Pattern Recognition 109 (2021) 107567

Contents lists available at ScienceDirect

Pattern Recognition

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

A comprehensive survey of multi-view video summarization

Tanveer Hussain[a], Khan Muhammad[b], Weiping Ding[c], Jaime Lloret[d], Sung Wook Baik[a,*], Victor Hugo C. de Albuquerque[e]

[a] Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul143-747, South Korea
[b] Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, Republic of Korea
[c] School of Information Science and Technology, Nantong University, Nantong 226019, China
[d] Politechnic University of Valencia, Valencia, Spain
[e] Laboratory of Bioinformatics, University of Fortaleza, Fortaleza, Brazil

ARTICLE INFO

Article history:
Received 22 August 2019
Revised 7 March 2020
Accepted 28 July 2020
Available online 29 July 2020

Index Terms:
Computer vision
Multi-view video summarization
Multi-sensor management
Multi-camera networks
Machine learning
Features fusion
Big data
Video summarization survey

ABSTRACT

There has been an exponential growth in the amount of visual data on a daily basis acquired from single or multi-view surveillance camera networks. This massive amount of data requires efficient mechanisms such as video summarization to ensure that only significant data are reported and the redundancy is reduced. Multi-view video summarization (MVS) is a less redundant and more concise way of providing information from the video content of all the cameras in the form of either keyframes or video segments. This paper presents an overview of the existing strategies proposed for MVS, including their advantages and drawbacks. Our survey covers the genericsteps in MVS, such as the pre-processing of video data, feature extraction, and post-processing followed by summary generation. We also describe the datasets that are available for the evaluation of MVS. Finally, we examine the major current issues related to MVS and put forward the recommendations for future research[1] .

1. Introduction

Due to the recent increase in amount of video data from surveillance cameras, it has become very challenging to process these data for various applications such as video browsing and retrieval [1], object segmentation [2], semantic/action recognition [3], and background subtraction [4]. The manual extraction of informative sections from video data and their processing are laborious tasks, and there is therefore a need for automatic techniques to remove redundancy and extract useful information. To solve these challenges, various techniques such as video skimming [5], video summarization [6], and video condensation [7] have been presented. A video skim is a short segment of the original video that reflects its overall representation. Video summarization is a technique used to extract salient frames or sequences of frames from a video; it offers fast browsing by shortening the input video into a synopsis and retaining only salient information. Video summarization is a hot area of research due to the massive growth in the amounts of video data captured by surveillance cameras or recorded via smartphones on a daily basis. It can be broadly divided into single-view video summarization (SVS) and MVS. SVS [8] is the process of creating a summary of single-view video and the generated summary should preserve three properties including minimum repetition, representativeness, and diversity [9]. Most of the summarization techniques are presented for single-view videos because their target is to produce a summary which is representative of the input video by considering only the intra-view correlations. SVS is less challenging compared to MVS due to lack of synchronization and illumination difference problem among different views. This article focuses on only MVS, thus SVS is outside the scope of this paper.

MVS is a rarely addressed problem in the literature of video summarization. As for SVS, the output generated from MVS is either a set of representative frames (keyframes), a short and comprehensive video (video synopsis) or video skims. Unlike SVS, however, the input for MVS is acquired from various cameras with different views. The basic pipeline for handling such videos include pre-processing, feature extraction, post-processing, and summary generation. The preprocessing of multi-view videos (MVVs) typically comprises redundancy removal steps such as segmentation [10] and shot boundary detection [11]. Pre-processing is followed by features extraction, object detection or tracking, the choice of

* Corresponding author.
E-mail addresses: tanveerkhattak3797@gmail.com (T. Hussain), khan.muhammad@ieee.org (K. Muhammad), sbaik@sejong.ac.kr (S.W. Baik).

[1] https://github.com/tanveer-hussain/MVS-Survey

https://doi.org/10.1016/j.patcog.2020.107567
0031-3203/

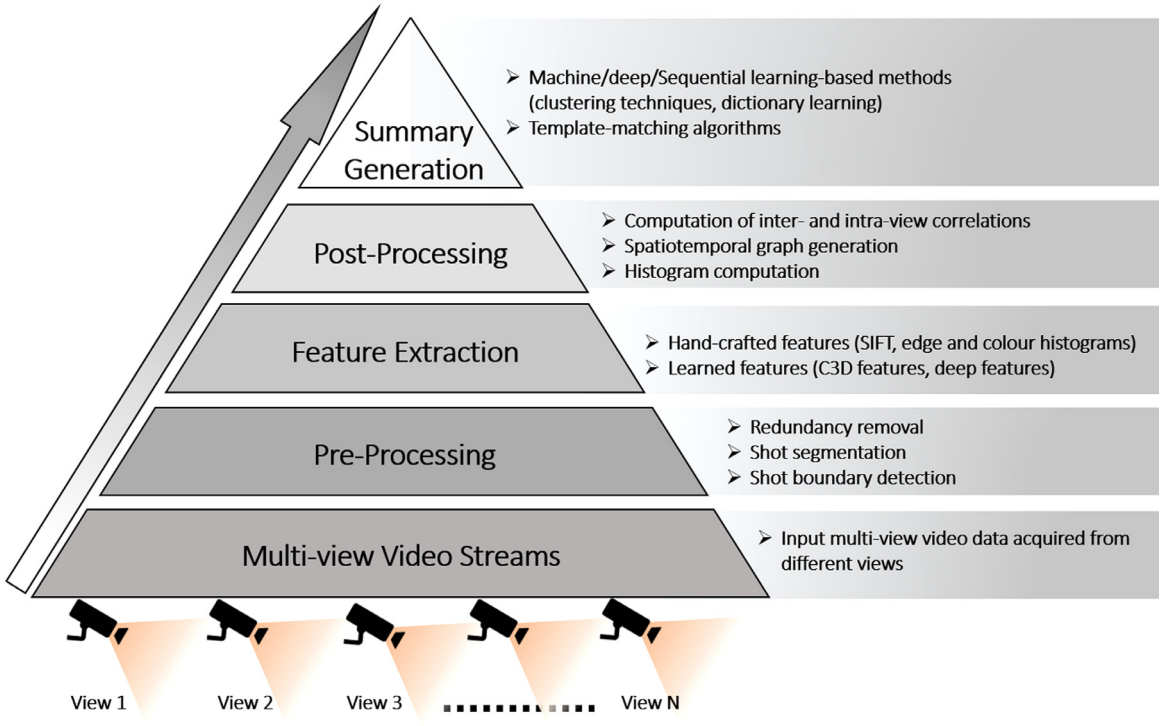

Fig. 1. General flow of MVS methods (condensation of multi-view videos into a short and comprehensive output summary).

which varies from one application to another. Feature extraction methods include handcrafted features [10,11] such as, colour histograms, edge histograms, colour layout descriptors, and learned features through CNNs, as presented in [12]. Post-processing of the extracted features refers to computation of intra- and inter-view correlations, as used in mainstream existing methods, followed by summary generation. The most common approaches for summary generation in MVS literature are based on machine learning techniques or template matching. The final generated output varies, and may consist of keyframes, video skim or video synopsis. The overall flow of MVS methods is shown in Fig. 1. MVS can be used in several applications including surveillance (both indoor and outdoor) for activity and event analysis [13], investigating accident scenarios [14], interesting and salient events extraction from sports, security and law enforcement for theft detection, and robbery events recognition. It also has applications in virtual reality in terms of creating a single 360 view using multi-view cameras.

There are several significant challenges that are encountered when summarizing MVVs, such as inter and intra-view correlations among different views, the problem of synchronisation, the presence of different lighting conditions for different views, and the possible overlapping of views. A simple multi-view network contains two cameras with two output videos. For example, consider two cameras acquiring video from two different views at 25 fps, which generates 180,000 frames (90000 for each camera) for one hour of video content. This huge amount of data is the big hurdle to the exploitation of important visual data and extraction of more interesting and specific aspects of videos is essential, rather than watching the whole video. The literature shows that several companies are already working on MVVs for different purposes such as car parking and the provision of supportive, comfortable, and safe driving. For instance, Honda Motor [15] has launched a multi-view camera system that offers different views with multiple wide angles from cameras within the vehicle. Similarly, many other companies are developing multi-view camera systems for different applications , such as the scheme in [16], which can record video from multiple views in excessive temperature , providing assistance for further analysis of the scene. Similarly, based on a multi-view camera network, a software developed in [17] can monitor and control users working inside a company and can analyse the ongoing processes within the company.

In the modern technological era, cameras are installed in offices, campuses, factories, streets, and public places. These cameras capture continuous data on daily basis that can be used for many purposes, such as persons tracking [18], disaster management [19], abnormal event detection [20], and other video analytics applications [21]. To date, the field of MVS has been less exploited and there is an urgent need for MVS techniques to process video data effectively for various applications. Mainstream current research works use traditional hand-engineered features for MVS, while deep learning has recently been applied to numerous computer vision applications such as disaster management [19], security [22], and abnormal activity recognition [23]. It is therefore recommended to develop deep learning-assisted intelligent methods for MVS. Likewise, in literature, there is a deficiency of standard publicly available datasets that are challenging and can be used for better evaluation of MVS techniques. The major contributions of our survey are summarized as follows:

1. We present the very first survey of MVS methods. To the best of our knowledge, there is no existing survey in the MVS literature and it lacks the attention of researchers. With this motivation, we present a comprehensive and compact tutorial of all the existing MVS methods.
2. In this survey, we cover trends in the MVS literature, the distribution on the basis of publishers, citations, types of research papers, and application-wise scattering of MVS approaches along with taxonomy of MVS methods. Further, this survey provides results of all the queries for searching MVS papers in various repositories and also offer remarks about the selection process of the retrieved papers in the survey.
3. This survey explores the current challenges of MVS methods, explores the evaluation metrics, datasets, and draw conclusions of the overall literature. Finally, our survey provides recommen-

dations and future research directions for further exploration of MVS field.

Rest of the paper is structurally divided into eight different sections. Section II covers the scope of this survey, provides its outline and coverage. Section III investigates the existing methods for summarization of MVV and provides categorical distribution and taxonomy of different MVS methods. The available MVV datasets with their characteristics are discussed in Section IV. Section V explores the evaluation metrics of different methods used in MVS literature. Section VI and VII highlight the major challenges of MVS and provides recommendations for future research. Section VIII concludes this survey with discussion about future research methodologies.

2. Scope, outline, and coverage of this survey

This study covers workshops, journals, and conference papers on MVS methods from diverse repositories including Google Scholar, ScienceDirect, IEEE Xplore, ACM, and Springer. We searched for related articles using different queries in all of these repositories and some of the retrieved items are excluded because of irrelevancy. The overall items explored with different searches made in several repositories are given in Table 1, together with reasons for not considering some papers in this survey. Table 2 shows the application-wise distribution of MVS methods. Distribution of MVS literature such as publication year, a number of articles per year in literature along with the citation of each paper is shown in Fig. 2 (a and b), while the distribution of MVS papers by publisher is shown in Fig. 2 (c). A scatter chart of MVS conference papers, journals, and other types of paper is visualized in Fig. 2 (d). In MVS literature, most of the papers are published in IEEE journals and top conferences. The year-wise trend of MVS is shown in Fig. 3, covering the overall literature. Initial research to MVS used low-level features (SIFT descriptors) and object detection based on handcrafted features with activity-based video segmentation. These early articles utilised clustering techniques (i.e., K-means clustering) for final summary generation. Next trend in literature also focus on low-level features with some improvements such as background subtraction for trajectories extraction and usage of machine learning techniques such as support vector machine (SVM) and K-means clustering for final summary generation. The final summaries of these works are of uniform length or user query based. There is a positive variation in the next MVS trend that used mid-level (i.e., motion and saliency) features along

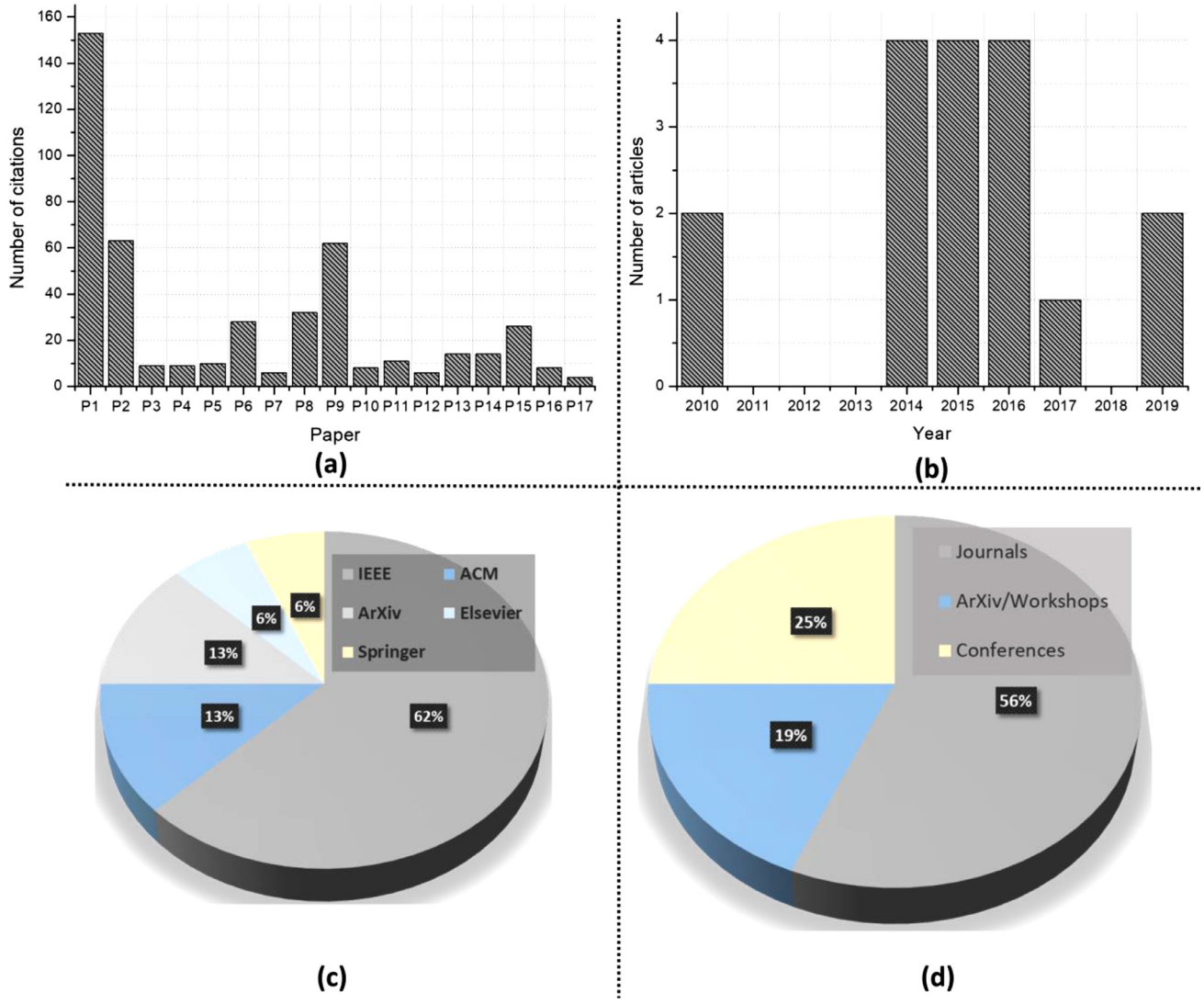

Fig. 2. Overall distribution of MVS literature (a) citation wise distribution of MVS research papers P1[26], P2[10], P3[33], P4[31], P5[50], P6[34], P7[37], P8[11], P9[35], P10[27], P11[43], P12[41], P13[40], P14[39], P15[45], P16[48], P17[49], (b) year-wise MVS publications to date, (c) publisher-wise distribution of MVS research paper, and (d) distribution of MVS on the basis of research paper's type.

Table 1
Comprehensive details of researched MVS literature distributed through various publishers.

No	Method	Year	Included	Remarks
1	[24]	2005	✖	Multi-view keyword in this paper is referred as human faces and video captioning problem individually based on low-level visual features. The authors considered human faces (one view), video captioning (second view) to generate summary of a single-view video.
2	[25]	2008	✖	This paper focuses on different events happening in an indoor office multi-camera setup. Optimal view selection in this article is followed by event sequence summarization. Finally, a fuzzy rule-based system is used to estimate the human decision making. No discussion about MVS and it's relevant steps.
3	[10]	2010	✔	The authors proposed an MVS framework using random walks, where hypergraphs are used to capture correlations among different views.
4	[26]	2010	✔	In this work, maximal marginal relevance (a concept of text summarization) is intelligently utilised to generate MVS.
5	[27]	2014	✔	The concepts of maximum-margin clustering [28] and disagreement minimisation criterion [29] is integrated together with metric learning for MVS.
6	[30]	2014	✖	This is a proposal paper, methodology is not explained, and the authors mentioned to present the actual MVS related work in future. No evaluation performed, and no standard datasets followed for experimentation.
7	[31]	2014	✔	Inspired by the previous work in [26], the authors proposed an online MVS system by integrating maximal marginal relevance with a bandwidth efficient distributed algorithm.
8	[32]	2014	✖	This method generated an SVS based on convex mixture models and spectral clustering.
9	[33]	2014	✔	In this framework, authors utilized user's view tendency for the selection of viewpoint for MVV contents. This framework showed good results for sport events and live concerts.
10	[34]	2014	✔	Motion regions based video segmentation is followed by activities recognition. Inter-inter-activity redundancy removal step is finally concluded by recognising anomalous patterns by finding rarely occurred activities.
11	[35]	2015	✔	The main theme is to reduce compression and transmission power. This scheme comprises online and offline modules for MVS.
12	[11]	2015	✔	This technique used semantic features in the form of a visual bag of words. Gaussian entropy, bipartite graph matching, and optimum-path forest algorithm are used for summary generation.
13	[36]	2015	✖	This method is focused on making decision about persons on the basis of their past activities in live surveillance video and it does not generate multi-view summary. Further it creates a separate synopsis for each view.
14	[37]	2015	✔	In this article, authors produced video synopsis from multi-view videos based on human actions in both indoor and outdoor scenarios.
15	[38]	2016	✖	This method is focused only on multi-video summarization, MVS is not covered in this article.
16	[39]	2016	✔	In this paper, authors presented a novel technique based on joint embedding and sparse coding for summarization of multi-view videos.
17	[40]	2016	✔	A multi-camera joint video synopsis is presented that finds object's appearance, merging, splitting, and disappearing moments in the frame sequence called as tube from each view. These tubes are joined by rearranging them so that temporal ordering remains same for all the cameras. The final multi-camera synopsis is created by stitching together the rearranged tubes and background images from the same camera.
18	[41]	2016	✖	This is the journal extension of [37] with similar contents.
19	[42]	2016	✖	It is the same work as presented in an archive paper [27], but published in a conference proceeding.
20	[43]	2016	✔	The authors proposed framework which makes sparse coding feasible in summarizing both single and multi-view videos by exploiting both intra- and inter-view content correlations.
21	[44]	2017	✖	This paper proposed an algorithm to generate summary of multi-video, they have not focused on MVS.
22	[45]	2017	✔	Authors captured multi-view correlations via embedding which helps for extracting diverse set of representation and used L1, and L2 sparse optimizations for selecting representative shots for the summary
23	[46]	2017	✖	[47] is an extended version of [46], and both techniques are focused only on SVS, the datasets used for evaluation are SVS datasets.
24	[47]	2018	✖	
26	[48]	2020	✔	This paper presents shot segmentation after object detection that is advanced to deep features extraction for MVS using multi-layer and bi-directional LSTM.
27	[49]	2020	✔	The research work in this paper performs targets detection and analysis, followed by transmission of annotated frames in encoded form towards the master resource-constrained device for information computation, where the final summary is generated on the basis of maximum information.

Table 2
Application wise distribution of MVS methods with specified output format and final objectives of each method.

Application	Method	Expected outputs			Objectives		
		Keyframes	Video skims	Video synopsis	Diversity	Action	Events
Law enforcement	[37]	-	-	✔	-	-	✔
	[35]	✔	-	-	✔	-	-
	[41]	-	-	✔	-	✔	✔
	[49]	✔	-	-	-	✔	-
Sports	[33]	-	✔	-	✔	-	-
	[52]	-	-	-	✔	-	-
Surveillance	[43]	✔	-	-	✔	-	-
	[40]	-	-	-	-	-	✔
	[45]	-	-	✔	✔	-	✔
	[50]	-	-	✔	✔	-	-
	[10]	-	✔	-	-	-	✔
	[11]	-	✔	-	-	-	✔
	[35]	✔	-	-	✔	-	-
	[27]	✔	-	-	✔	-	-
	[39]	✔	-	-	-	-	✔
	[31]	✔	-	-	✔	-	-
	[34]	✔	-	-	-	-	✔
	[48]	-	✔	-	-	-	✔

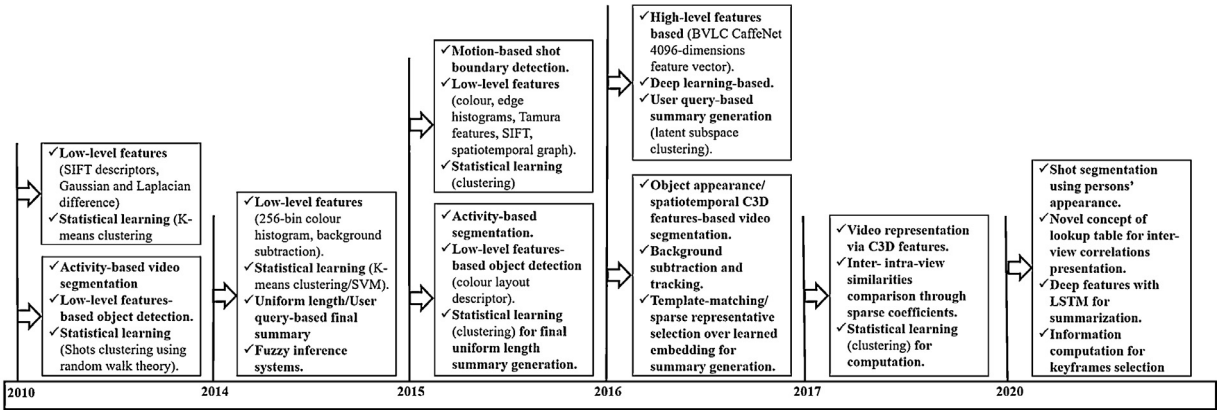

Fig. 3. Trend of MVS methods (year wise features distribution and learning mechanisms of representative articles from the literature).

with handcrafted features (i.e., colour-, edge-histograms, Tamura, and SIFT features). Similar to the previous trend, the summary of these methods is generated using clustering techniques. A breakthrough in MVS field is noticed after the usage of learned features and the generation of summaries by utilising deep features in the prerequisite steps. This trend is followed in 2016, where BVLC CaffeNet 4096-dim and spatio temporal C3D [51] features were used for sparse coding and video representation, respectively, for the first time in literature. Besides clustering for summary generation, template-matching and sparse representative selection over learned embedding are also used for final summary generation. A research in this trend segments video based on objects, followed by human action recognition using multiple kernel learning advanced to synopsis generation using fuzzy inference systems. Likewise, an article published in 2017 used C3D features for video representation. In this trend, inter- and intra-view similarities were computed via sparse coefficients and final summary is generated using a clustering scheme. A completely new approach is presented by Hussain et al. [48], performing shots segmentation using person's appearance. Inter-view correlations are computed using a novel concept of lookup table followed by deep features extraction that are fed into a multi-layer LSTM for final skims generation. More recently, we utilised resource-constrained devices for MVS [49], where these devices are equipped with cameras and are interconnected in internet of things (IoT) network. In this method, the captured video in real-time is processed to detect targets and the annotated frames with targets above acertain threshold are encoded and transferred to master device for information computation. The keyframes were selected based on the amount of information among the received frames.

3. Methods for summarization of multi-view videos

In literature, various techniques are presented for MVS and each method follows a generic process with some specific steps. The basic flow of MVS contains three steps: segmentation of MVVs, features extraction, and summary generation. Processing whole video and generating summary at once is a biased decision, since a video contains different shots, scenes, and diverse information that are distributed throughout the video. Thus, the first step in the majority of the MVS methods is video segmentation into smaller constituents, as briefly covered in the next section. This step is followed by features extraction that assists in final summary generation. In the final step, , specific schemes or validation criteria for keyframes selection are applied. These steps are elaborated in the subsequent sections individually and are comprehensively shown in Tables 3 and 4.

3.1. Multi-view videos segmentation

The segmentation of videos into parts is a pre-processing step for summary generation. Multi-view summary generated subsequently after segmentation of videos into several parts makes it more representative of all videos. In literature, many techniques

Table 3
Description of used features, shot segmentation technique, learning mechanism, and summary generation of existing MVS methods.

Method name/year	Visual features used				Shot segmentation		Summary generation						
							Technique				Length of generated summary		
	Low-level features	Saliency and motion	High-level features	Object detection	Uniform length	Features-based	Statistical classifiers	Deep learning classifier	Clustering	Other (template matching/sparse representative selection)	Uniform length	Importance-based	User query-based
[10] (2010)	✔	-	✔	✔	-	✔	-	-	✔	-	-	✔	-
[26] (2010)	✔	-	-	-	-	-	-	-	✔	-	✔	-	-
[33] (2014)	-	-	-	-	-	-	✔	-	-	-	-	-	✔
[31] (2014)	✔	-	-	-	-	✔	-	-	✔	-	✔	-	-
[50] (2014)	-	-	-	✔	-	-	-	-	✔	-	-	✔	-
[27] (2014)	✔	-	-	-	-	-	✔	-	-	-	✔	-	-
[34] (2014)	-	✔	-	-	-	✔	✔	-	-	-	-	✔	-
[37] (2015)	-	-	-	✔	-	-	✔	-	-	-	✔	-	-
[11] (2015)	✔	✔	-	-	-	✔	-	-	✔	-	-	✔	-
[35] (2015)	✔	-	-	✔	✔	-	-	-	✔	-	-	✔	-
[43] (2016)	-	-	✔	-	-	-	-	-	✔	-	-	-	✔
[40] (2016)	-	✔	-	-	-	-	-	-	-	✔	✔	-	-
[41] (2016)	✔	-	-	-	-	✔	-	-	-	✔	-	✔	-
[39] (2016)	✔	-	-	-	-	-	-	-	-	✔	-	✔	-
[45] (2017)	-	-	✔	-	-	✔	-	-	✔	-	-	✔	-
[48] (2020)	-	-	✔	-	-	✔	-	✔	-	-	-	✔	-
[49] (2020)	✔	-	✔	-	-	✔	-	-	-	✔	-	✔	-

Table 4
Detailed description of three main steps followed by each MVS method in literature.

Method	Segmentation	Feature extraction /object detection or tracking	Summary generation
[26] (2010)	-	Based on difference of Gaussian and Laplacian Gaussian and computation of SIFT descriptor, adopted from [53]	Clustering of SIFT descriptors into 500 groups by K-means to create a visual vocabulary for final MVS
[10] (2010)	Adopted activity-based video segmentation	Gaussian entropy fusion model, wavelet coefficients. Viola-Jones face detector is used to construct spatio-temporal shot graph	Shots clustering using random walk theory
[33] (2014)	-	Quality-Of-View is calculated from the distance between a camera and each object and the angle between them	SVM with RBF (Radial Basis Function) kernal function is used as the learning model
[31] (2014)	-	256-bin colour histogram in HSV colour space	K-means clustering for keyframes selection
[50] (2014)	Object segmentation and tracking	Background subtraction for trajectories extraction.	Key observation-based synopsis approach and K-means clustering for selection of pre-defined number of key actions
[34] (2014)	Motion regions based spatio-temporal cubes for video segmentation	Probabilistic latent component analysis (PLCA) to discover latent activities.	Inter- inter-activity redundancy removal, anomalous patterns recognition by finding rarely occurred activities.
[37] (2015)	-	Human detection using fuzzy inference system and then six different shape features are extracted from silhouette	SVM (Gaussian and Polynomial kernels are used in multiple kernel learning for classifying seven different actions)
[11] (2015)	Motion-based shot boundary detection	Colour histograms, edge histograms, Tamura features, and SIFT spatiotemporal graph	Unsupervised Optimum-Path Forest (clustering)
[35] (2015)	-	MPEG-7 colour layout descriptor, and score estimation of foreground object	Gaussian mixture model (clustering)
[27] (2015)	-	Low-level features	RBF kernel function for similarity measurement, maximal-margin clustering for summary generation.
[43] (2016)	-	BVLC CaffeNet (4096-dim CNN feature vector), sparse coding	Latent subspace clustering
[41] (2016)	Object detection and localisation	Action recognition using multiple kernel learning, as applied in [54]	Synopsis generation using fuzzy inference system applied to the tracked objects
[40] (2016)	Object appearance-based segmentation	Adaptive Scale Invariant Local Ternary Pattern (SILTP) for background subtraction and tacking object	Based on template-matching algorithm
[39] (2016)	Video segmentation via Spatio-temporal C3D [51] features	Two proximity matrices for inter- and intra-view correlations, pairwise Euclidean distances between frames	Sparse representative selection method over the learned embedding for summary generation.
[45] (2017)	Video representation via spatio-temporal C3D [51] features	Inter- and intra-view similarities are computed via sparse coefficients	Clustering for computation of data similarity
[48] (2020)	Person's appearance -ased shot segmentation	Deep features (AlexNet model) extraction from segmented shots in a lookup table	Multi-layer bi-directional LSTM for final summary generation.
[49] (2020)	Targets detection for shot segmentation	Low-level entropy features extraction	Information computation-based decision for keyframes selection

are based on segmentation of videos into shots, which are further processed for summary generation. In some techniques, shot segmentation is based on a uniform length [35] with a specific threshold for single shot selection, while in most of the techniques it is based on features [10,31] with non-uniform length. Segmentation of video into shots helps in removing redundancy such as suppressing the frames without persons [48], activity or motion and optimally choosing the frames with certain events or movements. Hence, the task of summarization can be achieved easily through features extraction and comparison after segmenting the shots that contain events, motion or saliency. In Table 3, third column shows the types of used shot segmentation techniques, i.e. whether shots are segmented with uniform length or they are features-based. The first column of Table 4 provides a detailed description about features or parameters used by each method for video segmentation.

Video segmentation of MVVs plays a vital role in the final summary generation, as the usage of the most representative segmented shots ensure a precise output summary. The available methods performing video segmentation can be divided into three major categories, based on motion/activity [10], visual features [39], and objects appearance [48]. The majority of the methods in MVS domain fall under the objects' appearance based category, where the video is segmented by considering shots having objects. Visual features are rarely used in the MVS field and C3D features are utilized to compute difference among different shots for video segmentation. Similarly, activity or motion [10] based video segmentation methods are also observable in MVS literature. The avaialable activity/motion-based methods have limited accuracy due to the adoption of weakly presented activity recognition algorithms, which can be extended in future with convincing accuracy after implementation of enhanced activity recognition and motion detection algorithms. High-level visual features-based methods and the objects appearance-based video segmentation algorithms result in properly segmented videos, which in turn yields representative summaries.

3.2. Feature extraction

The next prerequisite step for summary generation is features extraction which includes object detection or tracking [35,50]. The second column in Table 3 shows the features extracted by various methods in literature. Visual features are divided into four subcategories of low-level, mid-level (saliency and motion), high-level (deep) features, and object detection. Majority of the MVS methods are based on low-level features as shown in the third column of Table 4. Handcrafted features such as histogram (256 bin) features in HSV colour space [31], colour layout descriptors [35], entropy features [49], SIFT features [11,26], and other techniques such as background subtraction [50], foreground object estimation [35], and human detection [37] are utilized for summary generation. R. Panda et al. [43] used high-level features extracted from "BVLC CaffeNet" pre-trained model to find inter- and intra-view correlations in embedding space. Similarly, Hussain et al. [48] utilised learned

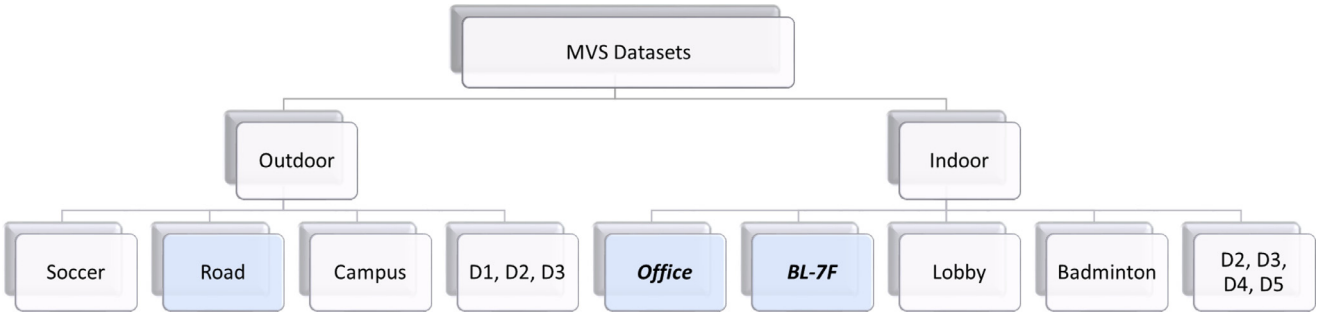

Fig. 4. Categorization of standard MVS datasets on the basis of recorded environment. Indoor recorded datasets include Office, BL-7F, Lobby, and Badminton. Outdoor datasets are Soccer, Road, Campus, and D1. D2 and D3 dataset include both indoor and outdoor scenarios. The datasets that are publicly available on the given links are filled with blue color and the ones with ground truth are made bold and italic.

features of a fully connected layer of an existing deep learning architecture, prior to final MVS.

The categorization of features extraction technologies given in Table 4 witnesses a variable set of features utilised as an intermediate recipient towards the final summary generation. The majority of the methods in MVS literature have considered statistical low-level features [49], while there are some techniques utilising high-level/learned features [48] or objects trajectories extraction [50] in the second step of MVS. The performance evaluation of MVS methods, as discussed in Section **V** notifies that the usage of learned features with effective prior shot segmentation techniques results in a representative summary. In contrast, the low-level features utilized in some techniques show comparatively convincing performance for final MVS, but the major techniques have poor output results.

3.3. Keyframes selection/summary generation

The final step involved in MVS pipeline is summary generation based on features extracted in the previous step. Final generated summary can be divided into further categories, such as uniform length summary generation [53], importance-based [10], and user query based [43] summaries, as represented in the last column of Table 3. Mainstream methods use statistical learning to generate summaries with different features and parameters. Clustering with random walks, K-means clustering, unsupervised optimum path, SVM, template-matching, and subspace clustering are widely used for summary generation. A description of summary generation methods is given in Table 4 along with references. A different approach for summary generation is presented in [48] which make use of sequential learning (LSTM) to choose subsequent frames (video skims) as a part of the final summary.

Among the present techniques for final summary generation, importance-based summary generation algorithms [48] perform well, when compared to statistical learning measures, template-matching, and many other similar techniques [53]. These methods generate a variable number of keyframes as final output on the basis of the information in the segmented shots, by using the features from the intermediate step of the general MVS pipeline [49].

4. Multi-view summarization datasets

A total of 11 multi-view datasets are available in literature; excluding action recognition datasets which are used for action- or event-based summary generation. The most popular MVS datasets are Office, Lobby, Campus [10], and BL-7F [35], that are covered in this section. The complete details and description of all the datasets are given in Table 5 and their distribution is presented in Fig. 4.

Table 5 contains the datasets information in sequential form including publication year, camera type, number of views, details about the indoor or outdoor environment, description about shots, a total number of videos, and their duration details. The last three columns represent task of the dataset (summary, action recognition, others), annotation of the dataset, and the papers that used the corresponding dataset for experiments. The popular datasets are discussed individually in next sub-sections.

4.1. Office dataset [10]

Office dataset is one the most popular datasets for MVS. This dataset is created using 4 stably-held cameras in an office, in positions that were not fixed. The four cameras were not synchronized with each other and there are different lighting conditions at different views in this dataset. Sample frames are shown in Fig. 5(a).

4.2. Lobby dataset [10]

Lobby dataset is provided by the same authors of Office dataset, captured by three cameras in a lobby area. All of the cameras in Lobby dataset were synchronized with each other, still, and non-fixed. There are very crowded scenes in this dataset, which makes it more challenging for summarization. Some sample images are shown in Fig. 5 (b).

4.3. Bl-7F [35]

Bl-7F is the largest among all MVS datasets. This dataset is created in Berrylam building in Taiwan University, where 19 surveillance cameras were installed on its 7^{th} floor. The recorded videos are very diverse and are challenging because of high-level overlapping between different views. Cameras installed are perfectly synchronised with each other and are still and fixed. Example frames are shown in Fig. 5 (c).

4.4. Campus [10]

The campus video dataset is recorded outdoor in a university campus with many trivial events. This dataset contains four views with 180-degree coverage. This dataset is created using web cameras or ordinary handheld cameras by non-specialists, meaning that it is unstable and obscure. Sample frames of campus dataset are visualised in Fig. 5 (d). The videos of campus dataset are challenging because they are not synchronised and contain motion of cameras.

5. Evaluation of MVS methods

The evaluation of MVS methods depends on the usage of datasets. Some methods in literature are evaluated using case

Table 5
Description, task and annotation of datasets used in MVS research provided with the papers references that used these datasets.

Dataset name	Year	Camera type	No. of views (Indoor/ Outdoor)	Shot description	No. of videos	Total duration	Task			Annotation	Cited by
							Summary	Action recognition	Others		
KTH [55]	2004	Fixed	Outdoor	-	2391	-	-	✔	-	Human actions	[37]
WEIZMANN [56]	2005	-	-	-	9	-	-	✔	-	-	[37]
PETS [57]	2009	-	-	-	85	-	-	✔	-	-	[37]
Lobby [10]	2010	Fixed	3/Indoor	✔	3	24 m 42s	✔	-	-	-	[10,35,43,45,39,31]
Office [10]	2010	Fixed	4/Indoor	✔	4	14 m 58s	✔	-	-	-	[10,35,43,27,45,39,31,48,49]
Campus [10]	2010	Fixed	4/Outdoor	✔	-	56 m 43 s	✔	-	-	-	[10,43,45,39]
[26]	2010	Internet website	-	-	88 sets	-	✔	-	-	Diverse videos	[26]
[58]	2012	Fixed	3	-		2 hours	✔	-	-	Human trajectory	[40]
PETS 2009 [59]	2012		4	-	4	3176 frames	-	-	✔	Human tracking	[41]
Soccer [33]	2014	Fixed	20/Outdoor	-	20	-	-	-	✔	Football match	[33]
D1 [50]	2014	Fixed	2/Outdoor/Indoor	-	-	6m 40s		✔		Pedestrian activity	[50]
D2 [50]	2014	Fixed	2 Outdoor/Indoor	-	-	10m 43s	-	-	✔	Vehicle surveillance	[40,50]
D3 [50]	2014	Fixed	2/Outdoor/Indoor	-	-	4m 45s	-	-	✔	Vehicle surveillance in night scenario	[50]
D4 [50]	2014	Fixed	3/Outdoor	-	-	3m 26s	-	-	✔	Vehicle surveillance in street scenario	[50]
D5 [50]	2014	Fixed	3/Outdoor	-	-	3m 31s	✔	-	-	Vehicle surveillance in a playground scenario	[50]
[33]	2014	Fixed	20/Outdoor	-	20	-	✔	-	-	Soccer videos	-
BL-7F [35]	2015	Fixed	19/Indoor	✔	19	7 m 10s	✔	-	-	-	[35,45,31]

Fig. 5. Sample frames from popular MVS datasets (a) Same frames of different views of Office dataset (b) three sample frames of different views of Lobby dataset (c) sequential frames with activity of BI-7F (four views), and (d) sample frames from handheld camera of campus dataset.

studies and there are some datasets that are publicly available along with the given ground truth of different events such sitting, putting on a coat, etc. The evaluation methods with reference of different techniques are explained below.

5.1. Subjective evaluation/user case studies

User case studies are utilised by many of the methods in MVS literature. The very first MVS method [10] provided both subjective evaluation and objective assessments. The major aim of objective evaluation is to assess three important aspects of the generated summary: enjoyability, informativeness, and usefulness. A finite number of participants are invited and asked for three questions. Q1: How about the enjoyability of the video summary? Q2: Do you think the information encoded in the summary is reliable compared to the original multi-view videos. Q3: Will you prefer the summary to original multi-view videos if stored in your computer? Inspired from [10],[48], and [49] used almost similar questions in subjective evaluation of their method.

5.2. Objective evaluation

The common evaluation metrics used by majority of the methods are Precision, Recall, and F1 score. Some methods also used event recall as a metric to show whether their method is able to extract the events as given in the ground truth. Precision indicates that how accurate a method is by calculating the number of false keyframes compared to the given ground truth as shown in Eq. 1. The value of recall corresponds to the ratio of matched keyframes with the ground truth frames as indicated in Eq. 2, while the F1 score considers both precision and recall to generate a representative score for both.

$$P = \frac{N_{matched}}{N_{extracted}} \tag{1}$$

$$R = \frac{N_{matched}}{N_{groundtruth}} \tag{2}$$

$$F1 = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{3}$$

P refers to precision, **R** indicates the value of recall, and **F1** shows the formula for F1 score given in Eqs. 1, 2, and 3, respectively. In these equations, "$N_{matched}$" refers to the total number of keyframes that are similar to the ground truth summary, "$N_{extracted}$" indicates the number of extracted frames by any method while "$N_{groundtruth}$" shows the total number of frames included in the ground truth summary. Although majority of the techniques did not provide any detail about how they calculated precision, recall, and F1 score but the formulas given in Eqs. 1, 2, and 3 are generic for video summarization methods. Some of the methods used different formulas for the computation of precision and recall, as given in Eqs. 4 and 5.

$$P = \frac{TP}{TP + NP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

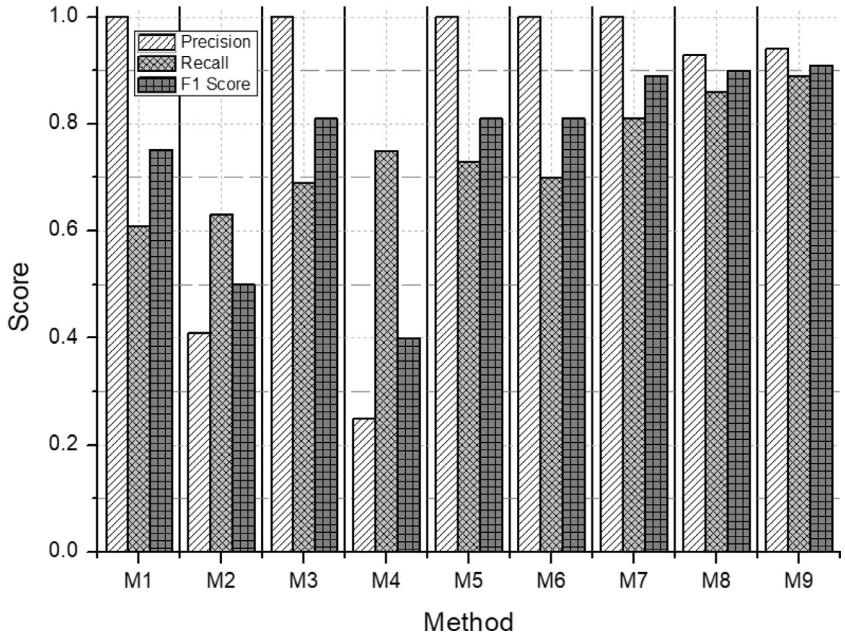

Fig. 6. Performance comparison of different MVS methods with respect to P, R, and F1 score on Office dataset. M1[10], M2[61], M3[11], M4[35], M5[43], M6[39], M7[45], M8[48], M9[49].

The above mentioned formulas are used by majority of researchers [35] for objective evaluation of their methods. Human operators marked the time periods of salient events in each video of MVS datasets and then those experts computed precision and recall value advanced to F1 score for their technique. The basic definition of the variables used in above equations are given below:

True positive (TP): A frame within the interval of marked salient event and also opted by given method

False positive (FP): A frame that is *not* present within the interval of salient event but is opted by the method as keyframes.

True negative (TN): A frame which is *not* in the range of salient event interval and is also not opted by the techniques.

False negative (FN): A frame in interval of salient event but is *not* opted by the technique.

Event recall is also calculated in this method [35]. An event is considered to be successfully extracted by the given method if it keeps more than 50% of its frames in the specific interval. Event recall is only calculated by [31,35] up to the present literature of MVS.

5.3. Performance evaluation of MVS methods

In this section, we discuss comparison of different MVS methods with respect to the metrics explained in the aforementioned paragraph. According to the best of our knowledge, there are overall 17 known research methods in MVS literature, where only eight of them share common datasets and perform evaluation using P, R, and F1 score. The datasets used by these methods for intercomparison are Office [10] and Bl-7f [35] and the ground truth for the events of these datasets is publicly available. Due to the subjective nature of video summarization it is impossible to compare methods based on the subjective evaluation of a dataset. Therefore, we only discuss the methods performing objective evaluation. The comparison of these methods is given in Fig. 6.

Office is one of the most popular datasets for MVS where most of the research scientists considers it for comparison with state-of-the-art. The performance comparison can be easily observed from Fig. 6, where the initiative paper of MVS achieves an F1 score of 0.75. An improvement in the F1 score can be observed in P3 where F1 score reaches 0.8. Similarly, improved results can be observed in recent methods, where a research in 2019 achieved the highest F1 score of 0.9 in the overall MVS literature. Another MVS dataset Bl-7f, that is presented by [35], initially acheived 0.6 F1 score that is improved by [11] to 0.85. This dataset is not used by any other method in literature to compare their results with the aforementioned methods. Time analysis of these MVS methods is not provided whereas the most recent method [48] reported overall summary generation time as 343.01 and 2048.88 seconds for cloud computing server and local computer, respectively. There are no further details about frames level processing of this method. A recent research method [49] described the execution time of their method for different frame rates. Authors also reported the execution time of each step involved in generating MVS. Readers are referred to Table 4 of the cited paper for detailed information. The time complexity of MVS methods is an important aspect for consideration of further research in literature which is not covered by majority of the techniques.

It is obvious from Fig. 6 that majority of the MVS methods achieve the highest possible value of precision, which indicates that all the keyframes extracted by these algorithms are in perfect match with the ground truth. In contrast, the recall value for most of the MVS methods has variation and sharp spikes due to the discrepancy between the matched keyframes and the number of frames present in the ground truth. The earlier MVS methods have lower recall value due to the usage of insufficiently presentable features of the input videos for final keyframes selection. Besides the usage of low-level statistical features, the algorithms utilising learned features [45] show convincing results for precision as well as recall value and finally produce a higher value of F1 score. Current state-of-the-art methods ([48] and [49]) perform on average for both precision and recall and thus final F1 score is comparatively greater than older MVS methods. The higher value achieved by these methods is due to the usage of human's appearance followed by information computation using deep learning [48] and low-level entropy features [49].

6. Challenges in MVS datasets

MVS datasets are more challenging as compared to SVS datasets as they have many issues such as instability of cameras while recording the video, lack of synchronisation among various cameras, and crowded scenes. The major challenges for MVS datasets along with references to relevant studies are discussed individually in sub-sequent sections.

6.1. Lack of synchronisation

Mainstream videos in MVS datasets of different views are not synchronised with each other. For instance, the videos of Office dataset [10] are highly un-synchronised. Soccer [33] dataset videos are also not synchronised during recording in the field, but later, they are synchronised manually. The problem of synchronization makes the summary generation a difficult task because it becomes challenging to compute the inter-view correlations among different unsynchronised videos. To tackle this problem, the authors of [35] manually aligned these videos and then performed experiments for summary generation. Besides this, some other techniques such as that in [45] intelligently found correlations through various features matching algorithms for summary generation.

6.2. Instability of camera

The cameras in some of the MVS datasets such as Bl-7F [35] are stable and fixed with no shuddering, whereas for Lobby and Office [10] dataset, cameras are stably held, but non-fixed. Videos captured by such cameras that are affected with motion blur makes it difficult to generate a good and satisfactory summary. To handle instability of camera and light condition problem, R. Panda et al. [43] used BVLC CaffeNet pre-trained model, and extracted 4096-dimensional CNN feature vector. This feature vector is acquired from top layer hidden unit activations of the CNN network, showing a global representation of the input image. These deep features show the best performance with videos of such instability and variable light conditions.

6.3. Crowded scenes

In MVS literature, Lobby [10] and Soccer [33] datasets are very crowded and contain richer activities as compared to Office [10] and Bl-7F [35], that are recorded in indoor environments. It is challenging to work with crowded scenes and heavy traffic as compared to dealing with simple scenes of limited persons doing some activities. The crowd in Lobby [10] dataset is the most challenging problem, while Office [10] and Bl-7F [35] are the smiplest datasets in MVS literature in terms of crowd density. In addition to these challenges, Office [10] dataset also contains unstable frame rate, and it suffers from highly variable light variations. The problem of a crowded scene is tackled in [11] by filtering the activities, thus the scenes with reduced activities are suppressed through Gaussian entropy.

7. Recommendations and future research directions

It has been observed that MVS problem is not adequately addressed as per need ofMVS for a vast amount of applications. It is apparent from the reviewed literature that to date, most of the research is based on handcrafted-features or mid-level features. Similarly, various clustering techniques and traditional machine learning-based classifiers are used in almost all reviewed techniques for final summary generation. Recommendations and future directions about MVS are provided as follows:

7.1. End-to-End deep learning models

Although there are some techniques in literature which select learned features for the summary generation as a prerequisite step, yet there are no such deep learning models that can input MVVs and directly generate their output summary. MVS literature lacks such CNN architectures that can achieve the task of correlations computation between different views and find salient frames intelligently. End-to-end deep learning models are used for various purposes, mainly for speech recognition [60] and others tasks [61]. Thus, in future work, it is highly recommended to propose such end-to-end deep learning models that can provide a summary for MVVs with satisfactory accuracy while preserving the properties of a good summary. Such end-to-end networks should input multi frames or sequence of frames from a network of cameras and process it through different layers such as convolutions, pooling etc. for final output summary. Most importantly correlation layers can be explored for inter-view correlation computation, as used in [62] for finding optical flow between two consecutive frames.

7.2. Standard datasets for benchmarking

The currently available datasets of literature involve many challenges such as lack of synchronisation, variable lighting conditions, unstable frame rates, and non-fixed cameras. But all these datasets only focus on some specific challenges, and they are not enough for better evaluation of MVS techniques. Furthermore, all these datasets are recorded in normal indoor or outdoor environments. In future work, researchers who are enthusiastic to work in MVS field should focus on creating standard datasets which cover all sorts of environments including normal and abnormal or uncertain conditions [63] (fire, fog, snow, rainfall, etc.). Similarly, outdoor environments capturing different actions and activities should be aimed while creating multi-view datasets such as majority videos in [64] are captured from outdoor surveillance. This can help building up MVS methods functional in every sort of scenario and deployable for real-time environments.

7.3. Intensive, efficient, and effective utilization of hardware resources

The hardware resource utilisation in existing MVS literature is inadequate, without any parallel processing mechanisms or multi-threading. Similarly, the modern available tools and resources are rarely utilised by MVS methods for effective and efficient output generation. The related tools and concepts that can be considered in future are given in subsequent sections.

7.3.1. Agents based MVS

In MVS literature, the state-of-the-art techniques generate summary on a solo processor i.e., performing every step of the algorithm on a single computer. As evident from the introduction section, SVS generates summary by processing a single video but MVS, in contrast, processes multiple videos to generate an output which makes the summary generation comparatively slow. Therefore, if the task of MVS is divided into different modules between certain agents, it will require less processing time and the tasks can be processed in parallel, boosting execution and the overall summary generation process. Several multi-agents-based systems for various applications are presented such as that in [65] for motive profiling of users in virtual worlds and gaming and method in [66] for energy minimization in cloud computing systems. Thus, researchers in future should exploit dividing the load into several agents and integrating their individual outputs to finally generate a summary.

7.3.2. Fog/cloud computing for MVS

Fog computing [67] is a layer of a distributed network that can work much faster than a single computer. It is used in differ-

ent computer vision applications such as healthcare [68], security systems [69], web applications [70], and video analytics [48]. So far, every state-of-the-art technique generates summary locally, but processing these videos through Cloud or Fog computing would make the output generation very faster. Generating a summary on a cloud server has several other advantages, such as it can be instantly used for other applications including abnormal action and activity recognition, events detection, and video retrieval.

7.3.3. Edge intelligence for MVS

Edge intelligence [71] refers to the processing, development, and analysis of data at the site where generated. It can be used for many types of sensors, such as a visual sensors to analyse the scenes captured in real time, or to detect abnormal activities, etc. For future research in MVS domain, embedded programs can be used in high definition cameras for MVS at the site independently. Therefore, instead of sending all the videos over wireless or local networks, it is better to send only generated summary which could assist analysing video in a short period of time, and could save bandwidth and the precious time of surveillance analysts [72].

7.4. Processing time for MVS

The experimental results provided by the current state-of-the-art MVS methods perform evaluation using only accuracy with no focus on their running time and computational feasibility for deployment in real-world surveillance networks. The MVS literature lacks an assessment of running time of algorithms provided with their system configurations. In current research efforts of SVS and other video analysis related fields, the contribution of new methods exists either in terms of accuracy or efficiency. Thus, it is recommended to evaluate the newly developed MVS methods from the perspective of running time and provide their full implementation details. It will also lead the MVS research community in a new direction to decide whether a system is efficient and satisfactory enough to be implemented in real-world scenarios.

7.5. MVS for resource constrained devices

Future research into MVS should also aim to generate automated summaries through resource-constrained devices such as Raspberry Pi, FPGA, Adriano, etc. These resource-constrained devices have abilities that can be utilised for wide range of applications [68]. MVS can be easily acheived through such devices as they can be attached together to build a small network of cameras, working independently for output creation. Thus, there will be reduced time complexity for MVS generation along with reduced computational power, if performed over resource-constrained devices.

8. Conclusion and future directions

The numbers of surveillance cameras providing single- or multi-view coverage are increasing exponentially. The distributed video cameras provide better courage of a scene and also generates comparatively huge amount of video data in contrast to single-view cameras. These Big Data contain rare events but most of them are redundant frames without any salient information. Extracting salient contents from such Big Data instigates the need of MVS techniques.

We have presented a complete survey of state-of-the-art techniques for MVS. As there exists limited number of articles in the MVS literature, therefore, in this survey we investigated each paper from several aspects. We presented how the MVS trend is deveoped from the very beginning of MVS literature in 2010 and described the generic working of MVS methods with each step explained separately along with references. Next, we presented an overview of the datasets used by different MVS techniques. Finally, after summarizing the whole literature, we provided recommendations and future directions for further work in the field of MVS. Similarly, the application wise distribution of MVS literature, generic flow of MVS methods, and a brief discussion of MVS datasets can lead researchers in different deployable directions of MVS such as industries, law and enforcement, and entertainment etc.

There are many future research directions for MVS in deep learning areas, resource-constrained computing, edge and cloud computing, among others, as explained in Section VII. Intelligent end-to-end deep learning models are required for efficient generation of MVS. The traditional approaches of low-level features based clustering or classification need to be replaced by learned features from deep learning models to achieve better accuracy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2B5B01070067)

References

[1] S. Antani, R. Kasturi, R. Jain, A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video, Pattern Recogn. 35 (2002) 945–965.
[2] W. Wang, J. Shen, X. Li, F. Porikli, Robust video object cosegmentation, IEEE Trans. Image Process. 24 (2015) 3137–3148.
[3] P. Wang, L. Liu, C. Shen, H.T. Shen, Order-aware convolutional pooling for video based action recognition, Pattern Recogn. 91 (2019) 357–365.
[4] M. Babaee, D.T. Dinh, G. Rigoll, A deep convolutional neural network for video sequence background subtraction, Pattern Recogn. 76 (2018) 635–649 04/01/ 2018.
[5] L. Zhang, L. Sun, W. Wang, Y. Tian, KaaS: A standard framework proposal on video skimming, IEEE Internet Comput. 20 (2016) 54–59.
[6] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, D. Dagan Feng, Video summarization via minimum sparse reconstruction, Pattern Recogn. 48 (2015) 522–533 02/01/ 2015.
[7] J. Zhu, S. Feng, D. Yi, S. Liao, Z. Lei, S.Z. Li, High-performance video condensation system, IEEE Trans. Circuit Syst. Video Tech. 25 (2014) 1113–1124.
[8] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino, V.H.C. de Albuquerque, Cost-Effective Video Summarization using Deep CNN with Hierarchical Weighted Fusion for IoT Surveillance Networks, IEEE Internet of Things J. 7 (5) (2019) 4455–4464.
[9] M. Ali, A. Anjum, M.U. Yaseen, A.R. Zamani, D. Balouek-Thomert, O. Rana, et al., Edge enhanced deep learning system for large-scale video stream analytics, in: 2018 IEEE 2nd International Conference on Fog and Edge Computing (ICFEC), 2018, pp. 1–10.
[10] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, Z.-H. Zhou, Multi-view video summarization, IEEE Trans. Multimedia 12 (2010) 717–729.
[11] S.K. Kuanar, K.B. Ranga, A.S. Chowdhury, Multi-view video summarization using bipartite matching constrained optimum-path forest clustering, IEEE Trans. Multimedia 17 (2015) 1166–1173.
[12] R. Panda, A. Das, A.K. Roy-Chowdhury, Embedded sparse coding for summarizing multi-view videos, in: 2016 IEEE international conference on image processing (ICIP), 2016, pp. 191–195.
[13] D. Wang, W. Ouyang, W. Li, D. Xu, Dividing and aggregating network for multi-view action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 451–467.
[14] Ö. Aköz, M.E. Karslıgil, Video-based traffic accident analysis at intersections using partial vehicle trajectories, in: 2010 IEEE 18th Signal Processing and Communications Applications Conference, 2010, pp. 499–502.
[15] A.A. Abdellatif, A. Mohamed, C.F. Chiasserini, M. Tlili, A. Erbad, Edge Computing for Smart Health: Context-Aware Approaches, Opportunities, and Challenges, IEEE Netw. 33 (2019) 196–203.
[16] N.H. Phuong, V. Kreinovich, Fuzzy logic and its applications in medicine, Int. J. Med. Inform. 62 (2001) 165–173 07/01/ 2001.
[17] M. Dursun, E.E. Karsak, M.A. Karadayi, Assessment of health-care waste treatment alternatives using fuzzy multi-criteria decision making approaches, Resour. Conserv. Recycl. 57 (2011) 98–107 12/01/ 2011.

[18] H. Wu, Y. Hu, K. Wang, H. Li, L. Nie, H. Cheng, Instance-aware representation learning and association for online multi-person tracking, Pattern Recogn. 94 (2019) 25–34 10/01/ 2019.
[19] K. Muhammad, J. Ahmad, S.W. Baik, Early fire detection using convolutional neural networks during surveillance for effective disaster management, Neurocomputing 288 (2018) 30–42 05/02/ 2018.
[20] Y. Yuan, Y. Feng, X. Lu, Structured dictionary learning for abnormal event detection in crowded scenes, Pattern Recogn. 73 (2018) 99–110 01/01/ 2018.
[21] H. Fu, D. Xu, B. Zhang, S. Lin, Object-based multiple foreground video co-segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3166–3173.
[22] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recogn. 92 (2019) 177–191 08/01/ 2019.
[23] D. Singh, C. Krishna Mohan, Graph formulation of video activities for abnormal activity recognition, Pattern Recogn. 65 (2017) 265–272 05/01/ 2017.
[24] Y. Zhuang, R. Xiao, F. Wu, Key issues in video summarization and its application, in: Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 2003, pp. 448–452.
[25] H.-S. Park, S.-B. Cho, A fuzzy rule-based system with ontology for summarization of multi-camera event sequences, in: International Conference on Artificial Intelligence and Soft Computing, 2008, pp. 850–860.
[26] Y. Li, B. Merialdo, Multi-video summarization based on Video-MMR, in: Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on, 2010, pp. 1–4.
[27] Y. Fu, L. Wang, Y. Guo, Multi-view metric learning for multi-view video summarization, arXiv preprint arXiv:1405.6434 (2014).
[28] L. Xu, D. Schuurmans, Unsupervised and semi-supervised multi-class support vector machines, in: AAAI, 2005, p. 13.
[29] B. Long, P.S. Yu, Z. Zhang, A general model for multiple view unsupervised learning, in: Proceedings of the 2008 SIAM international conference on data mining, 2008, pp. 822–833.
[30] M.Y. Zhang, W.Y. Cai, "Multi-view Video Summarization Algorithm for WMSN," in *2014 International Conference on Wireless Communication and Sensor Network*, 2014, pp. 213–216.
[31] S.H. Ou, Y.C. Lu, J.P. Wang, S.Y. Chien, S.D. Lin, M.Y. Yeti, et al., Communication-efficient multi-view keyframe extraction in distributed video sensors, in: 2014 IEEE Visual Communications and Image Processing Conference, 2014, pp. 13–16.
[32] A.I. Ioannidis, V.T. Chasanis, A.C. Likas, Key-Frame Extraction Using Weighted Multi-view Convex Mixture Models and Spectral Clustering, in: 2014 22nd International Conference on Pattern Recognition, 2014, pp. 3463–3468.
[33] Y. Muramatsu, T. Hirayama, K. Mase, Video generation method based on user's tendency of viewpoint selection for multi-view video contents, in: Proceedings of the 5th Augmented Human International Conference, 2014, p. 1.
[34] C. d. Leo, B.S. Manjunath, Multicamera video summarization and anomaly detection from activity motifs, ACM Trans. Sens. Netw. (TOSN) 10 (2014) 27.
[35] S.-H. Ou, C.-H. Lee, V.S. Somayazulu, Y.-K. Chen, S.-Y. Chien, On-line multi-view video summarization for wireless video sensor network, IEEE J. Select. Top. Signal Process. 9 (2015) 165–179.
[36] Y. Hoshen, S. Peleg, Live video synopsis for multiple cameras, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 212–216.
[37] A. Mahapatra, P.K. Sa, B. Majhi, A multi-view video synopsis framework, in: Image Processing (ICIP), 2015 IEEE International Conference on, 2015, pp. 1260–1264.
[38] L. Nie, R. Hong, L. Zhang, Y. Xia, D. Tao, N. Sebe, Perceptual Attributes Optimization for Multivideo Summarization, IEEE Trans. Cybernet. 46 (2016) 2991–3003.
[39] R. Panda, A. Das, A.K. Roy-Chowdhury, Embedded sparse coding for summarizing multi-view videos, in: Image Processing (ICIP), 2016 IEEE International Conference on, 2016, pp. 191–195.
[40] J. Zhu, S. Liao, S.Z. Li, Multicamera joint video synopsis, IEEE Trans. Circuit Syst. Video Tech. 26 (2016) 1058–1069.
[41] A. Mahapatra, P.K. Sa, B. Majhi, S. Padhy, MVS: A multi-view video synopsis framework, Sign. Process. Image Commun. 42 (2016) 31–44 03/01/ 2016.
[42] L. Wang, X. Fang, Y. Guo, Y. Fu, Multi-view metric learning for multi-view video summarization, in: 2016 International Conference on Cyberworlds (CW), 2016, pp. 179–182.
[43] R. Panda, A. Dasy, A.K. Roy-Chowdhury, Video summarization in a multi-view camera network, in: Pattern Recognition (ICPR), 2016 23rd International Conference on, 2016, pp. 2971–2976.
[44] R. Panda, N.C. Mithun, A.K. Roy-Chowdhury, Diversity-Aware Multi-Video Summarization, IEEE Trans. Image Process. 26 (2017) 4712–4724.
[45] R. Panda, A.K. Roy-Chowdhury, Multi-view surveillance video summarization via joint embedding and sparse optimization, arXiv preprint arXiv:1706.03121 (2017).
[46] J. Meng, S. Wang, H. Wang, Y.P. Tan, J. Yuan, Video Summarization via Multi-view Representative Selection, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 1189–1198.
[47] J. Meng, S. Wang, H. Wang, J. Yuan, Y.-P. Tan, Video summarization via multi-view representative selection, IEEE Trans. Image Process. (2018) 2134–2145.
[48] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S.W. Baik, V.H.C. d. Albuquerque, Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM, IEEE Trans. Indust. Inform. 16 (2020) 77–86.
[49] T. Hussain, K. Muhammad, J.D. Ser, S.W. Baik, V.H.C. d. Albuquerque, Intelligent Embedded Vision for Summarization of Multiview Videos in IIoT, IEEE Trans. Indust. Inform. 16 (2020) 2592–2602.
[50] X. Zhu, J. Liu, J. Wang, H. Lu, Key observation selection-based effective video synopsis for camera network, Mach. Vis. Appl. 25 (2014) 145–157.
[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
[52] F. Daniyal, A. Cavallaro, Multi-camera scheduling for video production, in: Visual Media Production (CVMP), 2011 Conference for, 2011, pp. 11–20.
[53] Y.-G. Jiang, C.-W. Ngo, Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, Comp. Vis. Image Understand. 113 (2009) 405–414.
[54] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, SimpleMKL, J. Mach. Learn. Res. 9 (2008) 2491–2521.
[55] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 2004, pp. 32–36.
[56] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 2005, pp. 1395–1402.
[57] J. Ferryman, A. Shahrokni, Pets2009: Dataset and challenge, in: Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on, 2009, pp. 1–6.
[58] X. Chen, K. Huang, T. Tan, Learning the three factors of a non-overlapping multi-camera network topology, in: Chinese Conference on Pattern Recognition, 2012, pp. 104–112.
[59] B. Yang, R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1918–1925.
[60] Y. Miao, M. Gowayyed, F. Metze, EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 167–174.
[61] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, et al., Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition, Pattern Recogn. 71 (2017) 196–206 11/01/ 2017.
[62] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2462–2470.
[63] K. Muhammad, S. Khan, M. Elhoseny, S.H. Ahmed, S.W. Baik, Efficient Fire Detection for Uncertain Surveillance Environment, IEEE Trans. Indust. Inform. 15 (2019) 3113–3122.
[64] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479–6488.
[65] L. Xuejie, K. Merrick, H. Abbass, Designing artificial agents to detect the motive profile of users in virtual worlds and games, in: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1–8.
[66] W. Wang, Y. Jiang, W. Wu, Multiagent-Based Resource Allocation for Energy Minimization in Cloud Computing Systems, IEEE Trans. Syst. Man Cybernet. Syst. 47 (2017) 205–220.
[67] H. El-Sayed, S. Sankar, M. Prasad, D. Puthal, A. Gupta, M. Mohanty, et al., Edge of things: The big picture on the integration of edge, IoT and the cloud in a distributed computing environment, IEEE Access 6 (2017) 1706–1717.
[68] M. Sajjad, K. Muhammad, S.W. Baik, S. Rho, Z. Jan, S.-S. Yeo, et al., Mobile-cloud assisted framework for selective encryption of medical images with steganography for resource-constrained devices, Multimed. Tool Appl. 76 (2017) 3519–3536.
[69] S. Alharbi, P. Rodriguez, R. Maharaja, P. Iyer, N. Bose, Z. Ye, FOCUS: A fog computing-based security system for the Internet of Things, in: 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), 2018, pp. 1–5.
[70] T. Wang, W. Zhang, C. Ye, J. Wei, H. Zhong, T. Huang, FD4C: Automatic Fault Diagnosis Framework for Web Applications in Cloud Computing, IEEE Trans. Syst. Man Cybernet. Syst. 46 (2016) 61–75.
[71] T.X. Tran, A. Hajisami, P. Pandey, D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *arXiv preprint arXiv:1612.03184*, 2016.
[72] J. Wu, B. Cheng, M. Wang, J. Chen, Energy-efficient bandwidth aggregation for delay-constrained video over heterogeneous wireless networks, IEEE J. Select. Area Commun. 35 (2016) 30–49.

Tanveer Hussain acknowledged his degree of Bachelor's in Computer Science from Islamia College Peshawar, Peshawar, Pakistan with Gold Medal distinction. Currently, he is enrolled in joint Master and Ph.D. program at Sejong University, Seoul, Republic of Korea and serving as a Research Assistant at Intelligent Media Laboratory (IM Lab). His major research domains are features extraction (learned and low-level features), video analytics, single/multi-view video summarization, IoT, and resource-constrained programming. He has published several journal articles in these areas in reputed journals including IEEE Network, TII, IoTJ, Elsevier PR, PRL, MDPI Sensors, and Wiley IJDSN. For further activities and implementations, visit: https://github.com/tanveer-hussain .

Khan Muhammad is an Assistant Professor at Department of Software and lead researcher of Intelligent Media Laboratory (IM Lab), Sejong University, Seoul, South

Korea. His research interests include medical image analysis (brain MRI, diagnostic hysteroscopy and wireless capsule endoscopy), information security (steganography, encryption, watermarking and image hashing), video summarization (single-view and multi-view), multimedia, computer vision, IoT and smart cities, and video analytics. He has published over 100 papers in peer-reviewed international journals and conferences in these research areas with target venues as IEEE COMMAG, NETWORK, TII, TIE, TSMC-Systems, IoTJ, Access, TSC, Elsevier INS, Neurocomputing, ASOC, PRL, FGCS, COMCOM, COMIND, JPDC, PMC, BSPC, CAEE, Springer NCAA, MTAP, JOMS, and RTIP, etc. He is serving as a professional reviewer for over 70 well-reputed journals and conferences including IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Industrial Informatics, IEEE Communication Magazine, IEEE Internet of Things (IoT), IEEE Network, IEEE Access, IEEE Transactions on Circuits and Systems for Video Technology, ACM Transactions on Multimedia Computing Communications and Applications, Elsevier Future Generation Computer Systems, Elsevier Neurocomputing, Elsevier Computer Methods and Programs in Biomedicine, Elsevier International Journal of Information Management, Elsevier Signal Processing, Elsevier Journal of Network and Computer Applications, Springer Multimedia Tools and Applications, Springer Mobile Networks and Applications, Springer Nonlinear Dynamics, Springer Medical & Biological Engineering & Computing, Springer Journal of Super Computing, Springer EURASIP Journal on Image and Video Processing, SPIE Journal of Electronic Imaging, T&F Journal of Experimental and Theoretical AI, SAGE International Journal of Distributed Sensor Networks, Hindawi Computational Intelligence and Neuroscience, CCCT 2015, ICNC 2017, AINA 2017, CMES-2018, INDIN 2019, IMC 2020, and NCAA 2020. He also acted as TPC member of AINA 2017 for the track "Multimedia Systems and Applications" and now as a program chair for ICNGC 2019. He is currently involved in editing of several special issues as GE/LGE. He is a member of the IEEE and ACM, visit: https://khan-muhammad.github.io/03_activities/

Weiping Ding is currently a Full Professor at Nantong University and was awarded the Excellent Paper (First Prize) in Computer Education by the Chinese National Committee of Computer Education and a Chinese Government Scholarship for Overseas Studies in 2011 and 2016. He was an Excellent Young Teacher in Jiangsu Province in 2014 and a High-level Talent in Jiangsu Province in 2016. Dr. Ding currently serves as an Associate Editor of IEEE Transactions on Fuzzy Systems (2015-) and Information Sciences (2016-). His main research interests include data mining, quantum co-evolutionary computing, and machine learning, and their applications in big data.

Jaime Lloret is an associate professor at Politechnic University of Valencia, Spain. He was Internet Technical Committee Chair during 2014–2015 and is the current Chair of IEEE 1907.1. He is the director of the Research Institute IGIC and head of the Innovation Group EITA-CURTE. He is co-Editor-in-Chief of Ad Hoc and Sensor Wireless Networks and Editor-in-Chief of Network Protocols and Algorithms. He has been General Chair of 36 international work-shops and conferences

Sung Wook Baik received the B.S degree in computer science from Seoul National University, Seoul, Korea, in 1987, the M.S. degree in computer science from Northern Illinois University, Dekalb, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Fairfax, VA, in 1999. He worked at Datamat Systems Research Inc. as a senior scientist of the Intelligent Systems Group from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, Korea, where he is currently a Full Professor and the Chief of Sejong Industry-Academy Cooperation Foundation. He is also the head of Intelligent Media Laboratory (IM Lab) at Sejong University. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games. He is a member of the IEEE.

Victor Hugo C. de Albuquerque received the graduation degree in mechatronics technology from the Federal Center of Technological Education of Ceará, Fortaleza, Brazil, in 2006, the M.Sc. degree in tele-informatics engineering from the Federal University of Ceará, Fortaleza, in 2007, and the Ph.D. degree in mechanical engineering with emphasis on materials from the Federal University of Paraíba, João Pessoa, Brazil, in 2010. He is currently an Associate Professor with the Graduate Program in Applied Informatics at the University of Fortaleza, Fortaleza. He has experience in computer systems, mainly in the research fields of applied computing, intelligent systems, visualization and interaction, with specific interest in pattern recognition, artificial intelligence, image processing and analysis, Internet of Things, Internet of Health Things, as well as automation with respect to biological signal/image processing, image segmentation, biomedical circuits, and human/brain–machine interaction, including augmented and virtual reality simulation modeling for animals and humans.