Computer Vision and Image Understanding 181 (2019) 26–38

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Video synopsis: A survey

Kemal Batuhan Baskurt *, Refik Samet

Bahcelievler Mahallesi, I Blok, Ankara Unv. 50. Yil Kampusu, 06830 Golbasi/Ankara, Turkey

ARTICLE INFO

Communicated by N. Paragios

Keywords:
Video surveillance
Video processing
Video synopsis
Motion detection
Object tracking
Optimization
Background generation
Stitching

ABSTRACT

Video synopsis is an activity-based video condensation approach to achieve efficient video browsing and retrieval for surveillance cameras. It is one of the most effective ways to reduce the inactive density of input video in order to provide fast and easy retrieval of the parts of interest. Unlike frame-based video summarization methods, the interested activities are shifted in the time domain to obtain video representation that is more compact. Although the number of studies on video synopsis has increased over the past years, there has still been no survey study on the subject. The aim in this article is to review state-of-the-art approaches in video synopsis studies and provide a comprehensive analysis. The methodology of video synopsis is described to provide an overview on the flow of the algorithm. Recent literature is investigated into different aspects such as optimization type, camera topology, input data domain, and activity clustering mechanisms. Commonly used performance evaluation techniques are also examined. Finally, the current situation of the literature and potential future research directions are discussed after an exhaustive analysis that covers most of the studies from early on to the present in this field. To the best of our knowledge, this study is the first review of published video synopsis approaches.

1. Introduction

Control and management of huge amounts of recorded video is becoming more difficult to deal with each passing day when considering the rapid increment in security camera usage in daily life. Efficient video browsing and retrieval are critical issues when considering the amount of raw video data to be summarized. The manpower required to monitor visual data is a challenging problem. Therefore, video condensation techniques are being widely investigated via a large number of applications in diverse disciplines.

A popular approach to solve video condensation problem is video synopsis, which has been investigated in the literature over the last decade. Video synopsis provides activity-based video condensation instead of frame-based techniques such as video fast-forward (Smith and Kanade, 1998), video abstraction (Truong and Venkatesh, 2007), and video summarization (Chakraborty et al., 2015). Video synopsis operates on an activity as a processing unit while frame-based approaches use a frame. Video synopsis achieves higher efficiency than frame-based video condensation techniques as smaller processing units provide the opportunity of better condensation because of more detailed video analysis. Activities can be shifted in the time domain and more than one activity can be showed simultaneously in a frame even though they come from different time periods.

The aim of video synopsis approaches is to find the best rearrangement of the activities in order to display most of them in the shortest time period. The biggest problem is handling activity collisions as they can lead to the loss of important content, thereby reducing efficiency. Collisions also cause a chaotic viewing experience which decreases the visual quality for surveillance applications. Displaying the maximum number of objects with minimal collisions means more computational complexity comparing to frame-based methods, because of processing the activities separately instead of processing the whole frame at once. Thus, video synopsis has become the hot spot in video summarization, especially with the support of technological improvement on computational capacity of current computers over the past years.

Existing video synopsis studies can be categorized by different aspects such as optimization type, camera topology, input data domain, and activity clustering. The aim of optimization is to find the best temporal positions of selected activities in order to obtain a more compact representation, which is the most important part of algorithm flow in video synopsis. Therefore, the most dominant criteria for categorization is optimization type, which is divided in two categories, namely on-line and off-line. A large part of the approaches performs off-line optimization of all activities to find the global optimum. However, latest approaches increasingly use on-line optimization that applies rearrangement on each new activity to find the local optimum. Aspects of camera topology have divided studies into two groups: single and multi-camera solutions. Most of the approaches are oriented toward the single-camera view that makes the optimization problem easier. Multi-camera approaches need to build a global energy definition which covers all of the camera network with the intention of finding the optimal solution for all. On the other hand, they provide the opportunity

* Corresponding author.
E-mail address: batuhanbaskurt@gmail.com (K.B. Baskurt).

https://doi.org/10.1016/j.cviu.2019.02.004
Received 13 December 2017; Received in revised form 9 February 2019; Accepted 11 February 2019
Available online 22 February 2019

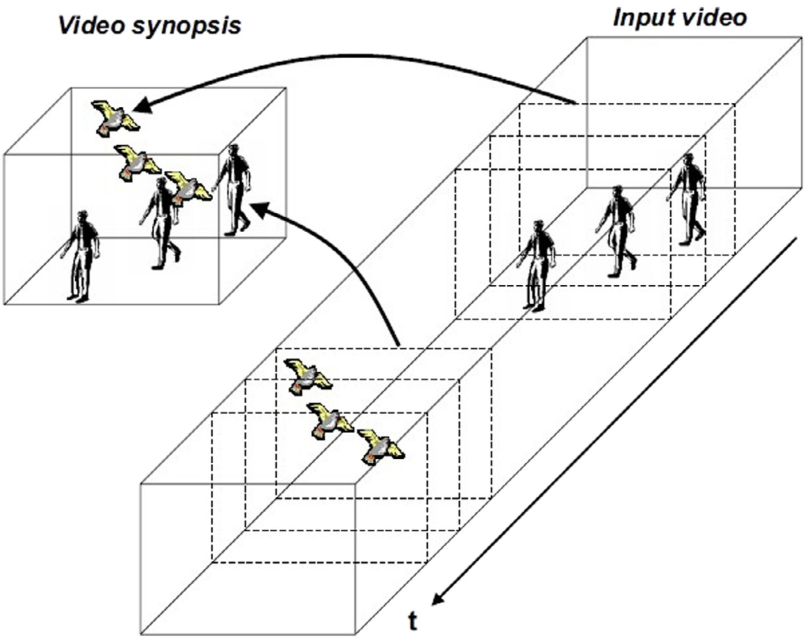


Fig. 1. Simultaneous display of activities (Rav-Acha et al., 2006).

to display and analyze activities in a wider perspective. Some studies focusing on run-time performance propose techniques applied directly to compressed data instead of losing time and computation power by transforming data to the pixel domain. Even though their run-time performance is significantly increased, condensation ratio cannot compete with pixel-domain methods. Besides, some studies apply activity clustering to group similar activities and display them together with the aim of providing better understanding of the scene as focusing on similar activities is easier for the user.

In this paper, we analyze 35 video synopsis approaches that cover all of the existing studies up to this point. Approaches are analyzed on the aforementioned aspects and the diversity of pre/post-processing methods used in existing video synopsis approaches are examined in detail.

The rest of the paper is organized as follows. Section 2 provides an overview of existing video synopsis approaches emphasizing on novelty and contribution to the field. Methods used in algorithm flow of video synopsis are described in Section 3. An analysis of the approaches according to optimization type, camera topology, input data domain, and activity clustering is described in Section 4. Evaluation criteria and commonly used datasets are presented in Section 5. Finally, Section 6 contains conclusions on the study.

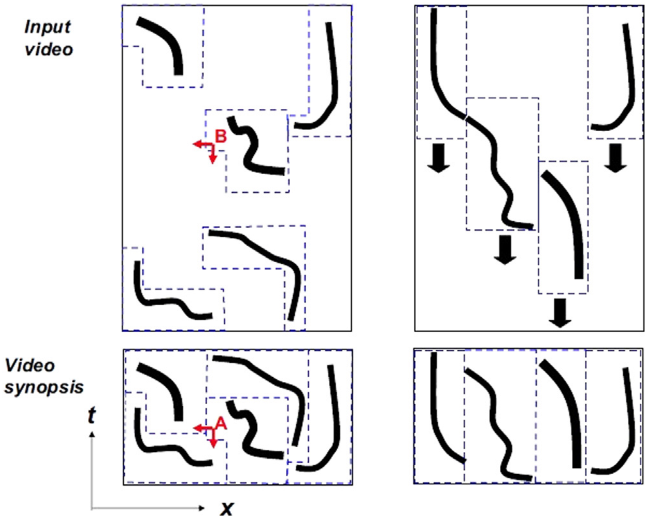


Fig. 2. Activity rearrangement (Rav-Acha et al., 2006).

2. Related works

Video synopsis is an activity-based video condensation technique and the main purpose is to display as many activities as possible simultaneously in the shortest time period. An activity represents a group of object instances belonging to a time period in which the object is visible. The activities extracted from the source are shifted in the time domain to calculate their optimal positions with the minimum number of collisions. Unlike frame-based video summarization techniques, activities from different time periods can be shifted into the same frame through pixel based analysis. Therefore, more efficient condensation performance is achieved compared to frame-based video summarization methods.

Activity-based video condensation was proposed by Rav-Acha et al. (2006) under the name of video synopsis, a novel approach that shifts detected activities in time domain to display them simultaneously over a shorter time period, as depicted in Fig. 1. Their approach contained two main phases: on-line and off-line. The on-line phase included activity generation and storing them into a queue. Subsequently, off-line phase started after selecting a time range of video synopsis with tube rearrangement, background generation, and object stitching. A global energy function containing activity, temporal consistency, and collision cost was defined, then simulated annealing method (Kirkpatrick et al., 1983) was applied for energy minimization, as illustrated in Fig. 2.

Their study is important as the video synopsis approach was proposed for the first time. Even though the study led to follow up ones, it is still a primitive version of video synopsis. In this manner, researchers continue to improve the approach by applying video synopsis to endless video streams, as reported by Pritch et al. (2007). The term ‘tube’ for representing activity consisting of object trajectories in video frames was first used in this study and has been widely used in the literature ever since.

They applied a better object detection method to improve the precision of video synopsis and proposed a more detailed energy function definition compared to Rav-Acha et al. (2006) using additional terms. However, these two studies only focused on theoretical improvement without any effort on practical implementation, and so the authors unified and expanded on their previous research in Pritch et al. (2008)

by providing an analysis of computation performance. Tubes were shifted by jumps of 10 frames and moving object detection was applied to every 10th frame, thereby reducing image resolution, etc. Even though it is not sufficient for full adaptation to real world applications, the proposed approach became more applicable to video surveillance scenarios by the performance improvement. Their study also made a positive contribution to the field by providing an analysis of run-time performance of both the on-line and off-line steps in the method.

Subsequently, they offered activity clustering in order to display similar activities together (Pritch et al., 2009). Appearance and motion features were used for clustering, and they provided the opportunity to display a video synopsis of the same person's activities or all of the activities in the same direction. Differently from previous approaches, long tubes were divided into 'tubelets', which were subsets with a maximum of 50 video frames. As clustering similar activities was novel in video synopsis at that time, they contributed to the field by providing a different perspective on existing studies.

The studies mentioned up to this point are by the authors who proposed video synopsis for the first time. Even so, there are still limitations such as time consuming optimization on video with dense activity, huge memory requirement, and uncertainty on determination of video synopsis length, although they improved on their first proposed approach with several subsequent studies. Their studies are important as they were pioneering to following studies and helped to build the principal methodology adopted by the following studies over a long period of time.

Xu et al. (2008) formulated the optimization problem of activities in terms of set theory, in which a universal set representing optimal temporal positions of the activities was obtained. The main difference from the preceding approaches is that temporal consistency was not considered on rearrangement of the activities. Even though a comparison of results with Pritch et al. (2007) was provided in which their method outperformed the classical one, their study did not attracted much attention and was not adopted by following studies. The probable reason for this was their simple optimization method to obtain local optima compared to global solution of Pritch et al. (2007).

Yildiz et al. (2008) applied a pixel-based analysis instead of an object-based one for activity detection. Input video was shrunk to only obtain the parts with high activity by extracting horizontal paths with minimum energy in video frames. They extracted the inactive parts of the video instead of temporal shifting of the activities. A pipeline-based framework was proposed to obtain real-time video synopsis with low memory consumption (Vural and Akgul, 2009). This study was extended to integrate with an eye tracking technology which was able to detect video parts that the operator did not pay attention to or vice versa. In this way, they provided the opportunity to cluster similar activities to be displayed together in the video synopsis. Their approach applied pixel-based optimization without object boundary information. Therefore, object unity might be broken in the video synopsis. Visual quality of the generated video synopsis was lower than object-based approaches, especially on scenes with high activity density.

Rodriguez (2010) contributed to the field by using an object detection method unaffected by camera motion, thus activities obtained from moving cameras could be displayed in the video synopsis. A template for a matching-based clustering method was also used to group similar activities used in the video synopsis. Chou et al. (2015) proposed the clustering of similar activities. Four regions in a camera view were first defined as possible entrance and exit locations, then activities were clustered by these regions. They used a method to cluster similar trajectories with different sampling rates, speeds, and sizes to achieve optimal results for their video synopsis. Lin et al. (2015) also proposed an approach using clustering activities with novel methods for anomaly detection, object tracking, and optimization in a video synopsis. Learning-based anomaly detection was applied to detect activities which were later clustered using predefined regions of the scene similar to the previous approach by Chou et al. (2015) using entrance and exit regions. Even though different activity clustering criteria are used in these mentioned methods, their main purpose was to make video synopsis easier to view by displaying activities with similar properties together. Besides using an additional activity clustering step in their methodology, they contributed to the field by the adaptation of clustering metrics to optimization. Their methods open new paths of investigation and possible improvements.

Another approach considering activity interactions on optimization phase was proposed by Fu et al. (2014). A motion structure term was added to the energy function to preserve interaction between activities. The motion structure term forces the retention of activities in the video synopsis that interacted in the source video. In this way, coherence of content is better preserved but optimization becomes more complicated as the length of activities shifted in the time domain increases by forcing the retention of activity interaction.

Kasamwattanarote et al. (2010) used the term 'tunnel' instead of tube and focused on real-time activity detection. Collisions between activities were embedded into recorded video during the video analysis phase. A renderer specified for video synopsis was proposed to extract embedded data in the recorded video and apply tunnel rearrangement. Ghost objects that decreased visual quality occurred in their video synopsis because of the inefficiency of the object detection method that they used. Despite the authors' declaration that their proposed video summarization method processed real-time, only the object detection part worked in real-time as most of the other offline optimization approach was off-line.

Differently from general tradition of temporal shifting in video synopsis, Nie et al. (2013) changed both the temporal and spatial positions of the activities in order to prevent collisions. Background belonging to the activities that had been spatially shifted was expanded to keep the background consistency. A synthetic background expansion was applied until there was enough space to put all activities into without any collisions, as shown in Fig. 3. Their method is the only one to shift the spatial position of the activities. Activity collisions were minimized in this way but their novelty also brought some shortcomings such as changing the background may damage the understanding of a scene since the background was extended to regions that did not have activity in the sample images. The mentioned extension could not be applied if there were no available regions without activity, thus application of the proposed method is limited to only specific scenes.

Li et al. (2016) proposed a different approach to solving the object collision problem in video synopsis in which colliding objects were scaled down in order to minimize the collision. A metric representing the scale down factor of each object was used in the optimization step. Even though the object collision problem was minimized technically, the proposed method might disturb the user. For instance, a reduction in object size causes an artificial view of the video synopsis as a car and a person that appear close in the scene might have similar sizes. Nevertheless, even this situation is prevented to a certain degree by an additional metric. He et al. (2017a,b) took activity collision analysis one step further by defining collision statuses between activities such as collision-free, colliding in the same direction, and colliding in opposite directions. They also proposed a graph-based optimization method by considering these collision states to improve the activity density and put activity collisions at the center of their optimization strategy.

Hence, a more detailed analysis of activity collision was provided compared to other video synopsis studies. Besides improvements by minimizing collisions, other metrics such as activity cost, chronological order, etc. were ignored. Therefore their optimization method still needs to be improved to find the optimal rearrangement.

Huang et al. (2014) emphasized the importance of on-line optimization techniques which enable tube rearrangement at the time of detection without any need to wait before starting optimization. Moreover, a synopsis table representing activities with their frame numbers for each pixel was proposed. Even though rearrangement obtained the local optimum, video synopsis could be generated a real-time video synopsis while activity analysis was being processed. The biggest problem with

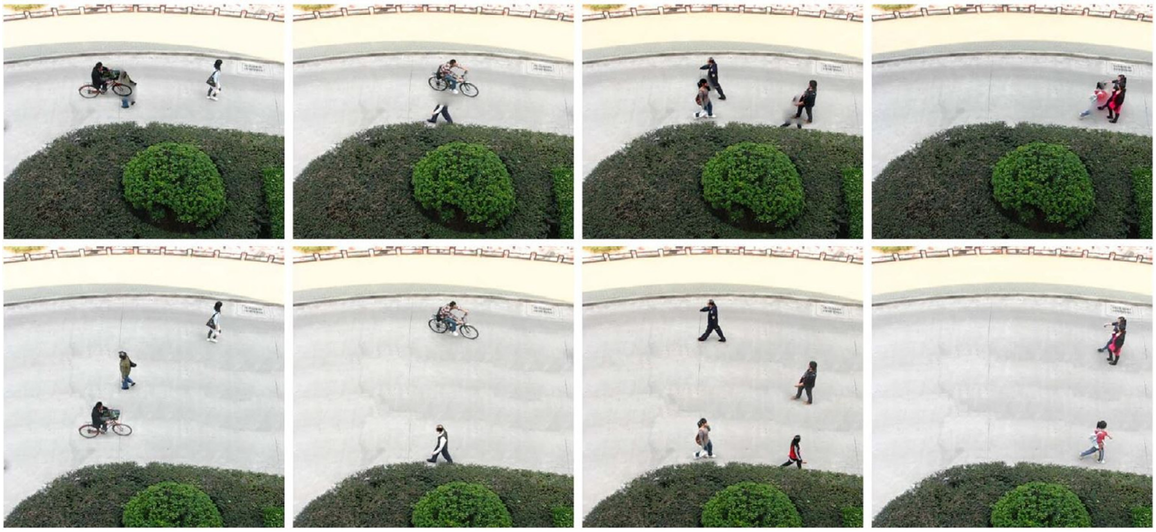


Fig. 3. Spatially shifting colliding objects (Nie et al., 2013).

their on-line method was completely ignoring activity collision situations in order to improve run-time performance, and another deficiency of the proposed optimization method was using manually determined threshold values instead of a more complex decision mechanism. With this in mind, a tradeoff between run-time performance and condensation ratio arose that decreased precision.

Zhu et al. (2014) mentioned deficiency in video synopsis due to a single-camera view since when considering video surveillance applications, an activity generally happens in more than one camera view. Thus, they proposed a multi-camera video synopsis approach with a panoramic view constructed using homography between partially overlapping camera views. Activities from different cameras were associated via trajectory matching in overlapping camera views. They also proposed a key frame selection approach for the activities whereby key frames of an activity in which the appearance or motion of an object is changed significantly are used instead of all of the frames for reducing redundancy of consecutive frames. Similarly, Zhu et al. (2016a) proposed a multi-camera video synopsis approach using a timestamp selection method to find critical moments of an activity. Key timestamps were defined as when objects first appear, the merge time with any other object, and the split and disappear time in the video. Unlike Zhu et al. (2014), object re-identification using visual information was applied between camera views. The energy function for optimization was also improved so as to be adaptable with multi-camera topology. The chronological order of objects was kept not only in one camera view but also among different camera views.

Hoshen and Peleg (2015) suggested a multi-camera video synopsis approach which defined a master camera and slave cameras around the master. Once an activity was detected in the master camera, a video synopsis containing activities of slave cameras belonging to related time period is generated. Although object re-identification between the cameras was not applied, they aimed to provide a wider perspective on the activity of master. Mahapatra et al. (2016) offered another video synopsis framework on multiple cameras having overlapping field-of-views for which a common ground plane via a homography between camera overlaps was generated. Activities were classified into seven categories, namely walking, running, bending, jumping, hand shaking, one hand waving, and both hands waving. Thus, they provided video synopsis of specific activity types.

Multi-camera video synopsis approaches are more applicable to real-world applications when considering distributed video surveillance networks. Nevertheless, optimization becomes more complicated with additional metrics used for the association of objects in different cameras. Another important point is overlapping of camera views. Studies applied to non-overlapping camera views seem more efficient as they have one less restriction on camera topology.

Different than the approaches explained up to now, Lin et al. (2017) mainly focused on acceleration of computing speed of video synopsis via a distributed processing model. Their framework included computing and storage nodes created for distributed computation in which the nodes represented different computers on a network or application threads. Their video synopsis algorithm was divided into several steps such as video initialization, and object detection, tracking, classification, optimization, etc., which were computed in a distributed fashion. Input video was segmented and each segment analyzed on a different node and tubes generated on each node were stored on storage nodes. Finally, another node generated the final video synopsis using data on the storage node. The region of interest of the scene was also defined in order to reduce the region of input processing. Furthermore, video size and frames per second were also reduced to increase performance without affecting the accuracy of object detection. This was the first study to perform a video synopsis with a distributed architecture and was innovative when considering the distributed camera topology of video surveillance applications. This study provided the opportunity to apply high precision but time consuming optimization methods close to real-time performance.

Besides, there are video synopsis approaches which work on compressed domains (Wang et al., 2013a,b; Zhong et al., 2014; Liao et al., 2017). They emphasized that video decoding increases the complexity of the approach and makes it hard to work in real-time, thus activity detection was carried out on compressed video and required that flags were set for use in the optimization step. Partial decoding was applied to improve the run-time performance of the approaches. Nevertheless, their object detection methods in the compressed domain were simple compared to pixel-based methods. Because inefficiency in object detection directly affects video synopsis performance, these methods need more improvement on precision.

The video synopsis approaches mentioned so far have commonly focused on the optimization step of the flow. Nevertheless, there have been studies that have focused on other steps such as background generation and object tracking specified for video synopsis. Feng et al. (2010) proposed a background generation approach aimed at choosing video frames with the most activity and representing changes in the scene. Thus, they later propose sticky tracking to minimize the object blinking problem which causes ghost objects in video synopsis (Feng et al., 2012). Objects with intersected trajectories were merged as a unique activity to be used in the video synopsis, the purpose is not to obtain perfect object tracking but to provide activity coherence.

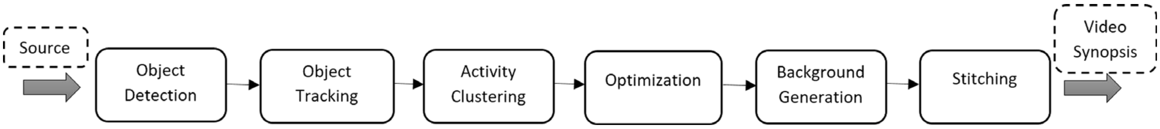


Fig. 4. Video synopsis methodology.

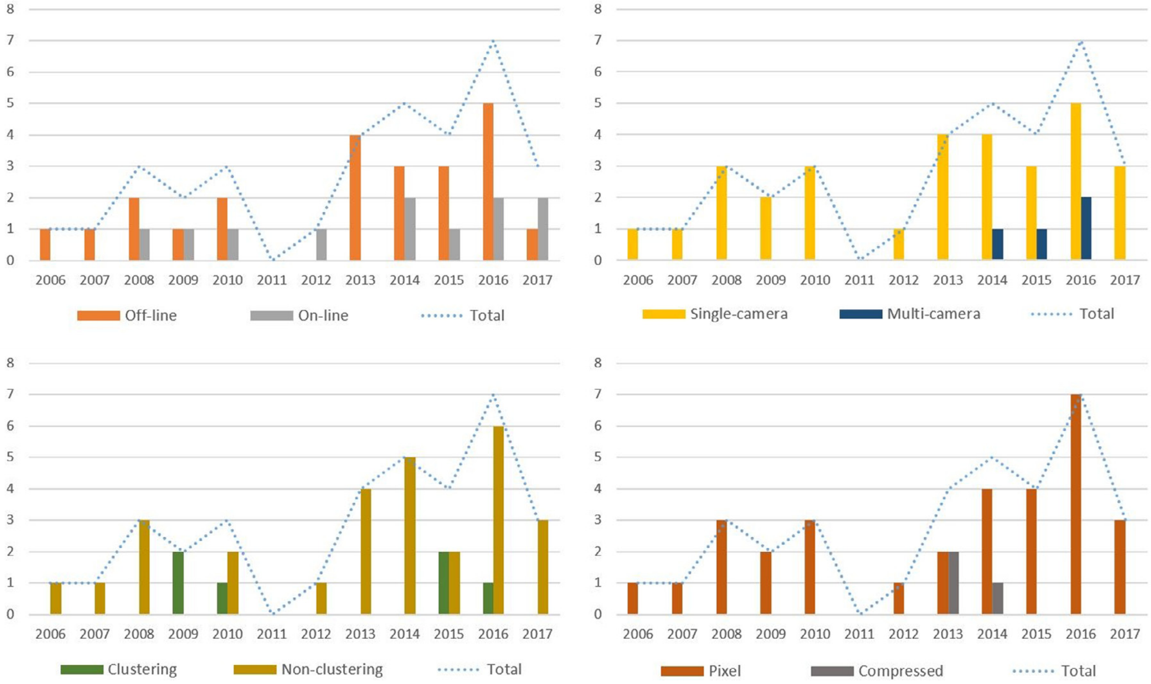


Fig. 5. Distribution of video synopsis studies over the years. The blue line shows the total number of studies belonging to the corresponding year. Colors representing each aspect are shown in the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Baskurt and Samet (2018) proposed another object tracking approach specified for requirements of video synopsis. Their approach focused on long term tracking to represent each target with just one activity in video synopsis. The target object was modeled with more than one correlation filter which represent the different appearances of the target during the tracking. Robustness across the environment challenges such as illumination variation, scale and appearance changes was obtained in this way. Lu et al. (2013) focused on object detection artifacts such as shadow and interruption of object tracking which reduce efficiency content analysis. They proposed support for both motion detection and object tracking methods with additional visual features in order to eliminate shadow and increase the robustness of the tracking method against collision. Baskurt and Samet (2017) also focused to increase robustness of object detection by proposing an adaptive background generation approach. Hsia et al. (2016) concentrated on efficiently searching an activity database to generate video synopsis. A novel range tree approach was proposed whose main purpose was to find the tubes selected by the user in an efficient way and to reduce the complexity of the algorithm.

These studies have made an important contribution to other video synopsis studies. Each step in the video synopsis pipeline feeds others, thus failure in the steps especially before optimization such as object detection and object tracking directly affect video synopsis output. Improving the optimization step is not enough to obtain the best results in a video synopsis. Therefore, the specific adaptation of commonly known methods from different fields such as object detection and tracking makes important contribution to the study of video synopsis.

Finally, Zhu et al. (2013, 2016b) emphasized using support of non-visual data in video synopsis. Information on weather forecasts, traffic monitoring, and scheduled public events were associated with visual data to cluster activities and achieve better video content analysis. Even though using non-visual data helped activity clustering or provided a better understanding of the activities, these studies did not mainly focus on video synopsis, rather on data acquisition and association with the activities.

To summarize this section, an overview emphasizing novelty and contribution of video synopsis approaches was presented. Studies were summarized with comments on both their pros and cons. It is evident that there is important variety in the studies as some of them focused on several steps in their methodology whereas others aimed to improve performance efficiency. While one branch of studies tried to move the video synopsis approach to multi-camera topology, others focused on contributing by changing the input data domain. Furthermore, some studies suggested performing an additional activity clustering step to display similar activities together. In this sense, recent literature on the field of video synopsis can be divided into several categories that are analyzed and discussed in Section 4.

3. The methodology of video synopsis

In this section, we analyze methodology of video synopsis described in Fig. 4. Video synopsis generation starts with object detection, then object tracking is applied to create activities. Next, activity clustering is applied to display similar activities together followed by optimization of the selected activities to obtain optimal temporal rearrangement. Afterwards, a time-lapse background representing the time period of the selected activities is created, and finally, activities are stitched to the generated background. Table 1 gives an overview of the methods used in object detection, object tracking and optimization which are the most critical steps of the methodology.

Table 1

Overview of object detection, object tracking and activity optimization methods used in video synopsis.

Object detection		Object tracking		Optimization	
Algorithm	Studies	Algorithm	Studies	Algorithm	Studies
Pixel difference with temporal median (Rav-Acha et al., 2006)	Rav-Acha et al. (2006), Nie et al. (2013), Li et al. (2016)	Multiple object tracking (Yang et al., 2005)	Xu et al. (2008), Feng et al. (2010)	Simulated annealing (Kirkpatrick et al., 1983)	Rav-Acha et al. (2006), Pritch et al. (2007), Rodriguez (2010), Feng et al. (2010), Wang et al. (2013a,b), Zhong et al. (2014), Zhu et al. (2014), Li et al. (2016), Mahapatra et al. (2016), Li et al. (2016), Liao et al. (2017)
Background cut (Sun et al., 2006)	Pritch et al. (2007, 2008, 2009), Feng et al. (2010, 2012), Zhu et al. (2014)	Tube generation (Pritch et al., 2008)	Pritch et al. (2008, 2009), Hoshen and Peleg (2015)	Mean shift (Xu et al., 2008)	Xu et al. (2008)
GMM (Stauffer and Grimson, 2000)	Xu et al. (2008), Lu et al. (2013), Fu et al. (2014), Zhong et al. (2014), Chou et al. (2015), Li et al. (2016), Hsia et al. (2016), Tian et al. (2016), Lin et al. (2017)	Clifford worm extraction (Rodriguez, 2010)	Rodriguez (2010)	Dynamic programming (Yildiz et al., 2008)	Yildiz et al. (2008), Vural and Akgul (2009)
Gradient magnitude (Yildiz et al., 2008)	Yildiz et al. (2008), Vural and Akgul (2009)	Euclidean distance (Kasamwattanarote et al., 2010)	Kasamwattanarote et al. (2010)	Greedy optimization (Pritch et al., 2008)	Pritch et al. (2008)
Optical flow (Papenberg et al., 2006)	Rodriguez (2010)	Sticky tracking (Feng et al., 2012)	Feng et al. (2012), Wang et al. (2013b), Zhu et al. (2016a), Fu et al. (2014)	Packing cost (Pritch et al., 2009)	Pritch et al. (2009)
HOG for human detection (Dalal and Triggs, 2005)	Kasamwattanarote et al. (2010), Hoshen and Peleg (2015)	Tracking with shadow detection (Kaewtrakulpong and Bowden, 2002)	Nie et al. (2013)	Film map generation (Kasamwattanarote et al., 2010)	Kasamwattanarote et al. (2010)
Quadtree (Finkel and Bentley, 1974)	Wang et al. (2013a)	Particle filter (Perez et al., 2002)	Lu et al. (2013)	Online tube filling (Feng et al., 2012)	Feng et al. (2012), Fu et al. (2014)
Motion vector based LBP (Wang et al., 2013b)	Wang et al. (2013b)	3D graph-cut (Zhong et al., 2014)	Zhong et al. (2014)	Alpha–beta swap graph-cut (Boykov et al., 2001)	Nie et al. (2013)
Hierarchical background modeling (Chen et al., 2007)	Huang et al. (2014)	Multi-feature graph-based object tracking (Taj et al., 2006)	Zhu et al. (2014)	Synopsis table-(Huang et al., 2014)	Huang et al. (2014)
Frame difference (Yao et al., 2014)	Yao et al. (2014)	Using proposed synopsis table (Huang et al., 2014)	Huang et al. (2014)	Genetic algorithm (Whitley, 1994)	Yao et al. (2014), Xu et al. (2015), Tian et al. (2016)
Min-cut (Kolmogorov and Zabin, 2004)	Fu et al. (2014)	Kalman filter (Kalman, 1960)	Yao et al. (2014), Li et al. (2016), Tian et al. (2016), Li et al. (2016)	Trajectory clustering (Chou et al., 2015)	Chou et al. (2015)
Adaptive background modeling (Stauffer and Grimson, 1999)	Hoshen and Peleg (2015)	Tracking (Chou et al., 2015)	Chou et al. (2015)	Fast greedy approach (Hoshen and Peleg, 2015)	Hoshen and Peleg (2015)
Abnormal activity detection (Lin et al., 2015)	Lin et al. (2015)	Blob sequence optimization (Lin et al., 2015)	Lin et al. (2015)	Abnormality-type based video synopsis (Lin et al., 2015)	Lin et al. (2015)
SILTP (Liao et al., 2010)	Zhu et al. (2016a)	LMS tracking (Hsia et al., 2011)	Hsia et al. (2016)	Low-complexity range tree (Lin et al., 2015)	Hsia et al. (2016)
Human detection (Mahapatra et al., 2014)	Mahapatra et al. (2016)	Clustered track extraction (Mahapatra et al., 2016)	Mahapatra et al. (2016)	Simple greedy approach (Zhu et al., 2016a)	Zhu et al. (2016a)
R-CNN (Girshick, 2015)	Jin et al. (2016)	Chi-square distance (Jin et al., 2016)	Jin et al. (2016)*	Table-driven approach, CBGC (Mahapatra et al., 2016)	Mahapatra et al. (2016)
ViBe (Barnich and Van Droogenbroeck, 2011)	He et al. (2017a,b)	Multiple pedestrian tracking (Zhang et al., 2015)	Lin et al. (2017)	Simple tube generation (Jin et al., 2016)	Jin et al. (2016)
Motion vectors (Liao et al., 2017)	Liao et al. (2017)	NN-based tracking (Choeychuen et al., 2006)	He et al. (2017a,b)	Huiyan (Lin et al., 2017)	Lin et al. (2017)
		VBF Deletion on 3D Graph (Liao et al., 2017)	Liao et al. (2017)	Graph coloring approach (He et al., 2017a)	He et al. (2017a)
				Potential collision graph (He et al., 2017b)	He et al. (2017b)

Object detection is used as the first step in the algorithm flow of video synopsis. The preference in most of the methods is to use motion for defining the objects. Simple motion detection methods such as pixel difference, temporal median, etc. show poor performance in complex scenes with dynamic background objects, dense motion, and significant variation of illumination. These environmental difficulties are handled better by more complex background modeling algorithms provided in Table 1. Human detection methods instead of motion detection are also used for object detection. They provide more precise results as the false detection ratio is lower. Motion detection methods are more likely to be affected by artifact as they provide lower level image analysis compared to human detection methods. On the other hand, using motion for object detection provides the opportunity of using different types of objects as targets. Motion detection methods are also scene independent compared to template matching or training-based methods that need target-specific training beforehand.

After detecting targets, object tracking associates detected objects in consecutive frames to build object trajectory, which represents an activity in a video synopsis. It has direct effect on video synopsis performance since tracking failures that cause broken trajectories, mismatch of colliding objects, etc. decrease their accuracy and creating more than one activity for the same object breaks the semantic completeness. These deficiencies also make the optimization problem more difficult as redundant activities will be generated. Therefore, robust object tracking methods specified for video synopsis significantly contribute to the accuracy of a video synopsis.

Some of the video synopsis approaches cluster the activities according to different criteria such as motion direction, action type, target type, etc. Their point is to improve visual quality of video synopsis as viewing similar activities together makes the video easier to trace by the user. Details of the approaches that apply activity clustering are discussed in Section 4.4.

Optimization step which is the most important part of video synopsis is applied after obtaining the activities of source video. Optimization aims to find best re-arrangement of the activities in order to display most of them in the shorter time period with minimum collision. Activities are shifted in time domain to place in optimal position in video synopsis. Finding optimal position of the activities are determined by some constraints such as background consistency, spatial collision, temporal consistency, etc. Detailed analysis of the optimization approaches used in video synopsis is provided in Section 4.1.

A time-lapse background representing activities and scene changes covering a corresponding time period needs to be created after finding the optimal places for the activities. Video synopsis output seems more natural with better background generation considering that the output is a synthetic video after rearrangement of the activities belonging to different time periods. Improvement of background generation provides a better user experience as visual inconsistency is minimized. Background generation does not affect the condensation performance of video synopsis, it just provides better visual quality. However, it has not been applied in most of the studies in the literature.

Stitching objects to a time-lapse background is the last step in the video synopsis flow. Stitching does not have an effect on the precision of the approaches, it just improves the visual quality of the output. Therefore, no great attention has been paid to improving this step. Most of the studies did not apply a specific stitching or blending algorithm other than pixel exchange of the object and the generated background. However, using a proper stitching method increases the quality of output as objects from different time periods are displayed at the same time over a unique background.

Methodology of video synopsis commonly applied in the literature was explained in this section. Next section categorizes the literature of video synopsis from different aspects such as optimization type, camera topology, input data domain and the activity selection criteria. Detailed analysis of the video synopsis approaches according to mentioned aspects is provided.

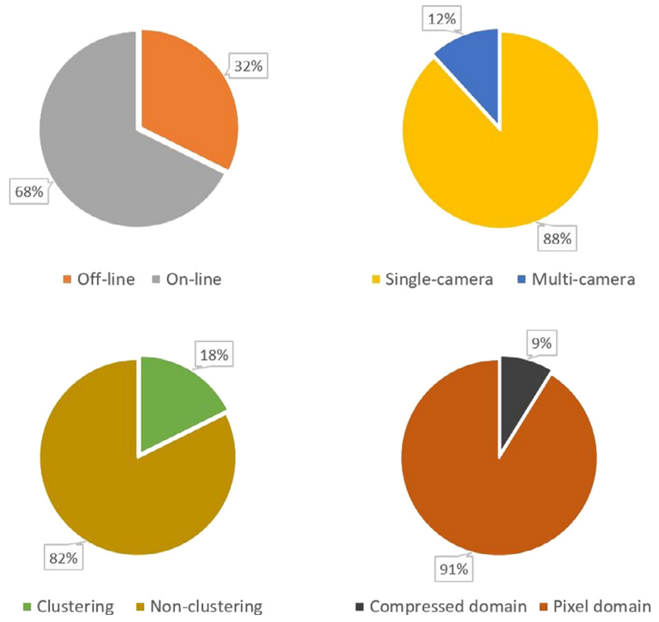


Fig. 6. Ratio of the number of studies for each optimization type, camera topology, activity clustering, and input data domain, respectively.

4. Classification of video synopsis approaches

Video synopsis approaches can be divided in four groups by content, namely optimization type, camera topology, input data domain, and activity clustering. The distribution of the studies over the years is provided in Fig. 5, and the ratio of publications according to four mentioned groups is shown in Fig. 6.

It is evident that off-line optimization approaches have been more dominant than on-line approaches. Although on-line approaches appeared early on, they have always been in a minority. Similarly, single-camera approaches are more popular against multi-camera approaches. There were no multi-camera approach until 2014 even though video synopsis was first proposed in 2006. Rare interest on approaches using the compressed domain appeared in 2013, 2014 and 2017. Also, there has been no consistent trend on video synopsis approaches that applies activity clustering as they appear in specific time periods. A general overview shows that while there is no significant trend on approaches to the compressed domain and activity clustering, number of on-line and multi-camera approaches has increased in recent years. This situation gives us a clue about future trends in the field of video synopsis. Following subsections provide detailed analyses on the four mentioned aspects.

4.1. Aspect 1: Optimization type

Optimization is the most important step in video synopsis. All optimization methods aim to obtain mapping of activities from the source video to proper positions in the video synopsis. The final goal is to display all of the activities in the shortest time period while avoiding collisions as much as possible. Generally, the optimization problem is defined as minimization of the global energy function that consists of several costs such as maximum activity, background and temporal consistency, and spatial collisions. While some studies used additional costs, others did not use all of them. A brief explanation of commonly used costs is provided as follows:

- The activity cost forces the inclusion of the maximum number of activities in a video synopsis. Activities staying outside are penalized by this term. Leaving out any activity in video synopsis approaches is not desired therefore, this term is used by almost all approaches.

- The aim of the background consistency cost is to guarantee stitching of tubes to background images having a similar appearance. This term measures the cost of stitching an object to the time-lapse background. Inconsistency between a tube and the background is penalized as it is assumed that each tube is surrounded by pixels from its original background.
- The role of the temporal consistency cost is to preserve the temporal order of the activities, therefore activity shifts that break the temporal order are penalized. Changing temporal order of the activities in optimization phase may provide more compact representation by increasing variation of activity sequences. On the other hand, preserving chronological order is important for causality relation of the activities. Analyzing the activities that have interaction in the source video is easier if the temporal consistency is preserved. Approaches generally use a weight parameter for this term in order to balance the semantic integrity and the optimal activity representation of the video synopsis.
- The collision cost prevents spatial collisions of the activities in order to provide better visual quality. Spatial collisions of the activities are penalized by increasing total energy. Handling spatial collision of the activities is main problem of the optimization step. Activities are generally collided with each other considering the crowded scenes captured by the surveillance cameras. Allowing collisions in video synopsis decreases the visual clarity and the traceability of the activities even it provides more compact output with higher number of activities in shorter time period. Nevertheless, video synopsis longer than source video may be created if the spatial collision is completely prevented especially for the crowded scenes. This term is placed in the center of activity optimization phase as it is the most challenging problem in the representation. Majority of the approaches focus on finding optimal solution for activity collision.

While the activity and the background consistency costs are calculated for each activity separately, the temporal consistency and the collision costs are calculated between the activities in video synopsis. Weight parameters are generally used especially for temporal consistency and the spatial collision costs to find optimal solution. An illustration of different activity representations that can be obtained after minimization of the same energy function with different weights of the temporal consistency cost is provided in Fig. 7. Scenarios for preserving chronological order absolutely (a), preserving chronological order partially (b) and ignoring chronological order (c) are represented. Fig. 7 shows that displaying activities in same chronological order of the source video costs longer video synopsis.

All the activities are represented in 28 frames in this case as illustrated in Fig. 7(a). Ignoring chronological order of the activities by lower weight parameter provides more compact representation (18 frames) as shown in Fig. 7(b). On the other hand, ignoring chronological order completely ends up with shortest representation of the activities as shown in Fig. 7(c). Video synopsis length is determined by the length of the longest activity (13 frames) in this case, but displaying all the activities at the same time causes a chaotic view because of the spatial occlusion. Even minimum energy is obtained in the third case, visual quality of the representation is the worst one. This illustration also proves the importance of using several costs together in energy function in order to find optimal solution for both compactness and the visual quality. For instance, considering collision cost with temporal consistency in this case would provide better solution even higher energy is obtained at the end of the optimization phase. As is seen, optimal representation of the activities depends on several conditions that makes the problem non-linear.

Finding minimum energy provides optimal solution as undesired situations are penalized by the costs described above. Thereafter, online or offline optimization methods are used to minimize the defined energy function. Off-line methods require the analysis of the entire video before starting optimization. All activities must be detected and ready for use in global optimization using all of the data at once. Two main problems with these approaches are the huge memory requirement for storing all of the activities and the time consuming processing phase to search all of them. The computational complexity of off-line methods is extremely high and exponentially proportional to the total number of activities. The constraints aforementioned make these methods difficult to apply to video surveillance cameras in real-time. Even though the activity detection part of off-line methods are generally performed on-line, these methods cannot be applied to real-world application efficiently because of the time consuming and computationally expensive optimization phase.

On-line methods follow a step-wise optimization strategy updated by each activity detection. Detected activities can be shifted by rearranging existing activities in memory. Unlike minimizing the global energy function of off-line methods, applying local optimization does not require huge memory and high computational power. Therefore, on-line video synopsis methods can be directly applied to endless video streams received from surveillance cameras.

To summarize, both methods have pros and cons. Off-line methods obtain better condensation ratios than on-line methods as more detailed optimization is performed but they are difficult to be applied directly to real-world applications compared to on-line methods. On the contrary, optimization precision needs to be improved in on-line methods. Detailed analysis of off-line and on-line methods used in video synopsis is provided in the following subsections.

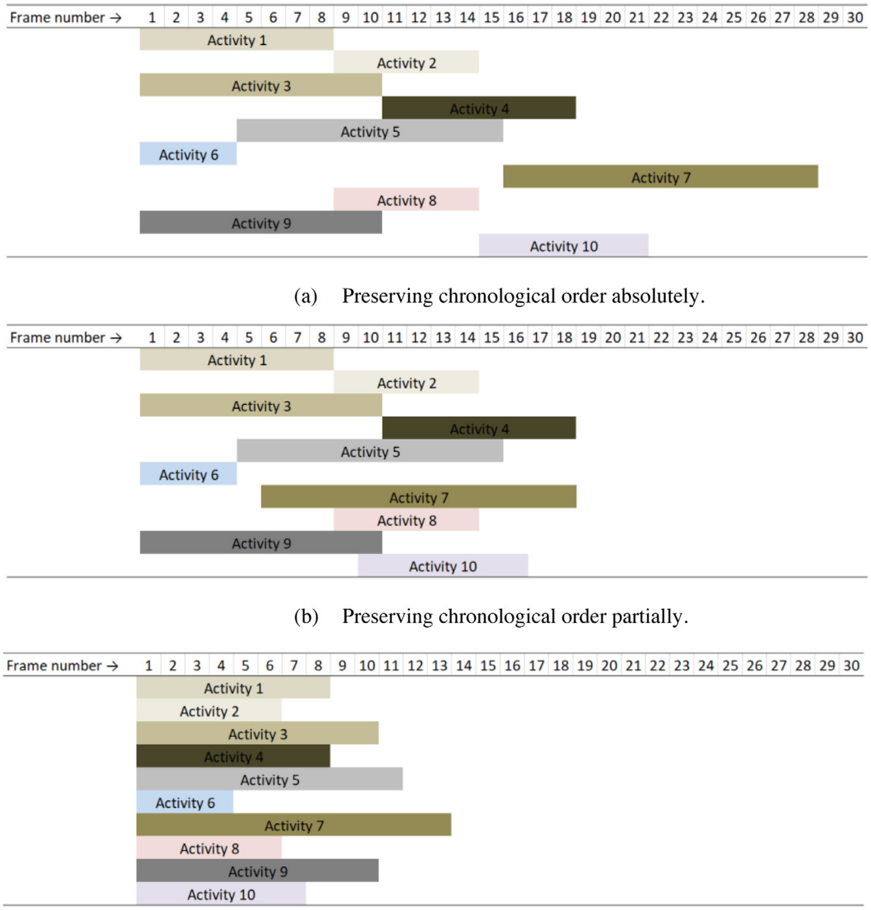
4.1.1. *Off-line optimization*

Simulated annealing (Kirkpatrick et al., 1983) is the predominantly used off-line optimization method in video synopsis studies, as shown in Table 1. Early studies used simulated annealing and so most of the following ones also adopted it. Simulated annealing, mean shift, greedy algorithms and the graph-cut based optimization methods used in video synopsis are similar approaches that aim to model all possible temporal mappings of the activities. In these methods, a random initial state is selected and the initial energy cost is calculated according to the defined energy function. After that, several iterations are applied to find optimal temporal mapping of the activities which is represented by minimum energy. These combinatorial optimization methods are effective to find global optimum, but they are time consuming and their convergence is very slow especially in the case of higher number of activities.

By comparison, genetic algorithms (Whitley, 1994) have been used and have shown better performance than simulated annealing for both condensation ratio and computational complexity. Although both genetic algorithms and the simulated annealing are combinatorial and based on randomness, genetic algorithms does not apply just a simple random search. Searching in genetic algorithms is directed towards the optimal solution by using a random method that creates relation between the optimal solutions of two iterations. Also, optimal solution of each operation does not affected by the initial state due to mutation operation. Therefore, genetic algorithms provide more compact video synopsis comparing to simulated annealing.

The ones mentioned above are optimization methods commonly used in different areas and also adapted into optimization step of video synopsis. Apart from these, there are optimization methods specifically proposed for finding optimal activity rearrangement of video synopsis. The packing cost proposed by Pritch et al. (2009) aims to find optimal re-arrangement of the activities that are already clustered according to their similarities. Their method gives priority to the longest activities first while putting them into video synopsis. Temporal overlap is calculated between the activity clusters that means all the activities belonging to the same cluster are placed at once. In this way, optimization problem is divided into two parts such as clustering and finding optimal position. Considering clusters as an activity makes the optimization easier with fewer components to be rearranged.

Film map generation (Kasamwattanarote et al., 2010) uses a direct shift collision method to calculate the occlusion of activities after which



(a) Preserving chronological order absolutely.

(b) Preserving chronological order partially.

(c) Ignoring chronological order completely

Fig. 7. Activity representations for different weights of temporal consistency cost.

a special representation consisting of the objects' depths and video frames is used to rearrange them. Optimal position of each activity is determined by comparing its depth map with the current depth map of the activities belonging to video synopsis. Dynamic programming is used to calculate occlusions. Condensation ratio of the method is lower than packing cost method according to the experimental results.

Trajectory clustering proposed by Chou et al. (2015) applies temporal shifts to the objects iteratively until no collisions remain, and they only used the collision cost on the tube rearrangement phase. Their iterative activity re-arrangement method does not perform as well as combinatorial methods mentioned above. Because brute force searching is applied instead of using any direction approach to reach optimal solution faster.

Abnormality-type-based video synopsis proposed by Lin et al. (2015) uses temporal consistency, collisions, and an additional cost representing the similarity of activities to define the energy function. A similarity metric enables the displaying of similar activities together in a video synopsis. They also assigned weights to each cost to find the minimum energy by trying different weight combinations.

For the low-complexity range tree proposed by Hsia et al. (2016), they defined a simple energy function by only considering the collision cost to find the optimal temporal position of the activities. The defined energy function was calculated between each tube, then the tubes were shifted repeatedly until the optimal position with no overlap or acceptable overlapping decided upon by the user was found. The range tree approach mainly focused on finding the tubes selected by the user in an efficient way rather than improving the optimization operation.

Consequently, computational complexity is a major problem of off-line methods. Time consuming global energy minimization makes these methods inapplicable to endless camera streams. Off-line optimization processes iterate in a loop to minimize global energy consisting of several costs penalizing undesired situations. Any change in tube arrangement in a loop requires re-computation of the energy function, which makes the solution time consuming. In addition, it is evident that computational cost is proportional to the number of activities.

4.1.2. *On-line optimization*

Online tube filling proposed by Feng et al. (2012) applied step-wise optimization by finding temporal shifts of the currently detected activity among activities collected up to that point. Two buffers named L1 and L2 were defined to store shifted activities, and while L1 had limited capacity, L2 did not. An energy function consisting of collision cost was defined and three step optimization consisting of a greedy algorithm, roulette wheel selection (Mitchell, 1998), and collision checking was applied.

The synopsis table proposed by Huang et al. (2014) represented each pixel in a video frame with an object index and the frame number that the object occupied in this position. The synopsis table was updated with the detected objects in each video frame. A posterior probability function was defined to estimate whether a detected object was an instance of an existing activity or a new one. Simple tube generation proposed by Jin et al. (2016) also used a synopsis table representing each pixel with an occupying object index similarly (Huang et al., 2014). Only one object was allowed to occur in a pixel. Therefore, colliding objects were stored in a buffer until they were available to display in the video synopsis.

The table-driven approach proposed by Mahapatra et al. (2016) defined a collision table in order to rearrange activities without any collisions. They also proposed a contradictory binary graph coloring approach for a multi-view video synopsis as a graph with each vertex representing an activity where each edge denotes a collision with another activity. Subsequently, the graph coloring approach was applied to generate a video synopsis without any collisions. He et al. (2017b) proposed an on-line video synopsis method by creating a potential collision graph to analyze the collision relationships between activities with collision cost as a unique term being evaluated. Two types of tube relationship: collision free and collision potential (divided in two cases: colliding in the same direction and colliding in the opposite direction) were defined. The created tubes were arranged on-line with regard to their collision relationship; collision-free tubes were placed in any position while collision-potential tubes were placed outside of the collision period. The authors later extend their work by proposing L(q)-coloring of a potential collision graph (He et al., 2017a), which is a graph created using all of the extracted tubes from the original video. Connections between graph nodes were created regarding the relationships between the related tubes, after which the tube rearrangement problem was formulated and solved as a graph-coloring problem.

On-line methods are aimed at performing rearrangement of each activity detected one by one. It starts with activity detection being realized on-line while a video stream is being received. Rearrangement is performed on an existing local activity set collected until that moment, thus all of the activities are not needed beforehand. This situation requires less memory and reduces computational complexity. On the other hand, achieved local optimization does not provide as high a condensation ratio as off-line methods.

4.2. Aspect 2: Camera topology

Most video synopsis approaches are applied to a single-camera view. Only four studies have been carried out on multi-camera topology. Zhu et al. (2014) applied video synopsis on multiple camera whose views were overlapping. In their study, all camera views were transformed to a common ground plane created using a homography between the camera views. Object association between cameras was performed by trajectory matching in the overlapping areas, although no visual information about the object was used for association. Mahapatra et al. (2016) proposed another multi-camera approach for an overlapping camera network. Similar to Zhu et al. (2014), a common ground plane using a homography was created. Activities were plotted onto a bird's eye view and a corresponding camera record was also displayed separately, as shown in Fig. 8. Trajectory matching on overlapping camera regions were applied to associate objects in different camera views.

Hoshen and Peleg (2015) proposed another multi-camera approach with several slave cameras supporting one master camera. Objects in the slave camera views belonging to the same time period of master camera view's objects were displayed separately to support contextual coherence. Moreover, there was no association between the objects in different camera views.

Zhu et al. (2016a)'s study is the only one that does not expect camera overlapping and performs a visual association of objects from different camera views. The activities in each camera were extracted separately, and then joint tube rearrangement was applied to merge instances of the same activity into a unique tube. Although the activities were merged, the activity parts belonging to each camera were displayed separately on the visualization phase.

As can be seen, video synopsis approaches applied to multi-camera topology is still limited. They are more efficient as this provides a wider angle of view to help understand an entire scenario. More than one camera is often used for video surveillance in daily life, even for small areas, and activities generally cover more than one camera view. On the other hand, applying video synopsis to multi-camera topology brings some difficulties compared to single-camera approaches. Robust object tracking among different cameras is still a serious problem in the field. Failure in multi-camera object tracking damages the contextual coherence of a video synopsis, which causes a disadvantage compared to single-camera approaches. Furthermore, the optimization step for multi-camera topology becomes more complicated because of additional metrics related to the association of activities among the cameras. Consequently, further investigation into improving multi-camera video synopsis approaches are needed.

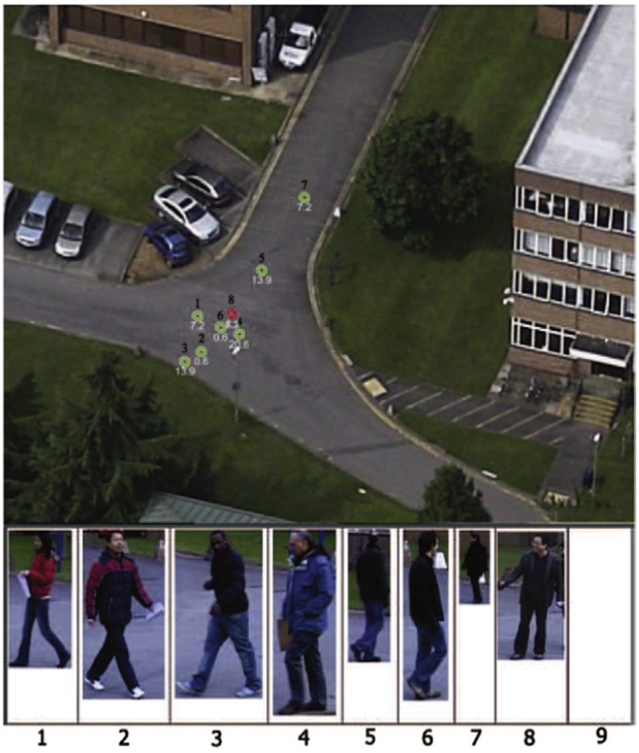


Fig. 8. Multi-camera video synopsis representation (Mahapatra et al., 2016). The image in the upper part is a bird's eye view of the scene. A preview of each corresponding object in the top view can be seen labeled with its unique id.

The distributed processing video synopsis model proposed by Lin et al. (2017) which is explained in detail in Section 2 is innovative when considering the distributed structure of video surveillance cameras all over the world and current improvements on cloud technology. A video synopsis approach performed on distributed architecture covering multiple cameras deserves further investigation according to technological hot spots in this area.

4.3. Aspect 3: Input data domain

Almost all video synopsis studies have been applied to the pixel domain. There are only four studies proposed by Wang et al. (2013a,b), Zhong et al. (2014) and Liao et al. (2017) that work directly on the compressed domain. These studies arose from high complexity of especially off-line video synopsis methods that cannot be applied directly to endless video streaming. Compressed domain based methods emphasize that video decoding causes extra computational cost and makes the algorithm hard to handle on-line streaming videos. They focus on decreasing computational complexity by performing some of the video synopsis steps on the compress domain, as illustrated in Fig. 9.

Information on each video frame such as synopsis information, the frame index in the video synopsis, and the number of activities are coded into the frames after analyzing the compressed data. Therefore, the memory requirement for storing the information to be used in the video synopsis generation phase is also minimized in compressed domain analysis.

However, condensation success with compressed-domain methods is lower than with pixel-domain ones as video analysis is performed using

Table 2
Activity clustering methods.

Existing study	Clustering
Pritch et al. (2009)	SVM (Cortes and Vapnik, 1995)
Vural and Akgul (2009)	Eye tracking (LC Technologies)
Rodriguez (2010)	Action MACH filter (Rodriguez et al., 2008)
Chou et al. (2015)	Longest common sequences (Vlachos et al., 2002)
Lin et al. (2015)	Abnormal activity classification (Lin et al., 2015)
Mahapatra et al. (2016)	SimpleMKL (Rakotomamonjy et al., 2008)

less information; even run-time performance is significantly higher. Compressed-domain methods seem more efficient for video browsing and retrieval as the required information is encoded into the frame. The performance decrease caused by the condensation ratio overrides the gain in run-time performance, thus studies on the compressed domain have not attracted too much attention. As of yet, these studies are beneficial to see effect of compressed domain analysis on video synopsis but do not seem to be promising in terms of efficiency.

4.4. Aspect 4: Activity clustering

The video synopsis studies listed in Table 2 use activity clustering before optimization whereby activities are categorized according to predefined metrics and similar activities are displayed together in the video synopsis. The first study in this area by Vural and Akgul (2009) used an eye gaze tracker to determine video frames that the operator paid attention to or vice versa, and activities were grouped in this direction and displayed together. The main focus of clustering is to prevent activity misses by the operator who is watching the surveillance cameras. Pritch et al. (2009) used the appearance and motion features of the activities. Support vector machine (SVM) proposed (Cortes and Vapnik, 1995) was used for activity clustering. Meanwhile, Rodriguez et al. (2008) employed an action MACH (Maximum Average Correlation Height) filter containing templates of frequency domains corresponding to defined activity groups such as running, picking up an object, entering a vehicle, etc. (Rodriguez, 2010). The detected activities were labeled according to these groups, after which those belonging to the group selected by the user were used to generate the video synopsis.

Unlike the aforementioned studies, Chou et al. (2015) used longest common sequence algorithm to cluster activities by their spatial position instead of appearance or motion features. Four different regions on the scene as possible entrance and exit regions were defined, then activities were clustered and displayed by their entrance and exit regions. Lin et al. (2015) proposed an approach to learn normal activities in a scene in a training phase, then abnormal activities were detected using the trained data. They also used spatial position of the activities in order to cluster them. Key regions were determined to define activity flow (from entrance region to exit region). Meanwhile, Mahapatra et al. (2016) used a multiple kernel learning method (Rakotomamonjy et al., 2008) for action recognition. Shape features of the activities were used to classify human actions like walking, running, bending, jumping, shaking hands, one hand waving, and both hands waving. After that, similar activities were displayed together.

Activity clustering may affect both positive and negative aspects depending on the application type. Displaying similar activities together improves the quality by providing better understanding as focusing on similar activities is easier for the user. On the other hand, the user may want to associate the activities of different groups, which is only possible by viewing all of them together. Activity clustering increases computational complexity as an additional step is included in the algorithm flow. Activity clustering provides variety in the display of the video synopsis that seems like an application level feature rather than an improvement in the methodology. Therefore it can be used as an optional step for any video synopsis application.

5. Performance metrics and datasets

Performance of video synopsis methods are generally compared according to following metrics; frame condensation ratio (FR), compact ratio (CR), overlap ratio (OR), chronological disorder ratio (CDR), time consumption, and visual quality.

FR represents the ratio of the number of frames in the synopsis to the source video; a higher reduction of frames means a smaller FR. CR measures the efficiency of tube rearrangement in terms of the activity density on each frame: the higher the CR, the more compact video synopsis. OR determines the collision degree of the activities; a smaller OR represents fewer collisions in the output, which is desired in video synopsis. CDR represents the number of activities which are chronologically disordered over all activities in the video synopsis; a smaller CDR indicates better preservation of the chronological order. Time consumption is measured per frame or the total optimization time depending on whether a study is on-line or off-line, respectively. Visual quality is a subjective metric used by some approaches such as those in He et al. (2017a), Fu et al. (2014), and Zhu et al. (2016a). Video synopsis results are viewed by randomly chosen users to compare the visual pleasure of the results.

Some of the public datasets used in the aforementioned studies are KTH, WEIZMAN, PETS 2009, LABV (Mahapatra et al., 2016), CAVIAR (Xu et al., 2015), Hall monitor, Day-time, F-building (Zhong et al., 2014; Wang et al., 2013b,a). Still, most of the studies evaluated their method on local datasets which are not publicly available.

There is no standard baseline for commonly used datasets or performance metrics, thus a performance comparison cannot be carried out directly on the measurements of each study. An experimental comparison study for all methods on the same domain to evaluate their performance would be a significant contribution.

6. Conclusions

A comprehensive review of video synopsis methods was presented in this paper. The reviewed studies cover all the literature on video synopsis starting from the first publications onwards. The current situation from different aspects such as on-line/off-line optimization, multi-camera/single-camera, compressed/pixel domain, and activity clustering were investigated, and the pros and cons of these aspects were examined in detail. Potential improvements and suggestions according to the performed analysis were mentioned, and statistics on publications regarding mentioned aspects were also shared to determine the current trend and potential future paths of study.

Video synopsis studies generally focus on the optimization step. Nevertheless, all of the other video synopsis methodology steps were analyzed in this study. These included object detection, object tracking, background generation and stitching, which also have a part to play in improving the results of video synopsis. Despite most of the studies considering these steps as pre/post-processing operations, recent methods showing better performance in each step can be adopted for video synopsis. Especially, object detection and tracking methods which are applied before the optimization have direct effect on the quality of video synopsis. A literature search on object detection showed that there are more precise methods (Goyette et al., 2012), and the adaptation of these methods into video synopsis will directly improve their quality. Even so, false detection directly affects the results of a video synopsis on both visual quality and computational complexity. Therefore, more attention must be paid on this step to obtain the best results for improvement of video synopsis. Similarly, object tracking is a very popular area in video processing, and research in this area progresses significantly day by day. Novel methods are being proposed that are especially robust against environmental difficulties, which are important problems in video synopsis (Kristan et al., 2015). Therefore, adaptation of the latest object tracking methods for video synopsis will also increase precision significantly.

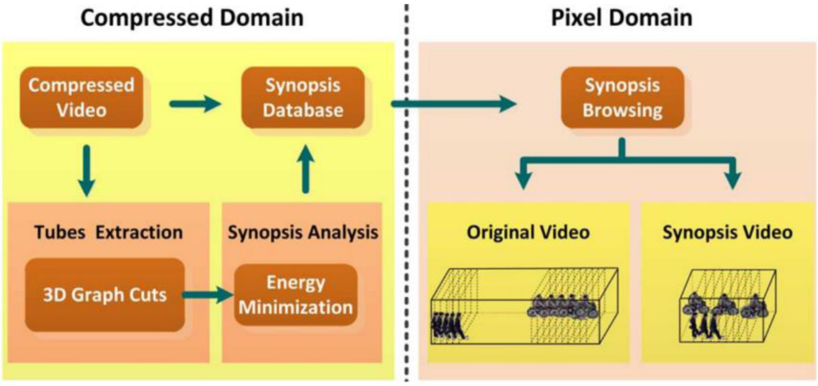


Fig. 9. Video synopsis framework in the compressed domain proposed by Zhong et al. (2014).

There is a tradeoff between condensation ratio and runtime performance of video synopsis. While off-line approaches show better precision, on-line approaches are more applicable to real-world applications when considering computational complexity. Off-line optimization methods are more efficient, especially for scenes with dense activity, as they provide more advanced rearrangement. On-line methods seem more proper for scenes with low density and simpler activities, especially when real-time application is desired. Even though the majority of approaches utilize a single-camera view, there is incremental interest in multi-camera approaches which provide efficient solutions with a wider perspective. Therefore, the trend in the literature shows that multi-camera approaches benefiting from the advantages of both on-line and off-line methods seem to be the main focus of future studies.

GPU-based methods for acceleration of off-line optimization seem to be a hot spot considering recent technological development in this area. Run-time performance, which is a major shortcoming of off-line optimization, can be solved by GPU-based implementation. Furthermore, the idea of distributed video synopsis frameworks that perform each step of the methodology on a different node seems promising considering its integration with multi-camera topology and current development of cloud technology. In this way, technical improvement could support the direct application of video synopsis on video surveillance systems.

In this study, we satisfied the need for a review of video synopsis methods. A systematic analysis of the current literature is provided with the aim of explaining how video synopsis methods work via a detailed analysis of the methodology, variations in the studies on different domains, and the advantages and bottlenecks specific to each step. Potential paths to overcome the bottlenecks are also discussed to guide future studies.

References

Barnich, O., Van Droogenbroeck, M., 2011. Vibe: A universal back-ground subtraction algorithm for video sequences. IEEE Trans. Image Process. 20, 1709–1724.

Baskurt, K.B., Samet, R., 2017. Improved adaptive background subtraction method using pixel-based segmenter. In: 2017 Computer Graphics, Visualization and Computer Vision (WSCG), 25th International Conference in Central Europe. WSCG, pp. 41–46.

Baskurt, K.B., Samet, R., 2018. Long-term multiobject tracking using alternative correlation filters. Turkish J. Electr. Eng. Comput. Sci. 26 (5).

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1222–1239.

Chakraborty, S., Tickoo, O., Iyer, R., 2015. Adaptive key frame selection for video summarization. In: 2015 IEEE Winter Conference on Applications of Computer Vision, WACV. IEEE, Piscataway, NJ, pp. 702–709.

Chen, Y.T., Chen, C.S., Huang, C.R., Hung, Y.P., 2007. Efficient hierarchical method for background subtraction. Pattern Recognit. 40, 2706–2715.

Choeychuen, K., Kumhom, P., Chamnongthai, K., 2006. An efficient implementation of the nearest neighbor based visual objects tracking. In: 2006 International Symposium on Intelligent Signal Processing and Communications. ISPACS'06. IEEE, Piscataway, NJ.

Chou, C.L., Lin, C.H., Chiang, T.H., Chen, H.T., Lee, S.Y., 2015. Coherent event-based surveillance video synopsis using trajectory clustering. In: 2015 IEEE International Conference on Multimedia & Expo Workshops. ICMEW. IEEE, Piscataway, NJ, pp. 1–6.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. IEEE, Piscataway, NJ, pp. 886–893.

Feng, S., Lei, Z., Yi, D., Li, S.Z., 2012. Online content-aware video condensation, in. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. CVPR. IEEE, Piscataway, NJ, pp. 2082–2087.

Feng, S., Liao, S., Yuan, Z., Li, S.Z., 2010. Online principal background selection for video synopsis. In: 20th International Conference on Pattern Recognition. ICPR. IEEE, Piscataway, NJ, pp. 17–20.

Finkel, R.A., Bentley, J.L., 1974. Quad trees a data structure for retrieval on composite keys. Acta Inform. 4, 1–9.

Fu, W., Wang, J., Gui, L., Lu, H., Ma, S., 2014. Online video synopsis of structured motion. Neurocomp 135, 155–162.

Girshick, R., 2015. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision. ICCV. IEEE, Piscataway, NJ, pp. 1440–1448.

Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P., 2012. Changedetection.net: A new change detection bench-mark dataset. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPRW. IEEE, Piscataway, NJ, pp. 1–8.

He, Y., Gao, C., Sang, N., Qu, Z., Han, J., 2017a. Graph coloring based surveillance video synopsis. Neurocomp 225, 64–79.

He, Y., Qu, Z., Gao, C., Sang, N., 2017b. Fast online video synopsis based on potential collision graph. IEEE Signal Process. Lett. 24, 22–26.

Hoshen, Y., Peleg, S., 2015. Live video synopsis for multiple cameras. In: 2015 IEEE International Conference on Image Processing. ICIP. IEEE, Piscataway, NJ, pp. 212–216.

Hsia, C.H., Chiang, J.S., Hsieh, C.F., 2016. Low-complexity range tree for video synopsis system. Multimedia Tools Appl. 75, 9885–9902.

Hsia, C.H., Wu, T.C., Chiang, J.S., Hsieh, C.F., 2011. VLSI Architecture design of moving objects detection using adaptive least-mean-square scheme. In: 2011 IEEE International Symposium on Intelligent Signal Processing and Communications Systems. ISPACS. IEEE, Piscataway, NJ, pp. 1–6.

Huang, C.R., Chung, P.C.J., Yang, D.K., Chen, H.C., Huang, G.J., 2014. Maximum a posteriori probability estimation for online surveillance video synopsis. IEEE Trans. Circuits Syst. Video Technol. 24, 1417–1429.

Jin, J., Liu, F., Gan, Z., Cui, Z., 2016. . online video synopsis method through simple tube projection strategy. In: 8th International Conference on Wireless Communications & Signal Processing. WCSP. IEEE, Piscataway, NJ, pp. 1–5.

Kaewtrakulpong, P., Bowden, R., 2002. An improved adaptive back-ground mixture model for real-time tracking with shadow detection. In: Remagnino, P., Jones, G.A., Paragios, N., Regazzoni, C.S. (Eds.), Video-Based Surveillance Systems. Springer, Boston, MA, pp. 135–144.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82, 35–45.

Kasamwattanarote, S., Cooharojananone, N., Satoh, S., Lipikorn, R., 2010. Real time tunnel based video summarization using direct shift collision detection. In: Advances in Multimedia Information Processing-PCM 2010. Springer, Berlin, Heidelberg, pp. 136–147.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Kolmogorov, V., Zabin, R., 2004. What energy functions can be minimized via graph cuts?. IEEE Trans. Pattern Anal. Mach. Intell. 26, 147–159.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., Vojir, T., Hager, G., Nebehay, G., Pflugfelder, R., 2015. The visual object tracking vot2015 challenge results. In: 2015 IEEE International Conference on Computer Vision Workshop. ICCVW. IEEE, Piscataway, NJ, pp. 1–23.

Li, X., Wang, Z., Lu, X., 2016. Surveillance video synopsis via scaling down objects. IEEE Trans. Image Process. 25, 740–755.

Liao, W., Tu, Z., Wang, S., Li, Y., Zhong, R., Zhong, H., 2017. 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, pp. 253–261.

Liao, S., Zhao, G., Kellokumpu, V., Pietikainen, M., Li, S.Z., 2010. Modeling pixel process with scale invariant local patterns for background sub-traction in complex scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition. CVPR. IEEE, Piscataway, NJ, pp. 1301–1306.

Lin, L., Lin, W., Xiao, W., Huang, S., 2017. An optimized video synopsis algorithm and its distributed processing model. Soft Comput. 21, 935–947.

Lin, W., Zhang, Y., Lu, J., Zhou, B., Wang, J., Zhou, Y., 2015. Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. Neurocomp 155, 84–98.

Lu, M., Wang, Y., Pan, G., 2013. Generating fluent tubes in video synopsis. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP. IEEE, Piscataway, NJ, pp. 2292–2296.

Mahapatra, A., Mishra, T.K., Sa, P.K., Majhi, B., 2014. Human recognition system for outdoor videos using hidden Markov model. AEU Int. J. Electron. Commun. 68, 227–236.

Mahapatra, A., Sa, P.K., Majhi, B., Padhy, S., 2016. MVS: A multi-view video synopsis framework. Signal Process. Image Commun. 42, 31–44.

Mitchell, M., 1998. An Introduction to Genetic Algorithms. MIT press.

Nie, Y., Xiao, C., Sun, H., Li, P., 2013. Compact video synopsis via global spatiotemporal optimization. IEEE Trans. Vis. Comput. Graph. 19, 1664–1676.

Papenberg, N., Bruhn, A., Brox, T., Didas, S., Weickert, J., 2006. Highly accurate optic flow computation with theoretically justified warping. Int. J. Comput. Vis. 67, 141–158.

Perez, P., Hue, C., Vermaak, J., Gangnet, M., 2002. Color-based probabilistic tracking. In: Proceedings of the 7th European Conference on Computer Vision. ECCV 2002. Springer-Verlag, London, UK, pp. 661–675.

Pritch, Y., Ratovitch, S., Hendel, A., Peleg, S., 2009. Clustered synopsis of surveillance video. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveil-Lance, 2009. AVSS'09. IEEE, Piscataway, NJ, pp. 195–200.

Pritch, Y., Rav-Acha, A., Gutman, A., Peleg, S., 2007. Webcam synopsis: Peeking around the world. In: 11th IEEE Interna-Tional Conference on Computer Vision, 2007. ICCV 2007. IEEE, Piscataway, NJ, pp. 1–8.

Pritch, Y., Rav-Acha, A., Peleg, S., 2008. Nonchronological video synopsis and indexing. IEEE Trans. Pattern Anal. Mach. In-tell 30, 1971–1984.

Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y., 2008. Simplemkl. J. Mach. Learn. Res. 9, 2491–2521.

Rav-Acha, A., Pritch, Y., Peleg, S., 2006. Making a long video short: Dynamic video synopsis. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR. IEEE, Piscataway, NJ, pp. 435–441.

Rodriguez, M., 2010. CRAM: Compact representation of actions in movies. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition. CVPR. IEEE, Piscataway, NJ, pp. 3328–3335.

Rodriguez, M.D., Ahmed, J., Shah, M., 2008. Action MACH: A spatiotemporal maximum average correlation height filter for action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008. IEEE, Piscataway, NJ, pp. 1–8.

Smith, M.A., Kanade, T., 1998. Video skimming and characterization through the combination of image and language under-standing. In: 1998 IEEE International Workshop on Content-Based Access of Image and Video Database. IEEE, Piscataway, NJ, pp. 61–70.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. In: 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway, NJ, pp. 246–252.

Stauffer, C., Grimson, W.E.L., 2000. Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. 22, 747–757.

Sun, J., Zhang, W., Tang, X., Shum, H.Y., 2006. Background cut. In: European Conference on Computer Vision. ECCV 2006. Springer, Berlin, Heidelberg, pp. 628–641.

Taj, M., Maggio, E., Cavallaro, A., 2006. Multi-feature graph-based object tracking. In: International Evaluation Workshop on Classification of Events, Activities and Relationships. Springer, Berlin, Heidelberg, pp. 190–199.

Tian, Y., Zheng, H., Chen, Q., Wang, D., Lin, R., 2016. Surveillance video synopsis generation method via keeping important relationship among objects. IET Comput. Vis. 10, 868–872.

Truong, B.T., Venkatesh, S., 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Comm. Appl. 3, http://dx.doi.org/10.1145/1198302.1198305.

Vlachos, M., Kollios, G., Gunopulos, D., 2002. Discovering similar multidimensional trajectories. In: 18th International Conference on Data Engineering. IEEE, Piscataway, NJ, pp. 673–684.

Vural, U., Akgul, Y.S., 2009. Eye-gaze based real-time surveillance video synopsis. Pattern Recognit. Lett. 30, 1151–1159.

Wang, S.z., Wang, Z.y., Hu, R.m., 2013a. Surveillance video synopsis in the compressed domain for fast video browsing. J. Vis. Commun. Image Represent. 24, 1431–1442.

Wang, S., Xu, W., Wang, C., Wang, B., 2013b. A framework for surveillance video fast browsing based on object flags. In: The Era of Interactive Media. Springer, Berlin, Heidelberg, pp. 411–421.

Whitley, D., 1994. A genetic algorithm tutorial. Stat. Comput. 4, 65–85.

Xu, M., Li, S.Z., Li, B., Yuan, X.T., Xiang, S.M., 2008. A set theoretical method for video synopsis. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. ACM, New York, NY, pp. 366–370.

Xu, L., Liu, H., Yan, X., Liao, S., Zhang, X., 2015. Optimization method for trajectory combination in surveillance video synopsis based on genetic algorithm. J. Ambient Intell. Humaniz. Comput. 6, 623–633.

Yang, T., Pan, Q., Li, J., Li, S.Z., 2005. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2005. IEEE, Piscataway, NJ, pp. 970–975.

Yao, T., Xiao, M., Ma, C., Shen, C., Li, P., 2014. Object based video synopsis. In: 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications. WARTIA. IEEE, Piscataway, NJ, pp. 1138–1141.

Yildiz, A., Ozgur, A., Akgul, Y.S., 2008. Fast non-linear video synopsis. In: 23rd International Symposium on Computer and Information Sciences, 2008. ISCIS'08. IEEE, Piscataway, NJ, pp. 1–6.

Zhang, P., Wang, L., Huang, W., Xie, L., Chen, G., 2015. Multiple pedestrian tracking based on couple-states Markov chain with semantic topic learning for video surveillance. Soft Comput. 19, 85–97.

Zhong, R., Hu, R., Wang, Z., Wang, S., 2014. Fast synopsis for moving objects using compressed video. IEEE Signal Process. Lett. 21, 834–838.

Zhu, X., Change Loy, C., Gong, S., 2013. Video synopsis by heterogeneous multi-source correlation. In: 2013 IEEE Inter-National Conference on Computer Vision. IEEE, Piscataway, NJ, pp. 81–88.

Zhu, J., Liao, S., Li, S.Z., 2016a. Multicamera joint video synopsis. IEEE Trans. Circuits Syst. Video Technol. 26, 1058–1069.

Zhu, X., Liu, J., Wang, J., Lu, H., 2014. Key observation selec-tion-based effective video synopsis for camera network. Mach. Vis. Appl. 25, 145–157.

Zhu, X., Loy, C.C., Gong, S., 2016b. Learning from multiple sources for video summarisation. Int. J. Comput. Vis. 117, 247–268.