(b^{\lfloor}d/2))$	Compact Synopsis
Hsia et al., 2013 [51]		✓	✓		1	S		Low Complexity Range Tree (LCRT)	Computationally expensive as well as occlusion can be noticed.	$O(logn + k)$	Retrieval
Huang et al., 2014 [49]		✓	✓		n	S/D		Maximum Posteriori Estimation	Space complexity and occlusion is highly noted.	$O(n4)$	Real-time Synopsis
Yao et al., 2014 [35]	✓		✓		1	S		OD, Object Tracking, Genetic Algorithm	Cannot detect and track continuously moving object. Thus, frames droped.	$O(gnm)$	Video Synopsis
Fu et al., 2014 [52]		✓	✓		n	S		Motion Structure, Hierarchical optimization	CE and does not support crowded videos.	$O(VE)$	Real-Time Synopsis
Zhu et al., 2015 [58]	✓	✓		✓	n	D	✓	Joint Tube Generation	Obtained video is confusing and redundant.	$O(ElogV + VlogV)$	Joint Synopsis
Olivera et al., 2015 [36]	✓		✓		1	S		Open source library	The resultant output suffers from jittering, flickering effects.	$O(VE)$	Video Synopsis
Hoshen et al., 2015 [59]	✓	✓		✓	n	D		TG, SA	Occlusion and jittering effect is observed. Moreover, frame drop.	$O(n4)$	Live Video Synopsis
Mahapatra et al., 2015 [60]		✓		✓	n	D		Clustered Track, Collision Detection	CE and chronology of objects is not maintained.	$O(ElogV + VlogV)$	Multiview Synopsis
Tian et al., 2016 [37]	✓		✓		1	S		Genetic Algorithm	Occurance of illumination and cluttering effect and CS.	$O(gnm)$	Video Synopsis
Ahmed et al., 2017 [38]	✓		✓		1	S		TG	Computationally expensive and the output is confussing.	$O(n4)$	Video Synopsis

Table 2. *Cont.*

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Yi et al., 2018 [40]	✓		✓		1	S		Spatio temporal	Computationally expensive and cannot handle illumination.	$O(log2(n))$	Video Synopsis
Li et al., 2018 [41]	✓		✓		1	S	✓	Group Partition, Greedy Approach	CE and does not support moving cameras.	$O(ElogV + VlogV)$	Video Complex Synopsis
Ghatak et al., 2019 [42]	✓		✓		1	S		HSATLBO	Framework dissents moving cameras and several frames are lost.	$O(VE)$	Video Synopsis
Zhang et al., 2020 [61]	✓		✓		n	S/D		Spatio-Temporal, Dynamic Programming	Browsing is not scalable and merging of objects can be seen.	$O(VE)$	Multiview Synopsis
Mona et al., 2020 [43]	✓		✓		1	S	✓	Yolo3, Swarm Algorithm	High memory consumption and numerous frames are dropped.	$O(nlogn)$	Video Synopsis
Ghatak et al., 2020 [53]	✓		✓		1	S		HSAJAYA	Quality of the video is compromised.	$O(nlogn)$	Video Synopsis
Chen et al., 2020 [54]	✓		✓		n	S		Attention-RetinaNet, Local Transparency	Computationally expensive and time consuming.	$O(logn)$	Video Synopsis
Nanitha et al., 2021 [55]	✓		✓		n	S/D	✓	Joint Tube Generation	High memory consumption and occlusion of object is observed.	$O(ElogV + VlogV)$	Video Synopsis
Kostadinov et al., 2022 [56]	✓		✓		n	S	✓	Object localization, Object tracking, reidentification	Resource intensive task thus consume large memory, flickering.	$O(n4)$	Video Synopsis
Xie et al., 2022 [62]	✓			✓	n	S/D		Video Spatialization, Spatiotemporal pipeline	CS as it deals with locating the camera position.	$O(ElogV + VlogV)$	Geospatial Synopsis
Li et al., 2022 [57]	✓		✓			S/D		Fourier Transform, Object tracking	Occlusion and jittering effect is observed.	$O(nlogn)$	Infrared Video synopsis
Priyadharshini. 2022 [63]	✓			✓	n	S/D	✓	Action recognition module, Tracking	High memory consumption and occlusion of object is observed.	$O(VE)$	Spherical video Synopsis

2.3. Action-Based Synopsis

Action-based synopsis is a technique that focuses only on extracting an object of interest in motion in order to construct a short video. In the object-based method, all of the moving or non-moving features are considered in the summary, potentially leading to more redundancy and higher computational cost. In action-based synopsis, action segregation and alignment are first used to extract the object, thereby reducing the redundancy. Finally, stitching and optimization are incorporated to overcome the high computational cost while shifting the object in the domain space. Hao et al. [64] implemented a GrabCut segmentation algorithm applied to a moving object matting sequence. Their study used user interaction to select the desired object; after selecting the object, GrabCut segmentation was used to create a synopsis of that object. The major drawback of this approach is that it requires user interaction to create the summary.

User interaction can be reduced by techniques such as rotoscoping and matting [65]. Similarly, Nie et al. [66] decomposed objects into several segments, with each segment corresponding to an action. Non-active elements were discarded during the process, and the selected action segments were stitched together to create a shorter compact video. In this approach, segmentation and action tracking take place with the help of hard and soft segmentation [67,68]. After the user has selected the object by drawing a curve, the object is rotoscoped using the timeline. After action segregation, the object is repaired to address any holes with respect to the background. When combining the action segments, the authors maintained a chronology of the action. The action parts were shifted using the vector, with the linear combination representing the energy function. Stitching was performed using a thinning algorithm applied to the pixel number and width. As a result, shorter videos can be obtained as compared to the original video sequence. This study's major drawback is that it is not applicable to crowded scenarios with more than one action object. A detailed analysis of action-based synopsis techniques is provided in Table 3 for Section 2.3.

Table 3. Comparison of action-based synopsis techniques.

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real-Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Hao et al., 2013 [64]	✓		✓		1	S		Grab Cut, Object Segmentation	Does not support multi-camera view and the quality is low.	$o(nlogn)$	Video Synopsis
Nie et al., 2014 [66]	✓		✓		1	S		MRF	Technique cannot be applied on moving cameras.	$O(mlog3n)$	Video Synopsis

2.4. Collision Graph-Based Synopsis

Differing from convention synopsis techniques, in this approach the tubes are shifted in order to reduce computational complexity. One example is a study Lu et al. [69] in which the authors proposed fluent tube generation by implementing two methods, namely, the Gaussian mixture and texture-based methods. While creating the tubes, they removed the shadows from the foreground, then used a filter to concatenate the tubes. The resulting synopses had better visual effects. In another approach, Wang et al. [70] shifted the object using background modeling and foreground segmentation. Their study provided scalable browsing and efficient synopsis generation. Similarly, Zhong et al. [71] proposed fast synopsis using compressed video; they abstracted tubes from the video using a graph-cut algorithm to perform parallel minimization on the energy function [72,73]. Their test result on the F-Building, Hall Monitor, and Daytime datasets showed better tube extraction with this approach. He et al. [74] mentioned a tube rearrangement technique to reduce potential collisions. Collision of tubes takes place when tubes occur at the same time and in the same space. They identified tube collisions by incorporating the collision relationship probability [75,76] to help determine the tube position [77–79]. However, while these techniques can reduce the computation cost and collision artifacts [80,81], they are challenging to implement on dynamic backgrounds. Nie et al. [82] incorporated attributes such as object size and speed in order to avoid collisions in the resultant synopsis, for which they used three variable optimization methods. An in-depth analysis of collision graph-based synopsis techniques is provided in Table 4 for Section 2.4.

Table 4. Comparison of collision graph-based synopsis techniques.

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real-Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Lu et al., 2013 [69]	✓		✓		1	S		Gaussian Mixture Model (GMM), TG	Synchronization and alignment of the tube is not seen.	$O((NKD)^3)$	Video Synopsis
Wang et al., 2013 [70]	✓		✓		1	S		Flag-Based, SA	Computationally expensive and loss of pixels.	$O(n4)$	Video Indexing
Zhong et al., 2014 [71]	✓		✓		1	S	✓	Graph Cut, SA	Cannot work with regular vidoes, movie and TV video.	$O(mlog3n)$	Fast Analysis
Li et al., 2016 [80]	✓		✓		1	S		TG, Greedy Approach	Chronology is not maintained and performance drop can be observed.	$O(ElogV + VlogV)$	Effective Synopsis

Table 4. *Cont.*

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real-Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Li et al., 2016 [83]	✓		✓		1	S		Temporal Domain, SA	Spatial domain is compromised giving rise to occlusion and frames are dropped.	$O(n4)$	Video Synopsis
Jin et al., 2016 [81]		✓	✓		1	S	✓	Projection Matrix	Quality is not up to mark and time consuming.	$o(nlogn)$	Real-Time Synopsis
He et al., 2017 [74]		✓	✓		1	S		Collision Graph	Computationally expensive and loss of frames.	$O(V+E)$	Online Video Synopsis
He et al., 2017 [75]	✓		✓		1	S		Graph Coloring	Chronological order, motion structure, activity preserving are compromised.	$O(m^V)$	Video Synopsis
Liao et al., 2017 [76]	✓		✓		1	S		3D Graph Cut	Computationally expensive and data lost can be seen.	$O(mlog3n)$	Synopsis Browsing
Ra et al., 2018 [77]		✓	✓		1	S		Fast Fourier Transform (FFT)	Computationally expensive and slow.	$o(nlogn)$	Real-Time Synopsis
Pappalardo et al., 2019 [78]	✓		✓		1	S		Graph Coloring	Object tracking and detection are not considered.	$O(m^V)$	Video Synopsis Dataset
Ruan et al., 2019 [79]		✓	✓		1	S/D		Dynamic Graph Coloring	Computationally expensive and time consuming.	$O(m^V)$	Online Video Synopsis

2.5. Abnormal Content-Based Synopsis

The abnormal content-based synopsis strategy is an application-specific method in which the abnormal case for constructing the synopsis is predefined. It only deals with shifting all abnormal foreground objects in the time and domain space to create a condensed video. Cho et al. [84] proposed an event-based video synopsis application, using a template-matching scheme to group similar activities. In their approach, they predefined the positions of cameras with entry and exit points. Then, they applied the template-matching scheme to the camera view, using cluster trajectories to detect the abnormal activities. Any event beyond the predefined trajectories is considered an abnormal event. A similar strategy was mentioned by Lin et al. [85]; they detected anomalies by first learning the local patch of object occurrence. Their study used blob sequence optimization to make it easier for the synopsis to display activity. Ahmed et al. [86] trained a CNN model to detect cars, bikes, and pedestrians in order to create a synopsis based on the requirements of user-specific queries. For background and foreground segmentation, they used an improved Gaussian mixture model (GMM) in which multiple object tracking was achieved using a sticky algorithm. Differently, Ingle et al. [87,88] used the LiDAR point cloud and image data to create a synopsis from drone video. Their study was highly reliant on the customized object detection model used to extract the objects; additionally, they used early fusion to perform stitching. Using pre-trained scenarios, these studies contribute a new approach to specific-event synopsis generation. Table 5 summarizes the studies referred to in Section 2.5 along with their characteristics.

Table 5. Comparison of abnormal content-based synopsis techniques.

Existing Studies	Deployment		Viewpoint		Analysis			Methods	Lacuna	Time Complexity	Application
	Non-Real Time	Real Time	Single Camera	Multiple Camera	No of Summary	Visualization	Best View Selection				
Chou et al., 2015 [84]	✓		✓		n	D		OD, Object tracking	Loss of frame. Object detection and tracking are left out.	$o(nlogn)$	Event Synopsis
Lin et al., 2015 [85]	✓		✓		1	S	✓	Local Patch Learning Based Abnormality Detection	Is not applicable to moving cameras and output is not accurate.	$o(nlogn)$	Activity Synopsis
Ahmed et al., 2019 [86]	✓			✓	N	S/D	✓	TG	Does not support crowded data and moving cameras.	$O(ElogV + VlogV)$	Intelligent Traffic

Based on our analysis of existing synopsis applications, Figure 2 showcases a chronological overview categorized as follows: (off-line + single camera view + keyframe-based); (off-line + single camera view + object-based); (off-line + multiple camera view + object-based); (offline + single camera view + collision graph-based); (offline + single camera view + object-based); (on-line + single camera view + object-based); (offline + single camera view + abnormal content-based).

Here, off-line means that the obtained live video feed is first stored on a storage device, then the synopsis process is carried out on the stored data to obtain the condensed video. In the on-line phase, the synopsis process is initiated directly on the obtained live video to construct the synopsis, which avoids the use of local storage space.

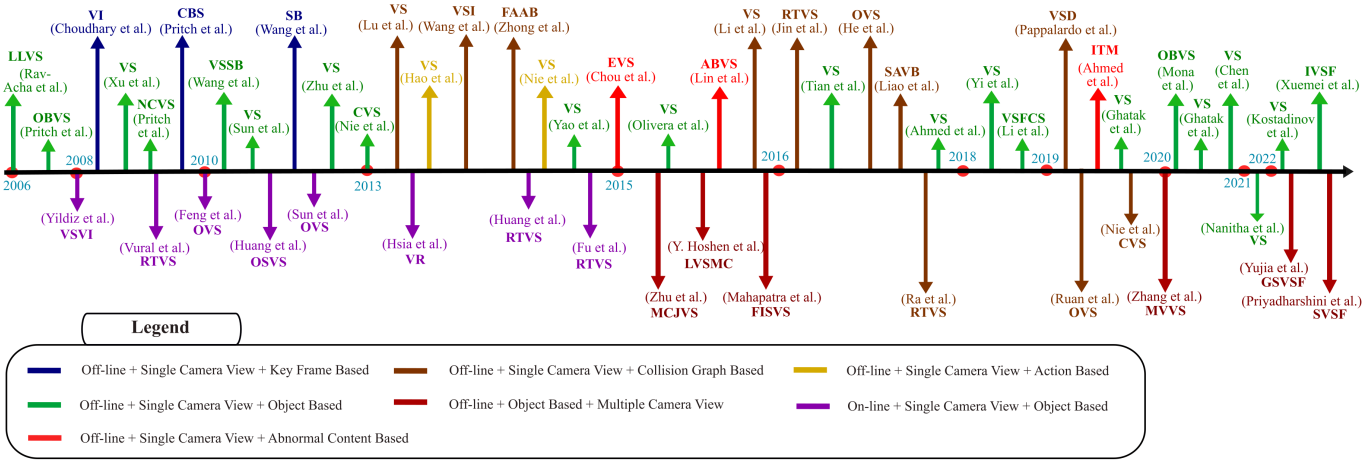

Figure 2. Chronological overview of the most relevant video synopsis studies. The chronology represents the names of the author and the respective timeline of their study.

3. Video Synopsis Framework

This section briefly describes video synopsis methodology components based on analysis and classification. These can mostly be distinguished based on camera view, that is, single-camera or multiple, as well as on anomaly detection pretraining in abnormal synopsis methodology. In the single-camera video synopsis framework, a shorter video is constructed for a single view in which the object is detected, and the extracted foreground is stored based on user query. Optimization and the visualization are carried out to obtain an optimal shorter video. In the multiple-camera video synopsis framework, there are multiple views from which the synopsis is created, and detection and extraction of the object are carried out for every single frame, thereby generating a tube for each object.

These object tubes are shifted in time and domain space, then stitched together with the corresponding background based on the alignment of the video sequence. Finally, blending is performed in the visualization phase to enhance the tube quality of the segmented objects to construct a better video synopsis. In the abnormal content video synopsis framework, a pre-trained CNN model is incorporated to detect the abnormal object only, for which the foreground is then extracted based on the criteria used for the synopsis. Optimization of this abnormal object is carried out along with corresponding background stitching to construct an abnormal video synopsis. A systematic illustration of different methods and their component is depicted in Figure 3.

A detailed explanation of the corresponding components and their methods is described below. Video synopsis generation begins with object detection and tracking to extract activity from the video sequence. In the single camera methodology, object detection and tracking occur only for a single video sequence, whereas in a multi-camera strategy there are multiple video sequences. In the abnormal method, the parameter to be detected is pretrained using a CNN detection model. This characteristic represents a significant difference from other practices. Suppose there is a user query-based interaction; in this case, a synopsis is generated for that object following the optimization process (i.e., rearranging or shifting the items) obtained from the object tracking database. This approach is able to deal with collisions between activities before they are stitched together. The visualization process blends the generated video for better video synopsis visual quality in single-camera methods; on the other hand, in the non-query-based approach a synopsis is created for the entire object.

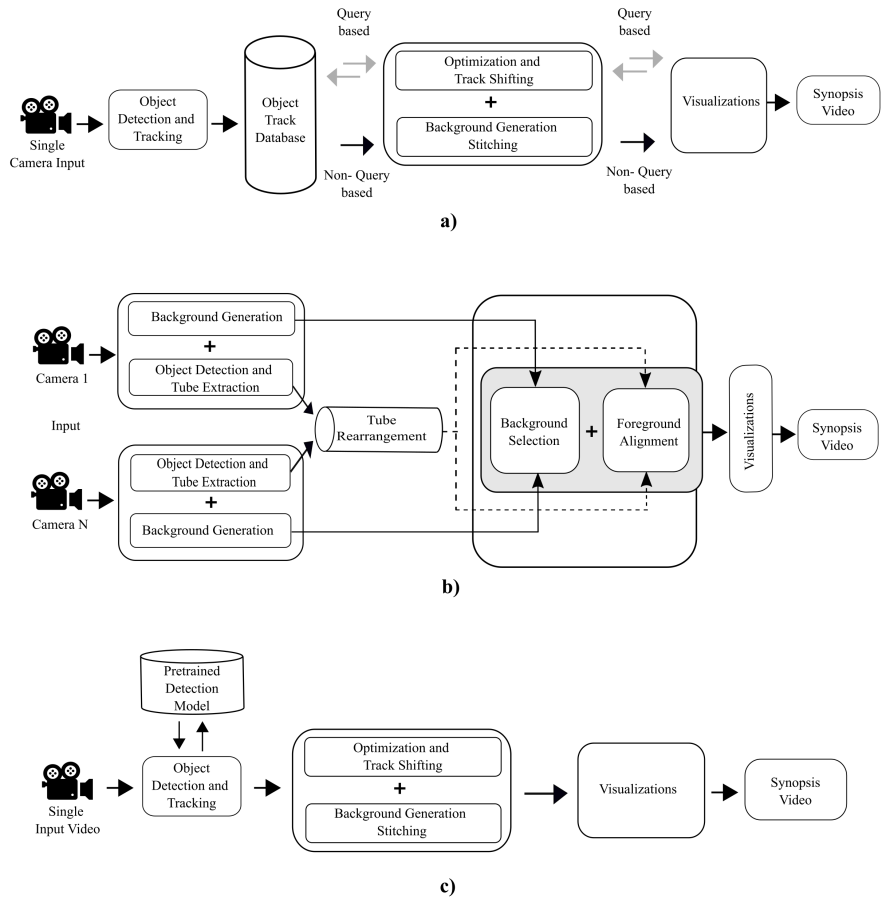

Figure 3. An illustration of different synopsis methodologies and their components: (**a**) single-camera video synopsis framework; (**b**) multi-camera video synopsis framework; (**c**) abnormal content video synopsis framework.

Object detection is an initial preprocessing stage used in obtaining a synopsis; to minimize preprocessing time, the preferred method used for detection is motion detection. Motion detection is a simple method that is initiated when there is a difference between the foreground and the background pixels; different methods of motion detection include the pixel difference [27], background cut [19], GMM [89], Gradient [43], PBAS [90], LOBSTER-BGS [91], and others. The lacuna in this method is their performance in cases with dynamic backgrounds or large crowds, as they cannot detect all of the objects in such scenarios.

As an alternative to motion detection, human detection methods can be used to detect different objects and humans against complex backgrounds. Such methods include CNN [92], Quadtree [93], Min-cut [94], and SILTP [95]. An abnormal activity detection method is applied to detect the anomalies in the video sequence. The pretrained template matching approach is regulated for abnormal activity detection. Object parallel tracking is initiated with detection, and is a critical stage for creating trajectories of frames. Based on these trajectories, objects are aligned in the resultant synopsis. Mathematical approaches such as Euclidean distance [96], Kalman filtering [97], and chi-square distance [98] can be applied in the frame-based tracking method to find the distance between two consecutive frames, ensuring that the frame of interest is tracked accordingly. The clustered track extraction process is initiated to group and track similar activity; on the other hand, entire objects can be tracked the form of a tube sequence method using Fourier transform [99], graph coloring [65], or key timestamp [100] approaches, among others. These methods are used in object-based tracking approaches. Alternatively, an action-based tracking approach can be used to track different actions using neural networks and multiple pedestrian tracking methods. Object tracking directly impacts the generated synopsis results, as any broken trajectories in tracking or collisions involving a track can decrease the performance of the entire methodology.

Optimization/energy minimization is vital for rearranging an object into a sequence with a minimum collision rate. When rearranging objects, it is important to perform object segmentation, which provides the desired object's closed region boundaries, in order to determine the object's position. Among the methods used to perform segmentation are edge segmentation [101], clustering segmentation [102], and region-based segmentation [103]. After determining the position of the activity of interest, it is shifted in the time domain to create a smaller merged video. When shifting the activity, it is necessary to determine certain parameters, such as the consistency and collision. When shifting an object, several different optimization methods can applied; these can be classified into clustering-based (e.g., packing cost [19], film map generation [96], mean shift [28], table-driven [60], etc.), tube-driven (e.g., TLBO [104], SA [105], etc.), tree-based (e.g., greedy approach [30], alpha–beta swap [106], genetic algorithms [107]), and dynamic programming [44] approaches. A detailed performance evaluation is provided below in Section 4. In a multiple-camera framework, two different tubes are extracted and arranged jointly in a common sequence, unlike the single-camera approach. In an abnormal synopsis framework, optimization and stitching are the same as in the single-camera approach.

The next stage in the synopsis process is background generation, followed by stitching, in which the obtained object is stitched with the time-lapse background. Here, it is crucial to generate a smooth background in order to ensure better visual quality of the resulting synopsis. In most of the existing literature, pixel-based rather than feature-based stitching is used. Stitching and background generation does not affect the performance of the final synopsis; rather, these steps are carried out to improve the visual quality. Unlike the single-camera approach, in multi-camera scenarios a common background must be selected before stitching.

We provide a comparison between various stitching algorithms in Table 6; the computational cost is dependent on the stitching method. Additionally, different blending algorithms can be used to improve the viewing quality during the visualization stage.

Table 6. Stitching algorithms used in video synopsis.

Method	Synopsis		Name and Reference	Technique	View Point		Distinguished		Computational Cost		
	Class	Type			Single Camera	Multiple Camera	Pros	Cons	L	M	H
Pixel Based	Frame	Off-line	Peleg et al. [108]	Optical flow	✓		Fast	Low accuracy		✓	
	Object / Action		Zhi Q et al. [109]	Depth and color	✓		Degree of depth	Complicated calculation			✓
			Uyttendale et al. [110]	Graph structutre	✓	✓	Eliminate ghosting	Complicated calculation			✓
Feature Based	Frame		Brown et al. [111]	Sparse matching	✓		Automated	Limited plane	✓		
			Lin et al. [112]	Varying affine	✓		Address parallax	Single affine		✓	
			Liu et al. [113]	Insertion view	✓		Degree of parallax	Complicated calculation	✓		
			Chang et al. [114]	Tranformation	✓	✓	Overlapping Region	Limited to parallel	✓		
	Object / Action		Li et al. [115]	Homography	✓	✓	Reduce distortion	Limited to parallel			✓
			Chen et al. [116]	Coarse fine	✓		Rotation correction	Local distortion			✓
			Zhang et al. [117]	Prior constraints	✓		Wide baseline	Complicated calculation			✓
			Xiang et al. [118]	Level Feature	✓	✓	Degree of texture	Local distortion			✓
	Object / Action / Collision	On-line	Rav-Ach et al. [119]	Embed the object	✓		Accurate alignment	Limited to camera		✓	
			Su et al. [120]	Optimization function	✓		Balance stabilization	Complicated calculation	✓		
			Nie et al. [121]	Background foreground	✓	✓	Improved matching	Complicated calculation	✓		✓
			Lin et al. [122]	Estimate parameter	✓		3D path	Limited to depth			✓

L-Low, M-Moderate, H-High.

In this section, we have explained the video synopsis framework components and classified the various methods involved. Additionally, we have provided a qualitatively analysis of different stitching algorithms and mentioned their respective pros and cons with respect to computational cost. The next section evaluates object detection, extraction, and optimization techniques used in state-of-the-art video synopsis methods.

4. Results and Discussion

We conducted exhaustive experiments using a standard dataset to evaluate state-of-the-art synopsis methodologies. An AMD Ryzen 5 3500X equipped with 16 GB RAM and an Nvidia Gigabyte GeForce RTX 2060 graphic card was used for experimentation. In testing, as a front-end we used Python programming language for most of the studies, while for others we used MATLAB version 2019a with C++. As the synopsis framework is composed of different components, each one of them was evaluated on the respective dataset with a standard metric. We analyzed and evaluated the state-of-the-art object detection, tracking, and optimization methods used in video synopsis, as discussed in Section 3. We tested this method on the five videos from the Hall Monitor dataset; the evaluation metric is mentioned in Section 4.1.2. We carried out this analysis in order to draw an outline of these methods. In all the videos, several humans are walking randomly from left to right. The video contains a single view and a static background with multiple objects. Finally, we provide a separate discussion of the experimental outcomes.

4.1. Datasets and Metrics

This section summarizes the different publicly available datasets for video synopsis, their respective challenges, and their evaluation methods.

4.1.1. Datasets

A diverse number of video surveillance datasets are available publicly. However, most video synopsis techniques are evaluated on a local dataset, which is typically not publicly available. Table 7 shows the list of datasets used for object detection and object tracking and segmentation in this study.

Table 7. A summary of existing datasets used for object detection and tracking, segmentation, and creation of video synopses.

Dataset	Year	View Type	Scenes	No. of Views	Application
PETS	2000	Single/multi	In/Outdoor	1, 2	activity monitoring, tracking, segmentation
WEIZMANN	2001	Single-view	Outdoor	1	detection, temporal segmentation
KTH	2004	Single-view	In/Outdoor	1	feature extraction, synopsis
CAVIAR	2007	Multi-view	In/Outdoor	1, 2	activity monitoring, tracking, segmentation, clustering
Hall Monitor	2014	Single-view	Indoor	1	object detection, tracking, segmentation, synopsis
Day-Time	2014	Single-view	Indoor	1	object detection, tracking, segmentation, synopsis
F-Building	2014	Single-view	In/Outdoor	1	object detection, tracking, segmentation, synopsis

PETS is a performance evaluation tracking and surveillance dataset, created in 2000 to evaluate tracking algorithms. All the video sequences in the PETS dataset are manually labeled using the bounding box to locate the objects. WEIZMANN is an event-based dataset created in 2001, and is specifically designed for evaluating different clustering and segmentation algorithms using a statistical measure; the dataset mainly contains video sequences with 6000 frames. It includes actions such as waving, running, and walking.

The KTH dataset was created in 2004; at that time, it was the most extensive human action dataset. The dataset contains indoor and outdoor video sequences, and includes walking, waving, jogging, running, boxing, and clapping actions. The CAVIAR dataset, created in 2007, consists of 80 indoor videos representing various gestures and positions, such as fighting, walking, shopping, etc. The Hall Monitor, Daytime, and F-Building datasets were created in 2014, and all contain indoor/outdoor video events that mainly include a static background with limited movement activities such as walking across the street or walking in an office building corridor.

4.1.2. Evaluation Metrics

Video synopsis performance is evaluated based on the different synopsis methodology stages, such as object detection and tracking, energy minimization, and computational cost. The metrics are precision, recall, F1 measure, similarity, frame condensation ratio (FCR), collision cost (CC), temporal consistency cost, chronological disorder ratio (CDR), and time of execution [123]. The precision metric is used to determine the accuracy of object prediction. In contrast, recall indicates the accuracy of detection based on the total number of objects, and the F1-score measures the test accuracy.

The similarity measure quantifies the similarity between two objects. FCR determines the total number of frames in the synopsis to that in the source video; the higher the frame reduction, the lower the FCR. CDR represents the total number of chronological disorder frame activities compared to the total number of activities. The smaller the CDR value, the better the chronological order in the synopsis video. The time of execution is determined based on the type, online or offline; it indicates the time required to create the synopsis video, which depends on the type.

Very few studies have evaluated their methods based on the video quality or camera usage (i.e., single-camera or multi-camera). Evaluation metrics for these can be formulated as follows:

$$FCR = T_S|T_1 \tag{1}$$

where T_S and T_1 are the length of the synopsis video and the input video, respectively. *Frame compact rate* (CR): the CR metric is used to determine whether the foreground is rearranged accurately in the synopsis, and is stated as follows:

$$CR = \frac{1}{w.h.T_S} \sum_{t=1}^{T_S} \sum_{x=1}^{w} \sum_{y=1}^{h} \{1|if\ p\ (x,\ y,\ t)\ \in\ foreground\ in\ V_s\} \tag{2}$$

where $p(x, y, t)$ indicates a pixel at the t-th frame such that w and h are the width and height of the synopsis frame. *Frame overlapping ratio* (FOR): the FOR defines the overlapping ratio between the collision degree of the foreground tubes:

$$FOR = \frac{1}{w.h.T_S} \sum_{t=1}^{T_S} \sum_{x=1}^{w} \sum_{y=1}^{h} \{1|if\ p\ (x,\ y,\ t)\ \in\ collision foreground\ in\ V_s\} \tag{3}$$

The CDR is defined as follows:

$$CDR = \frac{the\ number\ of\ chronological\ diordered\ of\ key\ time\ stamp\ pairs}{the\ total\ number\ of\ key\ time\ stamp\ pairs} \tag{4}$$

The precision, recall, and F1 measures determined as follows:

$$precision = \frac{True\ positives}{True\ positives + False\ positives} \tag{5}$$

$$recall = \frac{True\ positives}{True\ positives + False\ Negative} \tag{6}$$

$$F1\ measure = 2 \times \frac{precision \times\ recall}{precision\ +\ recall} \tag{7}$$

4.2. Analysis 1: Evaluation of Object Detection and Tracking

The object detection method solicited prior to the optimization phase directly impacts the features of the constructed synopsis video. Many object detection methods have been recently proposed and promised precise detection; thus, transforming their use in the synopsis framework can drastically change the quality of the synopsis video. A false negative prediction of an object can increase the computational cost; thus, adapting an appropriate detection method is crucial for the overall performance of the synopsis framework. An efficient object tracking method can significantly increase precision. We analyzed the effectiveness and efficiency of several algorithms: GMM [124], MAP-based algorithms [44], LCRT [51], LOBSTERBGS [91], Object Flag [20], SuBSENSE [125], 3D Graph-Cut and Pixel Domain [27], Graph-Cut with GMC and VQ [71], and MLBSA [126].

The analysis clearly shows that few algorithms suffered from noise. However, the observation signifies that in video1, video3, video4, and video5, the detected foreground mask achieved by MLBSA is slightly better than the others. In Video2, the 3D Graph-Cut and Pixel Domain approach achieves a better result. A visual assessment is provided for the GMM (T1), LOBSTERBGS (T2), MLBSA (T3), and 3D Graph-Cut Pixel Domain (T4) approaches, and is shown in Figure 4.

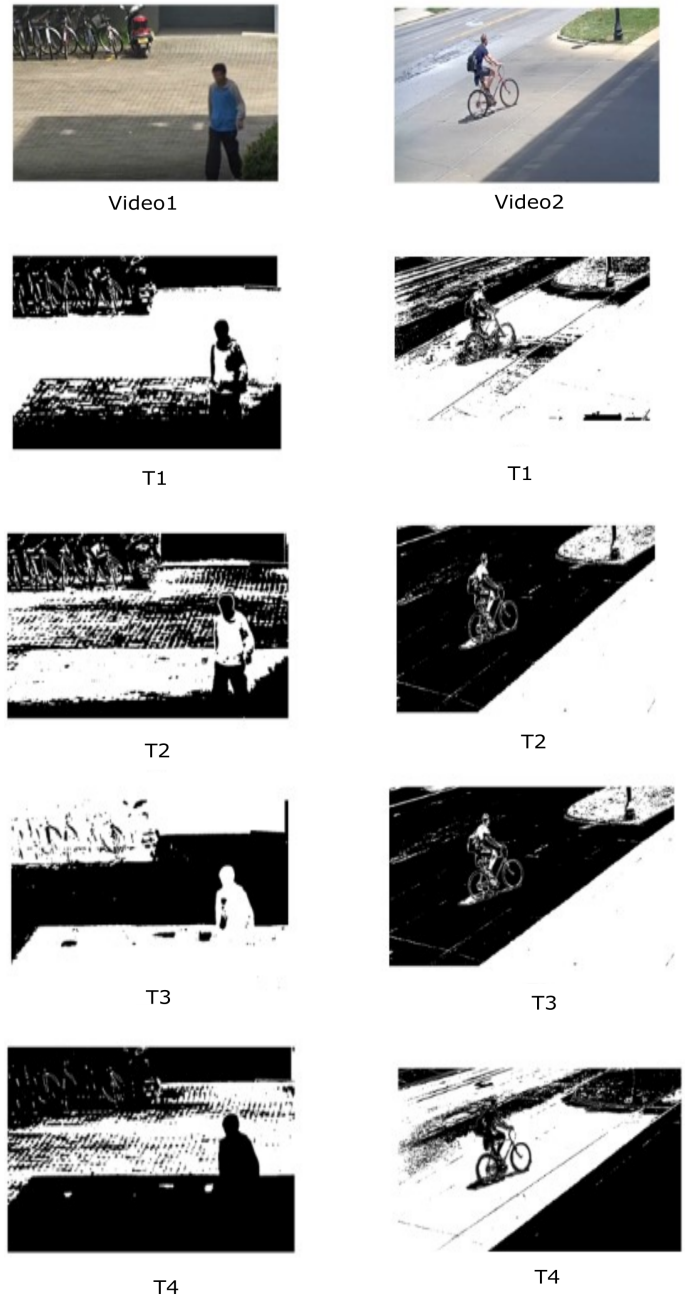

Figure 4. Illustration of different synopsis methodologies for generating foreground segmentation.

MLBSA performed better compared to others, as it leverages extraction of the binary pattern from the features, and as such is able to smoothly deal with illumination from moving objects. However, the synopsis generation time was seen lower when using 3D Graph-Cut and Pixel Domain, as this method converts the 3D Graph to a 2D Graph in order to determine the spatial location of the nodes. In the figure, Video1 and Video2 show the original input video frames, while T1, T2, T3, and T4 in each column depict the results for the respective input frame. Additionally, a quantitative analysis using the standard metrics of precision, recall, and F1-score is shown in Table 8.

Table 8. Quantitative analysis of different detection and tracking methods.

Dataset	Methods	Precision	Recall	F1	Time of Execution (s)
Hall Monitor Video-1	GMM	0.55	0.60	0.59	356.12
	MAP Based	0.66	0.69	0.71	214.89
	LCRT Algorithm	0.46	0.52	0.61	180.79
	LOBSTERBGS	0.54	0.66	0.60	269.01
	Object Flag	0.59	0.62	0.65	174.41
	SuBSENSE	0.69	0.75	0.61	266.32
	3D Graph Cut and Pixel Domain	0.67	0.72	0.69	**140.25**
	Graph Cut Algorithm	0.57	0.63	0.64	251.45
	GMC and VQ	0.60	0.64	0.61	154.38
	MLBSA	**0.75**	**0.76**	**0.72**	284.03
Hall Monitor Video-2	GMM	0.48	0.55	0.47	557.12
	MAP Based	0.61	0.66	0.52	348.25
	LCRT Algorithm	0.54	0.64	0.59	373.89
	LOBSTERBGS	0.64	0.66	0.60	545.43
	Object Flag	0.57	0.58	0.51	398.93
	SuBSENSE	0.59	0.60	0.55	436.25
	3D Graph Cut and Pixel Domain	**0.69**	**0.75**	**0.72**	311.71
	Graph Cut Algorithm	0.61	0.67	0.52	342.55
	GMC and VQ	0.59	0.60	0.54	**243.32**
	MLBSA	0.62	0.63	0.56	634.01
Hall Monitor Video-3	GMM	0.67	0.74	0.65	388.26
	MAP Based	0.78	0.83	0.77	247.03
	LCRT Algorithm	0.58	0.66	0.67	212.93
	LOBSTERBGS	0.66	0.80	0.66	301.15
	Object Flag	0.71	0.76	0.71	206.55
	SuBSENSE	0.81	0.89	0.67	298.46
	3D Graph Cut and Pixel Domain	0.79	0.86	0.75	**172.39**
	Graph Cut Algorithm	0.69	0.77	0.7	283.59
	GMC and VQ	0.72	0.78	0.67	186.52
	MLBSA	**0.87**	**0.90**	**0.78**	316.17
Hall Monitor Video-4	GMM	0.64	0.70	0.67	398.59
	MAP Based	0.75	0.79	0.79	257.36
	LCRT Algorithm	0.55	0.62	0.69	223.26
	LOBSTERBGS	0.63	0.76	0.68	311.48
	Object Flag	0.68	0.72	0.73	216.88
	SuBSENSE	0.78	0.85	0.69	308.79
	3D Graph Cut and Pixel Domain	0.76	0.82	0.77	**182.72**
	Graph Cut Algorithm	0.66	0.73	0.72	293.92
	GMC and VQ	0.69	0.74	0.69	196.85
	MLBSA	**0.84**	**0.86**	**0.80**	326.50
Hall Monitor Video-5	GMM	0.58	0.66	0.60	369.98
	MAP Based	0.69	0.75	0.72	228.75
	LCRT Algorithm	0.49	0.58	0.62	194.65
	LOBSTERBGS	0.57	0.72	0.61	282.87
	Object Flag	0.62	0.68	0.66	188.27
	SuBSENSE	0.72	0.81	0.62	280.18
	3D Graph Cut and Pixel Domain	0.70	0.78	0.70	**154.11**
	Graph Cut Algorithm	0.60	0.69	0.65	265.31
	GMC and VQ	0.63	0.70	0.62	168.24
	MLBSA	**0.78**	**0.82**	**0.73**	297.89

4.3. Analysis 2: Evaluation of Various Optimization Techniques

The runtime performance of a synopsis framework is inversely dependent on the condensation ratio. Offline optimization methods are more applicable to real-world problems, showing better performance. Efficient offline optimization methods can perform object rearrangement in the time and space domains on crowded videos. In contrast, online techniques can be more appropriate for less crowded videos. To demonstrate this difference, we compared the results of various optimization techniques: SA [105], TLBO [104], Graph Coloring [75], a greedy approach [59], Elitist-Jaya [127], ABSGCut [106], NSGA-II [128], Table-driven [60], GWO [129], and HSTLBO [42]. We considered the length of the generated video synopsis as equal to the generated tube length in order to condense the activities; thus, the activity cost is zero [83,130]. We first carried out a statistical analysis of the performance

in order to determine the superior algorithm, considering three parameters: collision cost, temporal cost, and time of execution. Table 9 depicts the performance comparison between the optimization methods. After this evaluation, it can be observed that the HSTLBO and TLBO algorithms perform better when considering the convergence parameters. NSGA-II mitigates non-elitism, computational complexity, and parameter sharing; thus, the optimization achieved was comparatively less. HSATLBO is a hybrid approach that rigorously searches for optimum solutions by minimizing the collision and activity costs.

Table 9. Performance comparison of various optimization techniques.

	Optimization Technique	Activity Cost	Collision Cost ($\times 10^3$)	Temporal Consistency Cost	Time of Execution (s)
Video-1	SA	0	16.21	11.2	356.12
	TLBO	0	16.01	11.5	214.89
	Graph Coloring	0	20.28	15.7	180.79
	Greedy Approach	0	18.05	13.1	269.01
	Elitist-JAYA	0	15.78	11.6	174.41
	ABSGCut	0	18.01	14.7	266.32
	NSGA-II	0	15.65	11.4	**140.25**
	Table-driven	0	17.47	12.3	251.45
	GWO	0	16.23	12.4	154.38
	HSTLBO	0	**14.03**	**10.8**	284.03
Video-2	SA	0	145.36	55.8	557.12
	TLBO	0	**137.32**	**49.4**	348.25
	Graph Coloring	0	190.01	70.3	373.89
	Greedy Approach	0	158.74	65.5	545.43
	Elitist-JAYA	0	150.21	61.7	398.93
	ABSGCut	0	159.65	72.3	436.25
	NSGA-II	0	148.47	54.4	311.71
	Table-driven	0	162.55	70.6	342.55
	GWO	0	151.17	67.8	**243.32**
	HSTLBO	0	146.87	58.7	634.01
Video-3	SA	0	18.34	12.6	388.26
	TLBO	0	18.14	12.9	247.03
	Graph Coloring	0	22.41	17.1	212.93
	Greedy Approach	0	20.18	14.5	301.15
	Elitist-JAYA	0	17.91	13.1	206.55
	ABSGCut	0	20.14	16.1	298.46
	NSGA-II	0	17.78	12.8	**172.39**
	Table-driven	0	19.60	13.7	283.59
	GWO	0	18.36	13.8	186.52
	HSTLBO	0	**16.16**	**12.2**	316.17
Video-4	SA	0	20.47	13.9	398.59
	TLBO	0	20.27	14.2	257.36
	Graph Coloring	0	24.54	18.4	223.26
	Greedy Approach	0	22.31	15.8	311.48
	Elitist-JAYA	0	20.04	14.3	216.88
	ABSGCut	0	22.27	17.4	308.79
	NSGA-II	0	19.91	14.1	**182.72**
	Table-driven	0	21.73	15.1	293.92
	GWO	0	20.49	15.1	196.85
	HSTLBO	0	**18.29**	**13.5**	326.50

Table 9. *Cont.*

	Optimization Technique	Activity Cost	Collision Cost ($\times 10^3$)	Temporal Consistency Cost	Time of Execution (s)
Video-5	SA	0	17.35	20.3	369.98
	TLBO	0	17.15	20.6	228.75
	Graph Coloring	0	21.42	24.8	194.65
	Greedy Approach	0	19.19	22.2	282.87
	Elitist-JAYA	0	16.92	20.7	188.27
	ABSGCut	0	19.15	23.8	280.18
	NSGA-II	0	16.79	20.5	**154.11**
	Table-driven	0	18.61	21.4	265.31
	GWO	0	17.37	21.5	168.24
	HSTLBO	0	**15.17**	**19.9**	297.89

Video synopsis is a complex problem consisting of several components working together to accomplish a single task. In this article, we have primarily focused on experimentally evaluating different detection, tracking, and optimization methods. However, several other parameters, such as the segmentation mask and the blending process, can be further assessed to determine a broader insight view. Most of the existing synopsis studies are application-oriented, and were designed to deal with a specific scenario; thus evaluating each study proved to be complicated and time-consuming, as each required particular types of video inputs and experimental setup. Certain studies required high-definition (HD) videos to minimize a significant drop in the detected object. In real-time synopsis, we used steady HD camera footage, which was computationally expensive when generating a tube. Our experiments were conducted in a controlled environment, and used a publicly available dataset to clearly define the cavity and component integration.

5. Challenges in Video Synopsis

Today, surveillance systems typically encompass multiple cameras aligned together using networking devices for surveillance. Therefore, intelligent surveillance systems are highly complex systems. These systems are responsible for monitoring daily activities 24/7 by using multiple cameras to extract a considerable amount of high-definition real-time video data. Mainly, these surveillance cameras have low computational capacity, and a set of cameras is connected to a common server for video data storage. Thus, extracting meaningful video data from different viewpoints to construct a video synopsis is tedious, contributing to many challenges. A number of challenges faced by researchers are listed below.

1. Edge-based synopsis: as next-generation surveillance cameras have slightly better computing, the summary can be accomplished on the edge device itself using technologies such as fog/cloud computing. However, state-of-the-art synopsis frameworks lack the required capabilities to create edge-based solutions.
2. Multi-view video synopsis: creating a synopsis for every single camera occupies a great deal of space and time; a better real-world solution is multi-view video synopsis, as it can create a single synopsis for multiple videos. However, a major problem that occurs is selecting a common background, as the acquired videos have different view angles and locations. Thus, the resulting synopsis view is complex and challenging to understand, as the tubes are shifted against a very different background.
3. Visual constituent redundancy: there have been many methods proposed for creating single-view camera summaries in past years. When a similar strategy is applied in the case of multi-view camera systems, the inter-video relations between visual content are ignored, leading to redundant content. Therefore, it is better to use a synopsis of each video and then stitch the frames to create a single summary for multi-view cameras.

4. Relationship association: as there are numerous objects present the constructed video synopses, it is difficult for a video analyzer to associate summary objects with the original video objects. A better option is to create a single-camera synopsis, which is not feasible in real-world surveillance system with multiple cameras. Thus, there is a need to find a mechanism that can link the desired synopsis object with the original video cameras.
5. Multi-model: as there are several components in the video synopsis framework, a multi-model learning approach can be used for better inclusion of these components. A single multitask learning model can perform segmentation, depth analysis, and background generation.
6. Interactive: as synopsis generation is predefined or application based, incorporation of an interactive user mode can help to generate user-defined parameters such as type of object, duration and speed of synopsis, etc.

6. Conclusions

In this article, we have provided a comprehensive survey and experimental analysis of different video synopsis methods. We cover all of the state-of-the-art synopsis methodologies, from the initial studies in the field until 2022. Based on their characteristics, we have classified the procedure into multiple techniques, namely, frame-based, object-based, action-based, collision graph-based, and abnormal content-based. Additionally, we have used various scenarios to discuss different synopsis frameworks while providing a taxonomy, and classified the methods applied in various video synopsis components. Focusing on each stage of the video synopsis process, we have provided a systematic comparison among the methods used in the detection, tracking, optimization, and stitching stages. Our analysis indicates that the MLSBA and 3D Graph-cut Pixel Domain procedures perform significantly better on object detection and tracking. At the same time, NSGA-II and GWO represent better optimization techniques for avoiding collisions, whereas the method proposed by Nie et al. is well-situated for multi-camera view synopsis stitching with minimum computational capacity. The benefits and drawbacks of each technique are associated with several other insights to provide a detailed understanding of synopsis methods for real-world application. Prominently, the many open challenges currently faced by researchers when dealing with synopsis have been brought to the forefront.

Author Contributions: The authors contributed to this paper as follows: P.Y.I. wrote this article, reviewed and designed the system framework, and conducted experimental evaluation; Y.-G.K. supervised and coordinated the investigation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (No.2019-0-00231, Development of artificial intelligence-based video security technology and systems for public infrastructure safety).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

ABGC	Alpha–Beta Graph Cut
CNN	Convolutional Neural Networks
DP	Dynamic Programming
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GSVSF	Geospatial video synopsis framework
SVSF	Spherical video synopsis framework

IVSF	Infrared video synopsis framework
KNN	k-nearest neighbor
LCRT	Low Complexity Range Tree
MRF	Markov Random Field
OD	Object Detection
RTVS	Real-time video synopsis framework
SA	Simulated Annealing
TA	Tube Generation

References

1. Reinsel, D.; Gantz, J.; Rydning, J. *Data Age 2025: The Evolution of Data to Life-Critical. Don't Focus on Big Data; Focus on the Data That's Big*; International Data Corporation (IDC) White Paper; 2017. Available online: https://www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf (accessed on 10 October 2022).
2. Sarhan, N. Automated Video Surveillance Systems. U.S. Patent 9,313,463, 12 April 2016.
3. Tsakanikas, V.; Dagiuklas, T. Video surveillance systems-current status and future trends. *Comput. Electr. Eng.* **2018**, *70*, 736–753. [CrossRef]
4. Truong, B.; Venkatesh, S. Video abstraction: A systematic review and classification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2007**, *3*, 3–es. [CrossRef]
5. Nam, J.; Tewfik, A. Video abstract of video. In Proceedings of the 1999 IEEE Third Workshop on Multimedia Signal Processing (Cat. No. 99TH8451), Copenhagen, Denmark, 13–15 September 1999; pp. 117–122. [CrossRef]
6. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, 23–26 August 2004; Volume 3, pp. 32–36. [CrossRef]
7. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef] [PubMed]
8. Sillito, R.; Fisher, R. Semi-supervised Learning for Anomalous Trajectory Detection. In Proceedings of the BMVC, Leeds, UK, 1–4 September 2008; Volume 1, p. 1–10.
9. Yang, B.; Nevatia, R. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1918–1925. Available online: https://www.computer.org/csdl/proceedings/cvprw/2012/12OmNANkoa6 (accessed on 18 October 2022).
10. Mahalingam, T.; Subramoniam, M. ACO–MKFCM: An Optimized Object Detection and Tracking Using DNN and Gravitational Search Algorithm. *Wirel. Pers. Commun.* **2020**, *110*, 1567–1604. [CrossRef]
11. Kille, B.; Hopfgartner, F.; Brodt, T.; Heintz, T. The plista dataset. In Proceedings of the 2013 International News Recommender Systems Workshop and Challenge, Hong Kong, China, 13 October 2013; pp. 16–23.
12. Wang, T.; Liang, J.; Wang, X.; Wang, S. Background modeling using local binary patterns of motion vector. In Proceedings of the 2012 Visual Communications and Image Processing, San Diego, CA, USA, 27–30 November 2012; pp. 1–5.
13. Baskurt, K.; Samet, R. Video synopsis: A survey. *Comput. Vis. Image Underst.* **2019**, *181*, 26–38. [CrossRef]
14. Ghatak, S.; Rup, S. Single Camera Surveillance Video Synopsis: A Review and Taxonomy. In Proceedings of the 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 19–21 December 2019; pp. 483–488. Available online: https://ieeexplore.ieee.org/xpl/conhome/9022805/proceeding (accessed on 20 November 2022).
15. Mahapatra, A.; Sa, P. Video Synopsis: A Systematic Review. In *High Performance Vision Intelligence*; Springer: Singapore, 2020; pp. 101–115.
16. Liu, T.; Zhang, X.; Feng, J.; Lo K. Shot reconstruction degree: A novel criterion for key frame selection. *Pattern Recognit. Lett.* **2004**, *25*, 1451–1457. [CrossRef]
17. Choudhary, V.; Tiwari, A. Surveillance video synopsis. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 207–212.
18. Zass, R.; Shashua, A. A unifying approach to hard and probabilistic clustering. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), NW, Washington, DC, USA, 17–21 October 2005; Volume 1, pp. 294–301.
19. Pritch, Y.; Ratovitch, S.; Hendel, A.; Peleg, S. Clustered synopsis of surveillance video. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy, 2–4 September 2009; pp. 195–200.
20. Wang, S.; Yang, J.; Zhao, Y.; Cai, A.; Li, S. A surveillance video analysis and storage scheme for scalable synopsis browsing. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1947–1954.
21. Petrovic, N.; Jojic, N.; Huang, T. Adaptive video fast forward. *Multimed. Tools Appl.* **2005**, *26*, 327–344. [CrossRef]
22. Smith, M.; Kanade, T. Video skimming and characterization through the combination of image and language understanding techniques. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; pp. 775–781. Available online: https://ieeexplore.ieee.org/xpl/conhome/4821/proceeding (accessed on 13 November 2022).

23. Pal, C.; Jojic, N. Interactive montages of sprites for indexing and summarizing security video. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, p. 1192
24. Wexler, Y.; Shechtman, E.; Irani, M. Space-time video completion. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 1.
25. Rother, C.; Bordeaux, L.; Hamadi, Y.; Blake, A. Autocollage. *ACM Trans. Graph. (TOG)* **2006**, *25*, 847–852. [CrossRef]
26. Kang, H.; Matsushita, Y.; Tang, X.; Chen, X. Space-time video montage. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1331–1338.
27. Rav-Acha, A.; Pritch, Y.; Peleg, S. Making a long video short: Dynamic video synopsis. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 1, pp. 435–441.
28. Pritch, Y.; Rav-Acha, A.; Gutman, A.; Peleg, S. Webcam synopsis: Peeking around the world. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio De Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
29. Xu, M.; Li, S.; Li, B.; Yuan, X.; Xiang, S. A set theoretical method for video synopsis. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, BC, Canada, 30–31 October 2008; pp. 366–370.
30. Pritch, Y.; Rav-Acha, A.; Peleg, S. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1971–1984. [CrossRef]
31. Wang, S.; Liu, H.; Xie, D.; Zeng, B. A novel scheme to code object flags for video synopsis. In Proceedings of the 2012 Visual Communications and Image Processing, San Diego, CA, USA, 27–30 November 2012; pp. 1–5.
32. Sun, H.; Cao, L.; Xie, Y.; Zhao, M. The method of video synopsis based on maximum motion power. In Proceedings of the 2011 Third Chinese Conference on Intelligent Visual Surveillance, Beijing, China, 1–2 December 2011; pp. 37–40.
33. Zhu, X.; Liu, J.; Wang, J.; Lu, H. Key observation selection for effective video synopsis. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 2528–2531.
34. Nie, Y.; Xiao, C.; Sun, H.; Li, P. Compact video synopsis via global spatiotemporal optimization. *IEEE Trans. Vis. Comput. Graph.* **2012**, *19*, 1664–1676. [CrossRef]
35. Yao, T.; Xiao, M.; Ma, C.; Shen, C.; Li, P. Object based video synopsis. In Proceedings of the 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), Ottawa, ON, Canada, 29–30 September 2014; pp. 1138–1141.
36. Olivera, J.; Cuadra, N.; Oliva, E.; Albornoz, E.; Martinez, C. Development of an open source library for the generation of video synopsis. In Proceedings of the 2015 XVI Workshop on Information Processing and Control (RPIC), Cordoba, Argentina, 6–9 October 2015; pp. 1–4.
37. Tian, Y.; Zheng, H.; Chen, Q.; Wang, D.; Lin, R. Surveillance video synopsis generation method via keeping important relationship among objects. *IET Comput. Vis.* **2016**, *10*, 868–872. [CrossRef]
38. Ahmed, A.; Kar, S.; Dogra, D.; Patnaik, R.; Lee, S.; Choi, H.; Kim, I. Video synopsis generation using spatio-temporal groups. In Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Sarawak, Malaysia, 12–14 September 2017; pp. 512–517.
39. Collins, R.; Osterdahl, K.; Shringi, A.; Corona, K.; Swears, E.; Meth, R.; Hoogs, A. *Data, Algorithms, and Framework for Automated Analytics of Surveillance Camera Networks*; Kitware Inc.: Clifton Park, NY, USA, 2018.
40. He, Y.; Han, J.; Sang, N.; Qu, Z.; Gao, C. Chronological video synopsis via events rearrangement optimization. *Chin. J. Electron.* **2018**, *27*, 399–404. [CrossRef]
41. Li, X.; Wang, Z.; Lu, X. Video synopsis in complex situations. *IEEE Trans. Image Process.* **2018**, *27*, 3798–3812. [CrossRef]
42. Ghatak, S.; Rup, S.; Majhi, B.; Swamy, M. An improved surveillance video synopsis framework: A HSATLBO optimization approach. *Multimed. Tools Appl.* **2020**, *79*, 4429–4461. [CrossRef]
43. Moussa, M.; Shoitan, R. Object-based video synopsis approach using particle swarm optimization. *Signal Image Video Process.* **2021**, *15*, 761–768. [CrossRef]
44. Li, T.; Ma, Y.; Endoh, T. A systematic study of tiny YOLO3 inference: Toward compact brainware processor with less memory and logic gate. *IEEE Access* **2020**, *8*, 142931–142955. [CrossRef]
45. Yildiz, A.; Ozgur, A.; Akgul, Y. Fast non-linear video synopsis. In Proceedings of the 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 27–29 October 2008; pp. 1–6.
46. Vural, U.; Akgul, Y. Eye-gaze based real-time surveillance video synopsis. *Pattern Recognit. Lett.* **2009**, *30*, 1151–1159. [CrossRef]
47. Feng, S.; Liao, S.; Yuan, Z.; Li, S. Online principal background selection for video synopsis. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 17–20.
48. Huang, C.; Chen, H.; Chung, P. Online surveillance video synopsis. In Proceedings of the 2012 IEEE International Symposium on Circuits and Systems (ISCAS), Seoul, Republic of Korea, 20–23 May 2012; pp. 1843–1846.
49. Huang, C.; Chung, P.; Yang, D.; Chen, H.; Huang, G. Maximum a posteriori probability estimation for online surveillance video synopsis. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 1417–1429. [CrossRef]
50. Sun, L.; Xing, J.; Ai, H.; Lao, S. A tracking based fast online complete video synopsis approach. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 1956–1959.

51. Hsia, C.; Chiang, J.; Hsieh, C.; Hu, L. A complexity reduction method for video synopsis system. In Proceedings of the 2013 International Symposium on Intelligent Signal Processing and Communication Systems, Okinawa, Japan, 12–15 November 2013; pp. 163–168.
52. Fu, W.; Wang, J.; Gui, L.; Lu, H.; Ma, S. Online video synopsis of structured motion. *Neurocomputing* **2014**, *135*, 155–162. [CrossRef]
53. Ghatak, S.; Rup, S.; Majhi, B.; Swamy, M. HSAJAYA: An improved optimization scheme for consumer surveillance video synopsis generation. *IEEE Trans. Consum. Electron.* **2020**, *66*, 144–152. [CrossRef]
54. Chen, S.; Liu, X.; Huang, Y.; Zhou, C.; Miao, H. Video synopsis based on attention mechanism and local transparent processing. *IEEE Access* **2020**, *8*, 92603–92614. [CrossRef]
55. Namitha, K.; Narayanan, A.; Geetha, M. Interactive visualization-based surveillance video synopsis. *Appl. Intell.* **2022**, *52*, 3954–3975. [CrossRef]
56. Kostadinov, G. Synopsis of video files using neural networks. In Proceedings of the International Conference on Engineering Applications of Neural Networks, Chersonisos, Crete, Greece, 17–20 June 2022; pp. 190–202.
57. Li, X.; Qiu, S.; Song, Y. Dynamic Synopsis and storage algorithm based on infrared surveillance video. *Infrared Phys. Technol.* **2022**, *124*, 104213. [CrossRef]
58. Zhu, J.; Liao, S.; Li, S. Multicamera joint video synopsis. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 1058–1069. [CrossRef]
59. Hoshen, Y.; Peleg, S. Live video synopsis for multiple cameras. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 212–216.
60. Mahapatra, A.; Sa, P.; Majhi, B.; Padhy, S. MVS: A multi-view video synopsis framework. *Signal Process. Image Commun.* **2016**, *42*, 31–44. [CrossRef]
61. Zhang, Z.; Nie, Y.; Sun, H.; Zhang, Q.; Lai, Q.; Li, G.; Xiao, M. Multi-view video synopsis via simultaneous object-shifting and view-switching optimization. *IEEE Trans. Image Process.* **2019**, *29*, 971–985. [CrossRef]
62. Xie, Y.; Wang, M.; Liu, X.; Wang, X.; Wu, Y.; Wang, F.; Wang, X. Multi-camera video synopsis of a geographic scene based on optimal virtual viewpoint. *Trans. GIS* **2022**, *26*, 1221–1239. [CrossRef]
63. Priyadharshini, S.; Mahapatra, A. PanoSyn: Immersive video synopsis for spherical surveillance video. *Sādhanā* **2022**, *47*, 167. [CrossRef]
64. Hao, L.; Cao, J.; Li, C. Research of grabcut algorithm for single camera video synopsis. In Proceedings of the 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), Beijing, China, 9–11 June 2013; pp. 632–637.
65. Agarwala, A.; Hertzmann, A.; Salesin, D.; Seitz, S. Keyframe-based tracking for rotoscoping and animation. *ACM Trans. Graph. (ToG)* **2004**, *23*, 584–591. [CrossRef]
66. Nie, Y.; Sun, H.; Li, P.; Xiao, C.; Ma, K. Object movements synopsis viaPart assembling and stitching. *IEEE Trans. Vis. Comput. Graph.* **2014**, *20*, 1303–1315. [CrossRef]
67. Wang, J.; Bhat, P.; Colburn, R.; Agrawala, M.; Cohen, M. Interactive video cutout. *ACM Trans. Graph. (ToG)* **2005**, *24*, 585–594. [CrossRef]
68. Bai, X.; Wang, J.; Simons, D.; Sapiro, G. Video snapcut: Robust video object cutout using localized classifiers. *ACM Trans. Graph. (ToG)* **2009**, *28*, 1–11. [CrossRef]
69. Lu, M.; Wang, Y.; Pan, G. Generating fluent tubes in video synopsis. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 2292–2296.
70. Wang, S.; Wang, Z.; Hu, R. Surveillance video synopsis in the compressed domain for fast video browsing. *J. Vis. Commun. Image Represent.* **2013**, *24*, 1431–1442. [CrossRef]
71. Zhong, R.; Hu, R.; Wang, Z.; Wang, S. Fast synopsis for moving objects using compressed video. *IEEE Signal Process. Lett.* **2014**, *21*, 834–838. [CrossRef]
72. Zhu, J.; Feng, S.; Yi, D.; Liao, S.; Lei, Z.; Li, S. High-performance video condensation system. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 1113–1124.
73. Chakraborty, S.; Tickoo, O.; Iyer, R. Adaptive keyframe selection for video summarization. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; pp. 702–709.
74. He, Y.; Qu, Z.; Gao, C.; Sang, N. Fast online video synopsis based on potential collision graph. *IEEE Signal Process. Lett.* **2016**, *24*, 22–26. [CrossRef]
75. He, Y.; Gao, C.; Sang, N.; Qu, Z.; Han, J. Graph coloring based surveillance video synopsis. *Neurocomputing* **2017**, *225*, 64–79. [CrossRef]
76. Liao, W.; Tu, Z.; Wang, S.; Li, Y.; Zhong, R.; Zhong, H. Compressed-domain video synopsis via 3d graph cut and blank frame deletion. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 253–261.
77. Ra, M.; Kim, W. Parallelized tube rearrangement algorithm for online video synopsis. *IEEE Signal Process. Lett.* **2018**, *25*, 1186–1190. [CrossRef]
78. Pappalardo, G.; Allegra, D.; Stanco, F.; Battiato, S. A new framework for studying tubes rearrangement strategies in surveillance video synopsis. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 664–668.
79. Ruan, T.; Wei, S.; Li, J.; Zhao, Y. Rearranging online tubes for streaming video synopsis: A dynamic graph coloring approach. *IEEE Trans. Image Process.* **2019**, *28*, 3873–3884. [CrossRef]

80. Li, K.; Yan, B.; Wang, W.; Gharavi, H. An effective video synopsis approach with seam carving. *IEEE Signal Process. Lett.* **2015**, *23*, 11–14. [CrossRef]
81. Jin, J.; Liu, F.; Gan, Z.; Cui, Z. Online video synopsis method through simple tube projection strategy. In Proceedings of the 2016 8th International Conference on Wireless Communications & Signal Processing (WCSP), Yangzhou, China, 13–15 October 2016; pp. 1–5.
82. Nie, Y.; Li, Z.; Zhang, Z.; Zhang, Q.; Ma, T.; Sun, H. Collision-free video synopsis incorporating object speed and size changes. *IEEE Trans. Image Process.* **2019**, *29*, 1465–1478. [CrossRef] [PubMed]
83. Li, X.; Wang, Z.; Lu, X. Surveillance video synopsis via scaling down objects. *IEEE Trans. Image Process.* **2015**, *25*, 740–755. [CrossRef] [PubMed]
84. Chou, C.; Lin, C.; Chiang, T.; Chen, H.; Lee, S. Coherent event-based surveillance video synopsis using trajectory clustering. In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Torino, Italy, 29 June–3 July 2015; pp. 1–6.
85. Lin, W.; Zhang, Y.; Lu, J.; Zhou, B.; Wang, J.; Zhou, Y. Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. *Neurocomputing* **2015**, *155*, 84–98. [CrossRef]
86. Ahmed, S.; Dogra, D.; Kar, S.; Patnaik, R.; Lee, S.; Choi, H.; Nam, G.; Kim, I. Query-based video synopsis for intelligent traffic monitoring applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3457–3468. [CrossRef]
87. Ingle, P.; Kim, Y. Real-Time Abnormal Object Detection for Video Surveillance in Smart Cities. *Sensors* **2022**, *22*, 3862. [CrossRef]
88. Ingle, P.; Kim, Y.; Kim, Y. Dvs: A drone video synopsis towards storing and analyzing drone surveillance data in smart cities. *Systems* **2022**, *10*, 170. [CrossRef]
89. Stauffer, C.; Grimson, W. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 747–757. [CrossRef]
90. Hofmann, M.; Tiefenbacher, P.; Rigoll, G. Background segmentation with feedback: The pixel-based adaptive segmenter. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 38–43.
91. St-Charles, P.; Bilodeau, G. Improving background subtraction using local binary similarity patterns. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Springs, CO, USA, 24–26 March 2014; pp. 509–515.
92. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
93. Finkel, R.; Bentley, J. Quad trees a data structure for retrieval on composite keys. *Acta Inform.* **1974**, *4*, 1–9. [CrossRef]
94. Kolmogorov, V.; Zabin, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 147–159. [CrossRef]
95. Liao, S.; Zhao, G.; Kellokumpu, V.; Pietikäinen, M.; Li, S. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1301–1306.
96. Kasamwattanarote, S.; Cooharojananone, N.; Satoh, S.; Lipikorn, R. Real time tunnel based video summarization using direct shift collision detection. In Proceedings of the Pacific-Rim Conference on Multimedia, Shanghai, China, 21–24 September 2010; pp. 136–147.
97. Kalman, R. A new approach to linear filtering and prediction problems. *J. Basic Eng. Mar.* **1960**, *82*, 35–45. [CrossRef]
98. Emran, S.; Ye, N. Robustness of Chi-square and Canberra distance metrics for computer intrusion detection. *Qual. Reliab. Eng. Int.* **2002**, *18*, 19–28. [CrossRef]
99. Bolme, D.; Beveridge, J.; Draper, B.; Lui, Y. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
100. Gorin, B.; Waxman, A. Flight test capabilities for real-time multiple target detection and tracking for airborne surveillance and maritime domain awareness. *Opt. Photonics Glob. Homel. Secur. IV* **2008**, *6945*, 205–217.
101. Rajab, M.; Woolfson, M.; Morgan, S. Application of region-based segmentation and neural network edge detection to skin lesions. *Comput. Med. Imaging Graph.* **2004**, *28*, 61–68. [CrossRef]
102. Salvador, S.; Chan, P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, 15–17 November 2004; pp. 576–584.
103. Hafiane, A.; Chabrier, S.; Rosenberger, C.; Laurent, H. A new supervised evaluation criterion for region based segmentation methods. In Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, Delft, The Netherlands, 28–31 August 2007; pp. 439–448.
104. Rao, R.; Savsani, V.; Vakharia, D. Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems. *Comput.-Aided Des.* **2011**, *43*, 303–315. [CrossRef]
105. Kirkpatrick, S.; Gelatt, C., Jr.; Vecchi, M. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
106. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1222–1239. [CrossRef]
107. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **1994**, *4*, 65–85. [CrossRef]

108. Peleg, S.; Rousso, B.; Rav-Acha, A.; Zomet, A. Mosaicing on adaptive manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1144–1154. [CrossRef]
109. Zhi, Q.; Cooperstock, J. Toward dynamic image mosaic generation with robustness to parallax. *IEEE Trans. Image Process.* **2011**, *21*, 366–378. [CrossRef]
110. Uyttendaele, M.; Eden, A.; Skeliski, R. Eliminating ghosting and exposure artifacts in image mosaics. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 2, p. II.
111. Brown, M.; Lowe, D. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vis.* **2007**, *74*, 59–73. [CrossRef]
112. Lin, W.; Liu, S.; Matsushita, Y.; Ng, T.; Cheong, L. Smoothly varying affine stitching. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 345–352.
113. Liu, W.; Chin, T. Correspondence insertion for as-projective-as-possible image stitching. *arXiv* **2016**, arXiv:1608.07997.
114. Chang, C.; Sato, Y.; Chuang, Y. Shape-preserving half-projective warps for image stitching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3254–3261.
115. Li, N.; Xu, Y.; Wang, C. Quasi-homography warps in image stitching. *IEEE Trans. Multimed.* **2017**, *20*, 1365–1375. [CrossRef]
116. Chen, Y.; Chuang, Y. Natural image stitching with the global similarity prior. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 186–201.
117. Zhang, G.; He, Y.; Chen, W.; Jia, J.; Bao, H. Multi-viewpoint panorama construction with wide-baseline images. *IEEE Trans. Image Process.* **2016**, *25*, 3099–3111. [CrossRef] [PubMed]
118. Xiang, T.; Xia, G.; Bai, X.; Zhang, L. Image stitching by line-guided local warping with global similarity constraint. *Pattern Recognit.* **2018**, *83*, 481–497. [CrossRef]
119. Rav-Acha, A.; Pritch, Y.; Lischinski, D.; Peleg, S. Dynamosaics: Video mosaics with non-chronological time. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 58–65.
120. Su, T.; Nie, Y.; Zhang, Z.; Sun, H.; Li, G. Video stitching for handheld inputs via combined video stabilization. In Proceedings of the SIGGRAPH ASIA 2016 Technical Briefs, Macao, China, 5–8 December 2016; pp. 1–4.
121. Nie, Y.; Su, T.; Zhang, Z.; Sun, H.; Li, G. Dynamic video stitching via shakiness removing. *IEEE Trans. Image Process.* **2017**, *27*, 164–178. [CrossRef]
122. Lin, K.; Liu, S.; Cheong, L.; Zeng, B. Seamless video stitching from hand-held camera inputs. *Comput. Graph. Forum* **2016**, *35*, 479–487. [CrossRef]
123. Panda, D.; Meher, S. A new Wronskian change detection model based codebook background subtraction for visual surveillance applications. *J. Vis. Commun. Image Represent.* **2018**, *56*, 52–72. [CrossRef]
124. Lee, D. Effective Gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 827–832.
125. St-Charles, P.; Bilodeau, G.; Bergevin, R. Flexible background subtraction with self-balanced local sensitivity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 408–413.
126. Yao, J.; Odobez, J. Multi-layer background subtraction based on color and texture. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
127. Rao, R.; Saroj, A. Constrained economic optimization of shell-and-tube heat exchangers using elitist-Jaya algorithm. *Energy* **2017**, *128*, 785–800. [CrossRef]
128. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
129. Mirjalili, S.; Mirjalili, S.; Lewis, A. Grey wolf optimizer. *Adv. Eng. Softw.* **2014**, *69*, 46–61. [CrossRef]
130. Wang, W.; Chung, P.; Huang, C.; Huang, W. Event based surveillance video synopsis using trajectory kinematics descriptors. In Proceedings of the 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; pp. 250–253.