A Comprehensive Survey on Multi-View Video Summarization

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

[a]Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-147, Republic of Korea
[b]School of Computer Science and Technology, Nantong University, Nantong 226019, China
[c]Politechnic University of Valencia, Spain
[d]Laboratory of Bioinformatics, University of Fortaleza, Brazil

***Abstract*—Nowadays, there is an exponential growth in visual data on a daily basis acquired from surveillance single-camera or multi-view camera networks. The massive amount of data requires efficient mechanisms such as video summarization to only receive significant data and reduce redundancy. Multi-view video summarization (MVS) provides less redundant and concise information of the video content from all the cameras either in the form of keyframes or video segments. This paper presents an overview of the existing strategies proposed for MVS, including their achievements and drawbacks. Our survey covers the generic steps of MVS like pre-processing of video data, feature extraction, post-processing followed by summary generation. Moreover, this paper describes the available datasets for the evaluation of MVS. Finally, this paper covers the major issues of current MVS and highlights the recommendations for future research.**

***Index Terms*—Computer Vision, Multi-View Video Summarization, Multi-Sensor Management, Multi-Camera Network, Machine Learning, Features Fusion, Big Data.**

I. INTRODUCTION

Due to the recent mounting of video data in surveillance cameras, it has become very challenging to process the whole data for various applications like crowd analysis [1, 2], video semantic recognition [3, 4], browsing, retrieval [5, 6], and analysis [7, 8]. Manual extraction of informative parts from video data and its processing is a laborious task, thus there is a need for automatic techniques to remove redundancy and useless information. To solve these challenges, various techniques such as video skimming [9-11], video summarization [12, 13], and video condensation [14-16] have been presented. Video skim is a short segment of the original video, covering its overall representation. Video summarization is a technique used to extract the salient frames or sequence of frames from a video. Further, it aims at fast browsing by shortening the input video into a synopsis and keeps the salient information inside the input video. Video summarization is a hot area of research because of the massive growth of video data captured for surveillance or recorded through smartphones on a daily basis. It is broadly divided into single-view video summarization (SVS) and MVS. SVS is the process of creating a summary of single video and the generated summary should preserve three properties including minimum repetition, representativeness, and diversity [17]. 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. In SVS there is no problem of synchronization or illumination difference among different views compared in MVS. 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. Similar to SVS the output generated from MVS is either number of representative frames (keyframes), short and comprehensive video (video synopsis) or video skims. Unlike SVS, the input of MVS is acquired from different views of various

cameras. The basic pipeline for processing such videos includes preprocessing, feature extraction, post-processing, and summary generation. Preprocessing of multi-view videos (MVV) comprises typically of redundancy removal steps such as segmentation [18-22] or shot boundary detection [23-25]. Preprocessing is followed by features extraction, object detection or tracking which varies from one application to another. Feature extraction methods include handcrafted features [23, 26] such as, color histograms, edge histograms, color layout descriptor, and learned features through CNNs as presented by [27, 28]. Post-processing of the extracted features refers to computing intra and inter-view correlations as used in most of the existing methods, followed by summary generation. The most common approaches for summary generation are based on machine learning techniques or template matching in MVS literature. The final generated output varies, and the possible forms are 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 [29, 30], investigating accident scenarios [31, 32], interesting and salient events extraction from sports, security and law enforcement for theft detection, and robbery events recognition. Despite these, it has applications in virtual reality for creating a single 360 view through multi-view cameras [33, 34].

There are two big challenges encountered while summarizing MVV: first is the inter-view and intra-view correlations among different views and second is the problem of synchronization, different lighting conditions for different views, and possible overlapping of views in MVS. A simple multi-view network contains two cameras with two output videos. For example, consider two cameras acquiring video from two different views with 25 fps, there are 15,000 (7500 for each camera) for one-hour video content. This huge amount of data is the big hurdle in exploitation of important visual data, therefore extracting the interesting and specific perspective of videos is essential instead of watching the whole video. Literature shows that several companies are already working on MVV for different purposes such as car parking, support comfortable, and safe driving. For instance, Honda Motor Co., Ltd [35] announced multi-view camera system that visualizes different views with multiple wide-angles from cameras fed in the vehicle. Similarly, many other companies are developing multi-view camera systems for different applications such as [36], which can record video from multiple views in a very warm environment, providing assistance for further analysis of the scene. Similarly, based on the multi-view camera network, a software developed by [37] can monitor and control users working inside a company and the ongoing processes within the company.

In today's technological era, cameras are installed in office, campus, industry, streets, and public places. These cameras capture 24 hours' data on daily basis which can be used for many purposes like persons tracking [38-40], disasters alerting [41, 42], action recognition [43], and different video analytics domains [44-46].To date, the field of MVS is less exploited while there is an urgent need of MVS techniques to process video data effectively for various applications. Most of the current research works use traditional hand-engineered features for MVS. Deep learning recently evolved in computer vision for various applications [47] such as disaster management [41], security [48-50], action recognition [51, 52], classification [53], and dynamic energy management [54]. Furthermore, deep learning has been implemented for consensus tracking problem in heterogeneous linear multi-agent systems [55], revenues prediction of movies [56], learning meaningful representations [57], and for various applications with integration to fuzzy systems [58, 59] in industries [60-62]. So, it is highly 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 present in MVS literature to date 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 MVS literature, its distribution on the basis of publishers, citations, type of research papers, and application-wise scattering of MVS approaches. Further, this survey provides results of all quires for searching MVS papers in various repositories. Moreover, we provide remarks about the selection process of the retrieved papers in the survey.

3. This survey explores the current challenges of MVS methods and datasets and concludes the overall literature. Finally, our survey provides recommendations and future research directions for further exploration of MVS field.

Rest of the paper is structurally divided into five 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. The available MVV datasets with their characteristics are discussed in Section IV. Section V highlights the major challenges of MVS and provides recommendations for future research. Section VI concludes this survey.

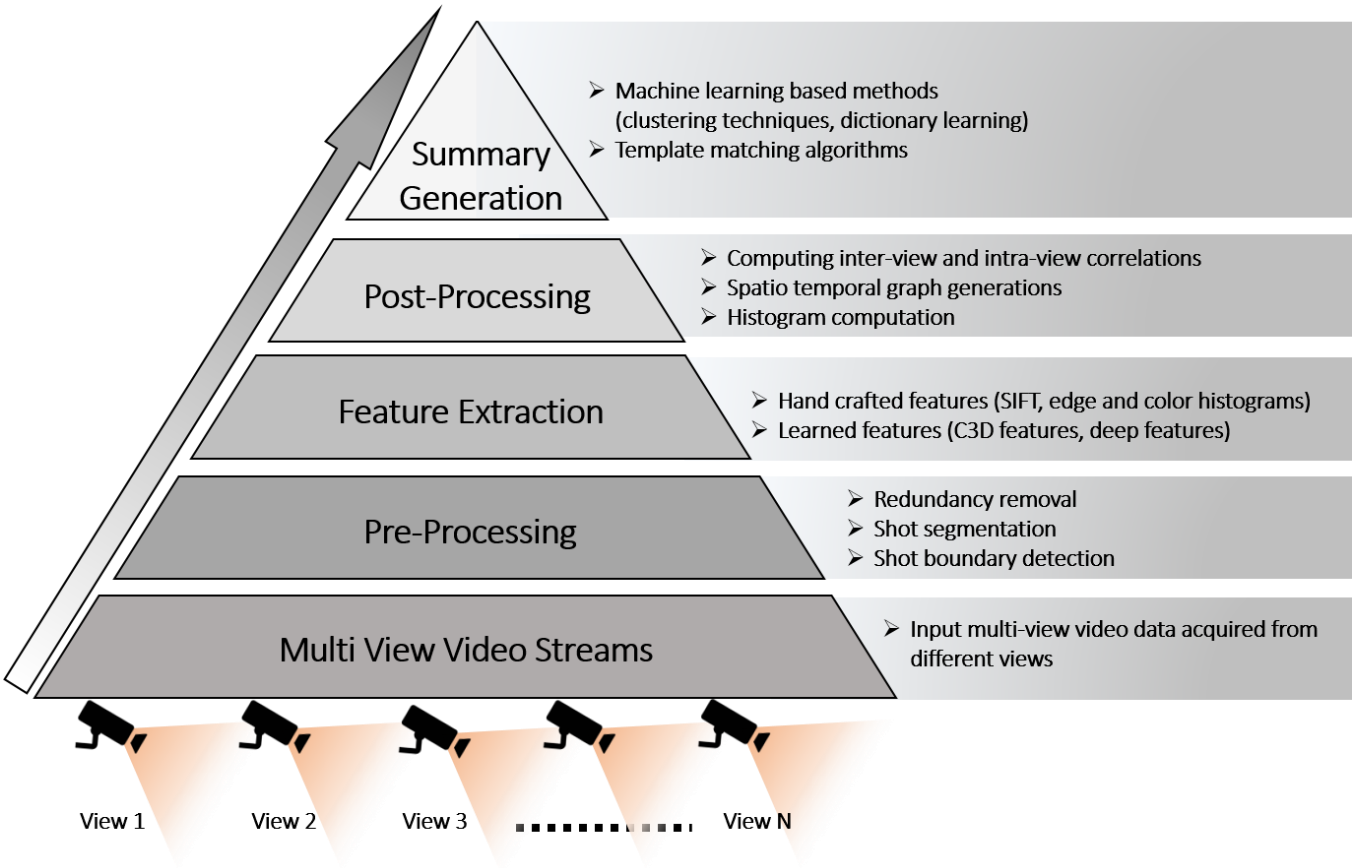

Fig. 1: General flow of MVS methods (condensation of multi-view videos into a short and comprehensive output summary)

II. Scope, Outline, and Coverage of This Survey

This tutorial covers multi-view (MV) workshops, journals, and conference papers from diverse repositories including Google Scholar, ScienceDirect, IEEE Xplore, and Springer. We searched related articles with different queries in all these repositories, some of the search items retrieved are excluded because of irrelevancy. The overall items explored with different searches made in several repositories are given in Table **I** along with remarks for not considering some papers in this survey. Table **II** shows 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). The scattered chart of

MVS conference papers, journals, and other category papers are visualized in Fig 2 (c). In MVS literature most of the papers are published in IEEE journals and top conferences. Distribution of MVS papers through various publishers is shown in Fig. **2** (d).

TABLE I

COMPREHENSIVE DETAILS OF SEARCHED LITERATURE DISTRIBUTED THROUGH VARIOUS PUBLISHERS IN MVS

S. No	Method	Year	Included	Remarks
1	[63]	2005	✗	Multi-view keyword in this paper is referred as human faces and video captioning problem individually based on low-level visual features. Thus, they considered human faces (one view), video captioning (second view) to generate summary of a single-view video.
2	[26]	2010	✓	Authors proposed an MVS framework using random walks where hypergraphs are used for capturing correlations among different views.
3	[64]	2010	✓	In this work, maximal marginal relevance (a concept of text summarization) is intelligently utilized to generate MVS.
4	[65]	2014	✓	Concept of maximum-margin clustering [66] and disagreement minimization criterion [67] is integrated together in this work with metric learning for generation of summary from multiple views.
5	[68]	2014	✗	This is only proposal paper, methodology is not explained, and the authors mentioned to present the work in future. No evaluation performed, and no standard dataset followed in experiments.
6	[69]	2014	✓	Inspired from previous work [64], the authors proposed an online MVS system by integrating maximal marginal relevance with a bandwidth efficient distributed algorithm.
7	[70]	2014	✗	This method generated a single view video summary based on convex mixture models and spectral clustering.
9	[71]	2014	✓	In this framework, authors utilized user's view tendency for the selection of viewpoint for multi-view video contents. This framework showed good results for sport events and live concerts.
10	[72]	2015	✓	In this article the main theme is to reduce compression and transmission power. This scheme comprises of online and offline modules for MVS.
11	[23]	2015	✓	This technique used semantic feature in the form of visual bag of words. Gaussian entropy, bipartite graph matching and optimum-path forest algorithm is used for summary generation.
12	[73]	2015	✗	This method is focused on making decision about persons on the basis of their past activities in live surveillance and it does not generates multi-view summary.
14	[74]	2015	✓	In this article, authors proposed a scheme to produce video synopsis from multi-view videos based on human actions, in both indoor and outdoor scenarios.
15	[75]	2016	✗	This method is focused only on multi-video summarization, MVS is not covered in this article.
16	[76]	2016	✓	In this paper, authors presented a novel technique based on joint embedding and sparse coding for summarization of multi-view videos.
17	[77]	2016	✓	A multi-camera joint video synopsis is presented that finds object's appearing, merging, splitting and disappearing moments in the frame sequence called as tube from each view. These tubes are joined by rearranging them such 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	[78]	2016	✓	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.
19	[79]	2017	✗	This paper proposed an algorithm to generate summary of multi-video, they have not focused on MVS.
20	[28]	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
21	[80]	2017	✗	[81] is the extended version of [80], and both techniques are focused only on SVS, the datasets used for evaluation are SVS datasets.
23	[81]	2018	✗	

The year-wise trend of MVS is shown in Fig. **3**, covering the overall literature. Initial researches in MVS use low-level features (SIFT descriptors) and object detection based on handcrafted features with activity-based video segmentation. These initiative articles utilized clustering (i.e., K-means clustering) techniques for final summary generation. Next trend in literature is also bent towards low-level features with some improvements such as background subtraction for trajectories extraction and the use of machine learning techniques such as support vector machine (SVM) and K-means clustering for final summary generation. The final summaries generated in this trend 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 with handcrafted features (i.e., color-, edge-histograms, Tamura, and SIFT features).

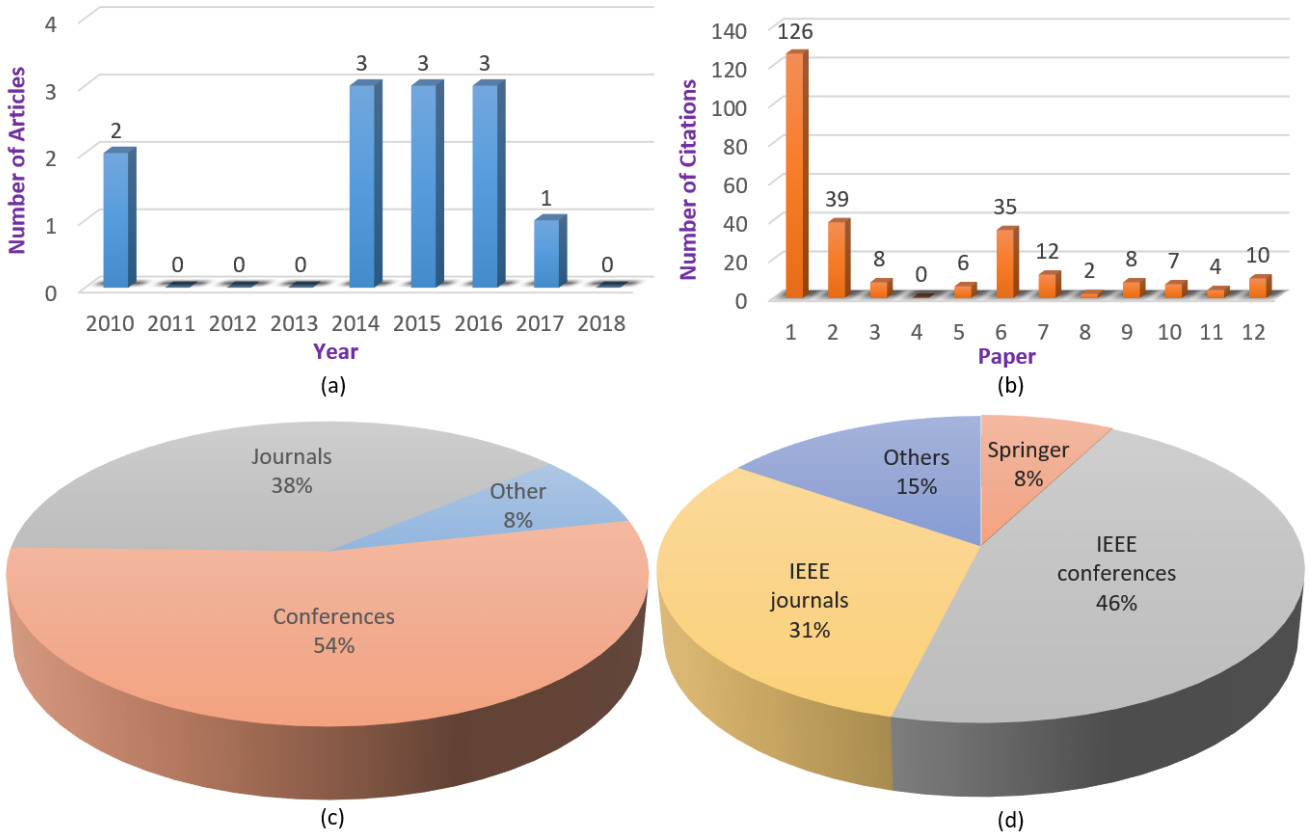

Fig. 2: Overall distribution of MVS literature (a) year-wise MVS publications till date (b) citation wise distribution of MVS research papers (c) distribution of MVS on the basis of research paper's type, and (d) publisher-wise distribution of MVS research paper

TABLE II
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	[82]	-	-	✓	-	-	✓
	[83]	✓	-	-	✓	-	-
Sports	[71]	-	✓	-	✓	-	-
	[84]	-	-	-	✓	-	-
Surveillance	[85]	✓	-	-	✓	-	-
	[77]	-	-	-	-	-	✓
	[28]	-	-	✓	✓	-	✓
	[86]	-	-	✓	✓	-	-
	[26]	-	✓	-	-	-	✓
	[23]	-	✓	-	-	-	✓
	[83]	✓	-	-	✓	-	-
	[65]	✓	-	-	✓	-	-
	[27]	✓	-	-	-	-	✓
	[69]	✓	-	-	✓	-	-

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 generating summaries by utilizing deep features in prerequisite steps. This trend is followed in 2016 where BVLC CaffeNet [87] 4096-dim and Spatio temporal C3D [88] features are used for sparse coding for the first time in literature, and

video representation, respectively. Besides clustering for summary generation, template matching and sparse representative selection over learned embedding are used for generating final summaries. Likewise, the previous trend, article published in 2017 uses C3D features for video representation. In this trend, inter and intra-view similarities are computed through sparse coefficients and final summary is generated using a clustering scheme.

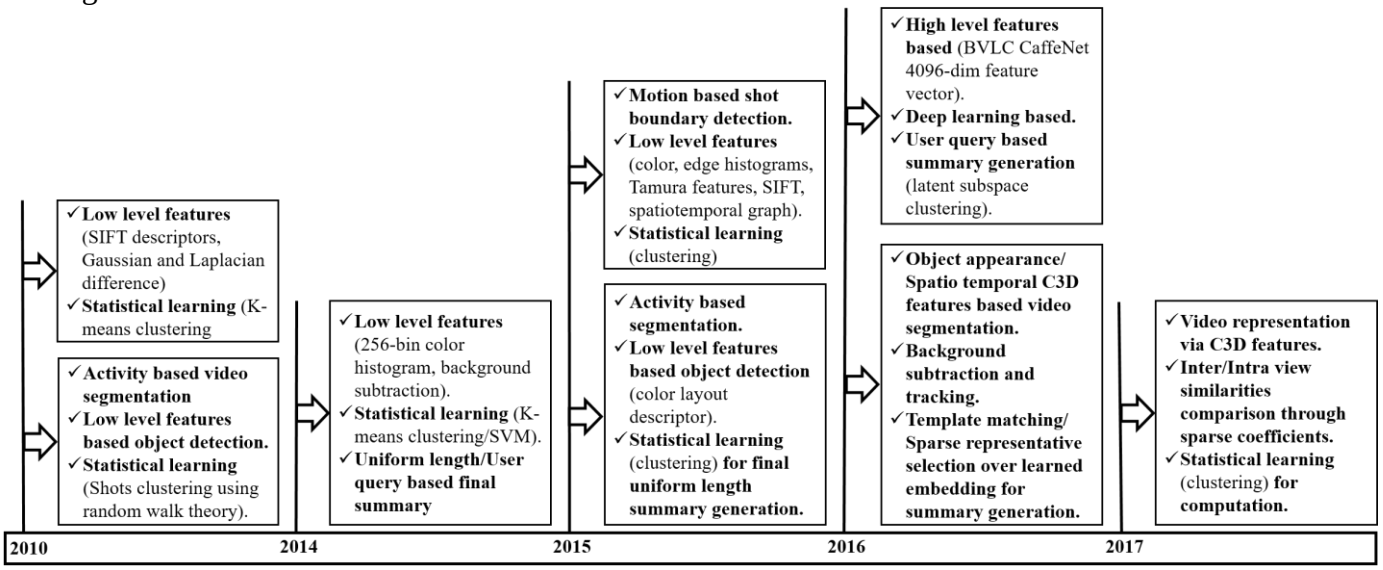

Fig. 3: Trend of MVS methods (year wise features distribution and learning mechanisms of representative articles from the literature)

III. Methods for Summarization of Multi-View Videos

In literature various techniques have been proposed for MVS, and each method follows a generic route with specific steps. The basic flow of MVS contains three steps: segmentation of multi-view videos, features extraction, and summary generation. Processing whole video and generating summary at once is a biased decision, as 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 to generate final summary. In the final step of summary generation, some schemes or validation criteria for keyframes selection are applied. These steps are elaborated in the subsequent sections individually and are comprehensively shown in Table **III** and **IV**.

A. Multi-View Videos Segmentation

Segmentation of videos into parts is a preprocessing step for summary generation. Multi-view summary generated subsequently after segmentation of videos into several parts, makes it more diverse and representative of all videos. In literature many techniques are based on segmentation of videos into shots, which are further processed for summary generation. In some techniques shot segmentation is uniform length [83] with specific threshold for single shot selection, while in most of the techniques it is based on features [26, 69] with non-uniform length. Segmentation of video into shots helps in removing redundancy such as suppressing the frames with no activity or motion and choosing the frames with certain events or motion. 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 **III**, the third column shows the methods using shot segmentation with a sub-category of whether shots are segmented with uniform length or features based. The first column of Table **IV** gives a detailed description about features or parameters used by the system for segmentation. In the current, literature video segmentation can be divided into activities based [26, 83], motion [23], objects [77], and deep high-level features based methods [27, 28] as explained in Table **IV**.

TABLE III

DESCRIPTION OF USED FEATURES, SHOT SEGMENTATION TECHNIQUE, LEARNING MECHANISM, AND SUMMARY GENERATION OF EXISTING MVS METHODS

Method name/year	Visual features used				Shot segmentation		Learning		Summary generation		
	Low level feature	Saliency and motion	High level feature	Object detection	Uniform length	Features based	Statistical	Deep learning	Uniformity	Importance	User query
[26] (2010)	✓	-	✓	✓	-	✓	✓	-	-	✓	-
[89] (2010)	✓	-	-	-	-	-	✓	-	✓	-	-
[71] (2014)	-	-	-	-	-	-	✓	-	-	-	✓
[69] (2014)	✓	-	-	-	-	✓	✓	-	✓	-	-
[86] (2014)	-	-	-	✓	-	-	✓	-	-	✓	-
[65] (2014)	✓	-	-	-	-	-	✓	-	✓	-	-
[82] (2015)	-	-	-	✓	-	-	✓	-	✓	-	-
[23] (2015)	✓	✓	-	-	-	✓	✓	-	-	✓	-
[83] (2015)	✓	-	-	✓	✓	-	✓	-	-	✓	-
[85] (2016)	-	-	✓	-	-	-	-	✓	-	-	✓
[77] (2016)	-	✓	-	-	-	-	-	-	✓	-	-
[27] (2016)	✓	-	-	-	-	-	-	✓	-	✓	-
[28] (2017)	-	-	✓	-	-	✓	-	✓	-	✓	-

B. Features Extraction

The next prerequisite step for summary generation is features extraction which includes object detection or tracking [83, 86]. The first column in Table **III** shows the features extracted by various methods in literature. Visual features are divided into four sub-categories 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 second column of Table **IV**. Handcrafted features such as histogram (256 bin) features in HSV color space [69], color layout descriptor [83], SIFT features [23, 90], and some other techniques such as background subtraction [86], foreground object estimation [83], and human detection [82] are utilized for summary generation. R. Panda et al. [85] used high-level features extracted from "BVLC CaffeNet" pre-trained model for finding inter and intra-view correlations in embedding space.

C. Keyframes Selection/ Summary Generation

The final step involved in MVS pipeline is summary generation based on features extracted in the previous section. Final generated summary can be divided into further categories such as uniform length summary generation [89], importance-based [26], and user query based [85] as represented in the last column of Table **III**. Mainstream methods use statistical learning for generating summaries with different features parameters. Clustering with random walks, K-means clustering, unsupervised optimum path, SVM, template matching, and subspace clustering are used for summary generation. A description of summary generation methods is given in Table **IV** along with references.

IV. MULTI-VIEW SUMMARIZATION DATASETS

A total of eleven 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 [26], and BL-7F [83] that are covered in this section. The complete detail and description about all the datasets are given in Table V and their distribution is presented in Fig. **4**.

TABLE IV

DETAILED DESCRIPTION OF THREE MAIN STEPS FOLLOWED BY EACH MVS METHOD IN LITERATURE

Method	Segmentation	Feature extraction /object detection or tracking	Summary generation
[90] (2010)	-	Based on difference of Gaussian and Laplacian Gaussian and then computing SIFT descriptor, adopted from [89]	Clustering of SIFT descriptors into 500 groups by K-means to acquire visual vocabulary
[26] (2010)	Adopted activity-based video segmentation	Gaussian entropy fusion model, wavelet coefficients. Viola-Jones face detector to construct Spatio-temporal shot graph	Shots clustering using random walk theory
[71] (2014)	-	Quality-Of-View (QOV). i.e., The QOV is calculated from the distance between a camera and each object and the angle between them.	SVM whose kernel is RBF (Radial Basis Function) as the learning model
[69] (2014)	-	256-bin color histogram in HSV color space	k-means clustering for keyframes selection
[86] (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
[82] (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)
[23] (2015)	Motion based shot boundary detection	Color histograms, edge histograms, Tamura features, and SIFT spatiotemporal graph	Unsupervised Optimum-Path Forest (clustering)
[83] (2015)	-	MPEG-7 color layout descriptor, and score estimation of foreground object	Gaussian mixture model (clustering)
[65] (2015)	-	Low-level features	RBF kernel function for similarity, maximal-margin clustering.
[85] (2016)	-	BVLC CaffeNet (4096-dim CNN feature vector), Sparse coding	Latent subspace clustering
[77] (2016)	Object appearance-based segmentation	Adaptive Scale Invariant Local Ternary Pattern (SILTP) for background subtraction and tacking object.	Based on template matching algorithm
[27] (2016)	Video segmentation via Spatio-temporal C3D [88] 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.
[28] (2017)	Video representation via Spatio-temporal C3D [88] features	Inter- and intra-view similarities are computed via sparse coefficients.	Clustering for computation of data similarity

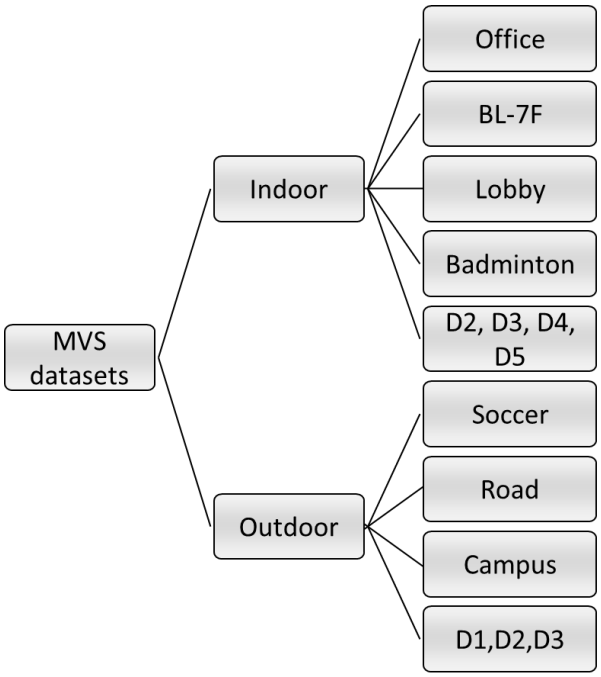

Fig. 4: Categorization of standard MVS datasets on the basis of recorded environment. Indoor recorded datasets include Office, BL-7F, Lobby, and Badminton. Outdoor datasets are Soccer, Road, Campus, and D1. D2 and D3 dataset include both indoor and outdoor scenarios

Table **V** 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 last column of Table V shows the papers that used the corresponding dataset for experiments. The popular datasets are discussed individually in next sub-sections.

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 Bl-7F four views, and (d) sample frames from handheld camera of campus dataset

A. Office Dataset [26]

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

B. Lobby Dataset [26]

Lobby dataset is provided by the same authors of Office dataset, captured by three cameras in a lobby area. All cameras in Lobby dataset are synchronized with each other, still, and non-fixed. There are much-crowded

scenes in this dataset which makes it more challenging for summarization. Some sample images are shown in Fig. **5** (b).

C. Bl-7F [83]

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

D. Campus [26]

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 which makes it unstable and obscure. Sample frames of campus dataset are visualized in Fig. **5** (d). The videos of campus dataset are challenging because they are not synchronized and contain motion of cameras.

V. Challenges in MVS Datasets

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

A. Lack of Synchronization

Mainstream videos in MVS datasets of different views are not synchronized with each other. For instance, the videos of Office dataset [26] are highly un-synchronized. Soccer [71] dataset videos are also not synchronized during recording in the field, but later, they are synchronized manually. The problem of synchronization makes the summary generation a difficult task because it becomes challenging to compute the inter-view correlations among different unsynchronized videos. To tackle this problem, [83] manually aligned these videos and then performed experiments for summary generation. Besides this, some other techniques such as [28] intelligently find correlations through various features matching algorithms for summary generation.

B. Instability of Camera

The cameras in some of the MVS datasets such as Bl-7F [83] are stable and fixed with no shuddering's, whereas in Lobby, Office [26] 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. [78] used BVLC CaffeNet [87] pre-trained model, and extracted 4096-dim CNN feature vector. This feature vector is acquired from top layer hidden unit activations of the 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.

C. Crowded Scenes

In MVS literature, Lobby [26] and Soccer [71] datasets are very crowded and contain richer activities as compared to Office [26] and Bl-7F [83] 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 [26] dataset is the most challenging problem, while Office [26] and Bl-7F [83] are the easiest datasets in MVS literature in terms of crowd density. Besides these challenges, Office [26] dataset also contains unstable frame rate challenge, and it suffers from highly variable light variations. The problem of a crowded scene is tackled in [23] by filtering the activities, thus the scenes with low activity are suppressed through Gaussian entropy.

VI. Recommendations and Future Research Directions

It has been observed that MVS problem is not addressed according to the current need of MVS for a vast amount of applications. It is apparent from the reviewed literature that till date, most of the research is based on handcrafted features or mid-level features. Similarly, various clustering techniques, 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:

TABLE V
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 [91]	2004	Fixed	Outdoor	-	2391	-	-	✓	-	Human actions	[82]
WEIZMANN [92]	2005	-	-	-	9	-	-	✓	-	-	[82]
PETS [93]	2009	-	-	-	85	-	-	✓	-	-	[82]
Looby [26]	2010	Fixed	3/ Indoor	✓	3	24 m 42s	✓	-	-	-	[26],[83], [85], [28], [27],[69]
Office [26]	2010	Fixed	4/ Indoor	✓	4	14 m 58s	✓	-	-	-	[26], [83], [85], [65], [28], [27], [69]
Campus [26]	2010	Fixed	4/ Outdoor	✓	-	56 m 43 s	✓	-	-	-	[26], [85], [28], [27]
[90]	2010	Internet website	-	-	88 sets	-	✓	-	-	Diverse videos	[90]
[94]	2012	Fixed	3	-		2 hours	✓	-	-	Human trajectory	[77]
Soccer [71]	2014	Fixed	20/ Outdoor	-	20	-	-	-	✓	Football match	[71]
D1 [86]	2014	Fixed	2/ Outdoor/ Indoor	-	-	6m 40s		✓		Pedestrian activity	[86]
D2 [86]	2014	Fixed	2 Outdoor/ Indoor	-	-	10m 43s	-	-	✓	Vehicle surveillance	[77], [86]
D3 [86]	2014	Fixed	2/ Outdoor/ Indoor	-	-	4m 45s	-	-	✓	Vehicle surveillance of night scenario	[86]
D4 [86]	2014	Fixed	3/ Outdoor	-	-	3m 26s	-	-	✓	Vehicle surveillance of street scenario	[86]
D5 [86]	2014	Fixed	3/ Outdoor	-	-	3m 31s	✓	-	-	Vehicle surveillance of playground scenario	[86]
[71]	2014	Fixed	20/ Outdoor	-	20	-	✓	-	-	Soccer videos	-
BL-7F [83]	2015	Fixed	19/ Indoor	✓	19	7 m 10s	✓	-	-	-	[83], [28], [69]

A. *End-to-End Deep Learning Models*

Although there are some techniques in literature which pick up learned features for the summary generation as a prerequisite step, yet there are no such deep learning models that can input multi-view videos and directly generate their output summary. MVS literature lacks such CNN architectures that achieve the task of searching correlations between different views and find salient frames intelligently. End-to-end deep learning

models are used for various purposes, mainly for speech recognition [95-97]. Thus, in future work, it is highly recommended to propose such end-to-end deep learning models that can provide a summary for MV videos with satisfactory accuracy while preserving the properties of a good summary.

B. Standard Datasets for Benchmarking

The currently available datasets of literature address many challenges such as lack of synchronization, variable lighting conditions, unstable frame rate, 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 environment for indoor or outdoor. In future work, researchers who are enthusiastic to work in MVS should focus on creating standard datasets which cover all sort of environments including normal and abnormal or uncertain (fire, fog, snow, rainfall etc.)

C. Agents Based MVS

In MVS literature, till date 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 from processing a single video but MVS in contrast processes multi videos to generate an output which makes the output generation steady. Therefore, if the task of MVS is divided into different modules between different 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 [98] for motive profiling of users in virtual worlds and gaming and [99] for energy minimization in cloud computing systems. Thus, researchers in future should try dividing the load of MVS into several agents and integrating their individual outputs to finally generate a summary.

D. Fog/Cloud Computing for MVS

Fog computing [100-102] is a layer of distributed network that can work much faster than a single computer. It is used in different computer vision applications such as healthcare [103-105] security systems [106, 107], web applications [108], video analytics [109], forecasting systems [110], and image retrieval [111] problems. So far every state-of-the-art technique generates summary locally, but processing these videos through Cloud [112] or Fog computing will make the output generation very faster.

E. Edge Intelligence for MVS

Edge intelligence [113-116] refers to processing, development, and analysis of data at the site where the data is generated. It can be used for many sensors such as a visual sensor to analyze the scenes captured in real time or to detect abnormal activities etc. For future research in MVS, embedded programs can be used in high definition cameras for MVS generation at the site independently. Therefore, instead of sending all the videos on wireless or local networks, it is better to send only generated summary which can assist analyzing video in short period of time and saving bandwidth and the precious time of surveillance analysts [117, 118].

F. Processing Time for MVS

The experimental results provided by the current state-of-the-art MVS methods perform evaluation using the only accuracy with no focus on their running time and computational feasibility for deployment in real-world surveillance networks. MVS literature lacks the assessment of running time of algorithms provided with their system configuration. 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.

G. MVS for Resource Constrained Devices

Future research in MVS should also aim to generate automated summaries through resource constrained devices such as Raspberry pi, FPGA, and Adriano etc. These resource constrained devices have many capacities that can be utilized for several applications [104, 119]. Generating summaries through such devices

has the advantage of attaching these devices with a small network of camera which will work independently for output creation.

VII. Conclusion

We have presented a complete survey of state-of-art techniques for MVS. MVS has become an interesting topic with recent gigantic increase in multi-view cameras and surveillance networks installed at public and private places. Because there exists a countable number of articles in MVS literature. Thus, we investigated each paper with several aspects. In this survey, we presented how the MVS trend is followed from the very beginning of MVS literature in 2010. Furthermore, we described the generic working of MVS methods with each step explained individually and referenced tables correspondingly. Next, we made an overview of the datasets used for MVS by different techniques. Finally, summarizing the whole literature we provided recommendations and future directions for further work in MVS.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2016R1A2B4011712).

References

[1] H. Fradi and J.-L. Dugelay, "Towards crowd density-aware video surveillance applications," *Information Fusion*, vol. 24, pp. 3-15, 2015/07/01/ 2015.
[2] S. Rho, W. Rahayu, and U. T. Nguyen, "Intelligent video surveillance in crowded scenes," *Information Fusion*, vol. 24, pp. 1-2, 2015/07/01/ 2015.
[3] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised Feature Selection via Spline Regression for Video Semantic Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 252-264, 2015.
[4] Z. Gheid, Y. Challal, X. Yi, and A. Derhab, "Efficient and privacy-aware multi-party classification protocol for human activity recognition," *Journal of Network and Computer Applications*, vol. 98, pp. 84-96, 2017.
[5] F. F. Chamasemani, L. S. Affendey, N. Mustapha, and F. Khalid, "Speeded up surveillance video indexing and retrieval using abstraction," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2017, pp. 374-378.
[6] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad, "On influential trends in interactive video retrieval: Video Browser Showdown 2015-2017," *IEEE Transactions on Multimedia*, pp. 1-1, 2018.
[7] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, pp. 489-504, 2009.
[8] Y.-J. Yu, A.-C. Pang, and M.-Y. Yeh, "Video encoding adaptation for QoE maximization over 5G cellular networks," *Journal of Network and Computer Applications*, vol. 114, pp. 98-107, 2018.
[9] L. Zhang, L. Sun, W. Wang, and Y. Tian, "KaaS: A Standard Framework Proposal on Video Skimming," *IEEE Internet Computing*, vol. 20, pp. 54-59, 2016.
[10] V. K. Vivekraj, D. Sen, and R. Balasubramanian, "Vector ordering based multimodal video skimming for user videos," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2017, pp. 775-780.
[11] M. Yu-Fei and Z. Hong-Jiang, "A model of motion attention for video skimming," in *Proceedings. International Conference on Image Processing*, 2002, pp. I-I.
[12] H. Nam and C. D. Yoo, "Content adaptive video summarization using spatio-temporal features," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4003-4007.
[13] A. H. Meghdadi and P. Irani, "Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 2119-2128, 2013.
[14] S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Online content-aware video condensation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2082-2087.
[15] J. Zhu, S. Feng, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "High-Performance Video Condensation System," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, pp. 1113-1124, 2015.
[16] S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Online content-aware video condensation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2082-2087.
[17] "http://encyclopedia.jrank.org/articles/pages/6930/Video-Summarization.html (Cited on 05 Sep, 2018)."
[18] G. Yihong and L. Xin, "Video shot segmentation and classification," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, pp. 860-863 vol.1.
[19] L. Ranathunga, R. Zainuddin, and N. A. Abdullah, "Conventional video shot segmentation to semantic shot segmentation," in *2011 6th International Conference on Industrial and Information Systems*, 2011, pp. 186-191.
[20] C. K. Mohan, N. Dhananjaya, and B. Yegnanarayana, "Video Shot Segmentation Using Late Fusion Technique," in *2008 Seventh International Conference on Machine Learning and Applications*, 2008, pp. 267-270.
[21] W. Zhang, Y. Wang, and X. Jiang, "A shot segmentation algorithm for H.264 compressed videos," in *2013 6th International Congress on Image and Signal Processing (CISP)*, 2013, pp. 81-85.
[22] F. Husain, B. Dellen, and C. Torras, "Consistent Depth Video Segmentation Using Adaptive Surface Models," *IEEE Transactions on Cybernetics*, vol. 45, pp. 266-278, 2015.
[23] S. K. Kuanar, K. B. Ranga, and A. S. Chowdhury, "Multi-view video summarization using bipartite matching constrained optimum-path forest clustering," *IEEE Transactions on Multimedia*, vol. 17, pp. 1166-1173, 2015.
[24] G. L. P. G and S. D, "Walsh–Hadamard Transform Kernel-Based Feature Vector for Shot Boundary Detection," *IEEE Transactions on Image Processing*, vol. 23, pp. 5187-5197, 2014.
[25] C. Bi, Y. Yuan, J. Zhang, Y. Shi, Y. Xiang, Y. Wang, *et al.*, "Dynamic Mode Decomposition Based Video Shot Detection," *IEEE Access*, vol. 6, pp. 21397-21407, 2018.
[26] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, pp. 717-729, 2010.

[27] R. Panda, A. Das, and 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.

[28] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *arXiv preprint arXiv:1706.03121*, 2017.

[29] C. Benedek, B. Gálai, B. Nagy, and Z. Jankó, "Lidar-Based Gait Analysis and Activity Recognition in a 4D Surveillance System," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 101-113, 2018.

[30] K.-P. Chou, M. Prasad, D.-L. Li, N. Bharill, Y.-F. Lin, F. Hussain, *et al.*, "Automatic Multi-view Action Recognition with Robust Features," Cham, 2017, pp. 554-563.

[31] A. Ö and M. E. Karsligil, "Video-based traffic accident analysis at intersections using partial vehicle trajectories," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 4693-4696.

[32] Y. Ki and D. Lee, "A Traffic Accident Recording and Reporting Model at Intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, pp. 188-194, 2007.

[33] S. Ramagiri, R. Kavi, and V. Kulathumani, "Real-time multi-view human action recognition using a wireless camera network," in *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 2011, pp. 1-6.

[34] M. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group Level Activity Recognition in Crowded Environments across Multiple Cameras," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 56-63.

[35] "https://techcrunch.com/2008/09/23/hondas-new-multi-view-camera-system-makes-driving-safer/ (Accessed on 03 Sep, 2018)."

[36] "http://www.thermoteknix.com/products/cement/multi-view-thermascope-hd-software/ (Accessed on 03 Sep, 2018)."

[37] "https://www.softwareadvice.com/accounting/multiview-enterprise-profile/ (Accessed on 03, Sep, 2018)."

[38] J. Satake and J. Miura, "Stereo-Based Multi-person Tracking Using Overlapping Silhouette Templates," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 4304-4307.

[39] H. Yongkai, H. Liusheng, X. Hongli, and X. Ben, "Cluster-based location report for person tracking in wireless sensor networks," in *2008 11th IEEE International Conference on Communication Technology*, 2008, pp. 105-108.

[40] X. Luo, R. T. Tan, and R. C. Veltkamp, "Multi-person tracking based on vertical reference lines and dynamic visibility analysis," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 1877-1880.

[41] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30-42, 2018.

[42] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional Neural Networks Based Fire Detection in Surveillance Videos," *IEEE Access*, vol. 6, pp. 18174-18183, 2018.

[43] A. Iosifidis, A. Tefas, and I. Pitas, "View-Invariant Action Recognition Based on Artificial Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 412-424, 2012.

[44] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4894-4903.

[45] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, pp. 38-49, 2018.

[46] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 715-731.

[47] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11-26, 2017.

[48] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, *et al.*, "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognition Letters*, 2018.

[49] D. Peralta, I. Triguero, S. García, Y. Saeys, J. M. Benitez, and F. Herrera, "On the use of convolutional neural networks for robust classification of multiple fingerprint captures," *International Journal of Intelligent Systems*, vol. 33, pp. 213-230, 2018.

[50] R. Olmos, S. Tabik, and F. Herrera, "Automatic handgun detection alarm in videos using deep learning," *Neurocomputing*, vol. 275, pp. 66-72, 2018/01/31/ 2018.

[51] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. Baik, *Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features* vol. PP, 2017.

[52] G. Varol, I. Laptev, and C. Schmid, "Long-Term Temporal Convolutions for Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1510-1517, 2018.

[53] A. Gómez-Ríos, S. Tabik, J. Luengo, A. Shihavuddin, B. Krawczyk, and F. Herrera, "Towards Highly Accurate Coral Texture Images Classification Using Deep Convolutional Neural Networks and Data Augmentation," *arXiv preprint arXiv:1804.00516*, 2018.

[54] P. Zeng, H. Li, H. He, and S. Li, "Dynamic Energy Management of a Microgrid using Approximate Dynamic Programming and Deep Recurrent Neural Network Learning," *IEEE Transactions on Smart Grid*, pp. 1-1, 2018.

[55] H. Jiang, X. Liu, H. He, C. Yuan, and D. Prokhorov, "Neural Network Based Distributed Consensus Control for Heterogeneous Multi-agent Systems," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 5175-5180.

[56] Y. Zhou and G. G. Yen, "Evolving Deep Neural Networks for Movie Box-Office Revenues Prediction," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1-8.

[57] Y. Sun, G. G. Yen, and Z. Yi, "Evolving Unsupervised Deep Neural Networks for Learning Meaningful Representations," *IEEE Transactions on Evolutionary Computation*, pp. 1-1, 2018.

[58] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale Convolutional Neural Networks for Fault Diagnosis of Wind Turbine Gearbox," *IEEE Transactions on Industrial Electronics*, pp. 1-1, 2018.

[59] C.-F. Juang, Y.-C. Chang, and I. Chung, "Evolutionary hexapod robot gait control using a new recurrent neural network learned through group-based hybrid metaheuristic algorithm," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 2018, pp. 111-112.

[60] C. Juang and C. Chen, "An Interval Type-2 Neural Fuzzy Chip With On-Chip Incremental Learning Ability for Time-Varying Data Sequence Prediction and System Control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 216-228, 2014.

[61] C. Juang and C. Chen, "Data-Driven Interval Type-2 Neural Fuzzy System With High Learning Accuracy and Improved Model Interpretability," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1781-1795, 2013.

[62] C. Juang, W. Chen, and C. Liang, "Speedup of Learning in Interval Type-2 Neural Fuzzy Systems Through Graphic Processing Units," *IEEE Transactions on Fuzzy Systems*, vol. 23, pp. 1286-1298, 2015.

[63] Y. Zhuang, R. Xiao, and F. Wu, "Key issues in video summarization and its application," pp. 448-452.

[64] Y. Li and B. Merialdo, "Multi-video summarization based on Video-MMR," pp. 1-4.

[65] Y. Fu, L. Wang, and Y. Guo, "Multi-view metric learning for multi-view video summarization," *arXiv preprint arXiv:1405.6434*, 2014.

[66] L. Xu and D. Schuurmans, "Unsupervised and semi-supervised multi-class support vector machines," in *AAAI*, 2005, p. 13.

[67] B. Long, P. S. Yu, and 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.

[68] M. Y. Zhang and W. Y. Cai, "Multi-view Video Summarization Algorithm for WMSN," in *2014 International Conference on Wireless Communication and Sensor Network*, 2014, pp. 213-216.
[69] 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.
[70] A. I. Ioannidis, V. T. Chasanis, and 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.
[71] Y. Muramatsu, T. Hirayama, and 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.
[72] S. H. Ou, C. H. Lee, V. S. Somayazulu, Y. K. Chen, and S. Y. Chien, "On-Line Multi-View Video Summarization for Wireless Video Sensor Network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 165-179, 2015.
[73] Y. Hoshen and S. Peleg, "Live video synopsis for multiple cameras," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 212-216.
[74] A. Mahapatra, P. K. Sa, and B. Majhi, "A multi-view video synopsis framework," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1260-1264.
[75] L. Nie, R. Hong, L. Zhang, Y. Xia, D. Tao, and N. Sebe, "Perceptual Attributes Optimization for Multivideo Summarization," *IEEE Transactions on Cybernetics*, vol. 46, pp. 2991-3003, 2016.
[76] R. Panda, A. Das, and 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.
[77] J. Zhu, S. Liao, and S. Z. Li, "Multicamera joint video synopsis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, pp. 1058-1069, 2016.
[78] R. Panda, A. Dasy, and A. K. Roy-Chowdhury, "Video summarization in a multi-view camera network," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2971-2976.
[79] R. Panda, N. C. Mithun, and A. K. Roy-Chowdhury, "Diversity-Aware Multi-Video Summarization," *IEEE Transactions on Image Processing*, vol. 26, pp. 4712-4724, 2017.
[80] J. Meng, S. Wang, H. Wang, Y. P. Tan, and J. Yuan, "Video Summarization via Multi-view Representative Selection," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1189-1198.
[81] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multi-view representative selection," *IEEE Trans. on Image Processing*, pp. 2134-2145, 2018.
[82] A. Mahapatra, P. K. Sa, and B. Majhi, "A multi-view video synopsis framework," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 1260-1264.
[83] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 165-179, 2015.
[84] F. Daniyal and A. Cavallaro, "Multi-camera scheduling for video production," in *Visual Media Production (CVMP), 2011 Conference for*, 2011, pp. 11-20.
[85] R. Panda, A. Dasy, and 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.
[86] X. Zhu, J. Liu, J. Wang, and H. Lu, "Key observation selection-based effective video synopsis for camera network," *Machine vision and applications*, vol. 25, pp. 145-157, 2014.
[87] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, *et al.*, "Caffe: Convolutional architecture for fast feature embedding," pp. 675-678.
[88] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.
[89] Y.-G. Jiang and C.-W. Ngo, "Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval," *Computer Vision and Image Understanding*, vol. 113, pp. 405-414, 2009.
[90] Y. Li and 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.
[91] C. Schuldt, I. Laptev, and 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.
[92] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2005, pp. 1395-1402.
[93] J. Ferryman and 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.
[94] X. Chen, K. Huang, and T. Tan, "Learning the three factors of a non-overlapping multi-camera network topology," in *Chinese Conference on Pattern Recognition*, 2012, pp. 104-112.
[95] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," pp. 173-182.
[96] Y. Miao, M. Gowayyed, and 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.
[97] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
[98] L. Xuejie, K. Merrick, and 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.
[99] W. Wang, Y. Jiang, and W. Wu, "Multiagent-Based Resource Allocation for Energy Minimization in Cloud Computing Systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, pp. 205-220, 2017.
[100] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," in *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, 2015, pp. 73-78.
[101] S. Yi, C. Li, and Q. Li, "A survey of fog computing: concepts, applications and issues," in *Proceedings of the 2015 workshop on mobile big data*, 2015, pp. 37-42.
[102] 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*, vol. 6, pp. 1706-1717, 2017.
[103] C. Thota, R. Sundarasekar, G. Manogaran, R. Varatharajan, and M. Priyan, "Centralized fog computing security platform for IoT and cloud in healthcare system," in *Exploring the convergence of big data and the internet of things*, ed: IGI Global, 2018, pp. 141-154.
[104] 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," *Multimedia Tools and Applications*, vol. 76, pp. 3519-3536, 2017.
[105] I. Mehmood, M. Sajjad, and S. W. Baik, "Mobile-cloud assisted video summarization framework for efficient management of remote sensing data generated by wireless capsule sensors," *Sensors*, vol. 14, pp. 17112-17145, 2014.

[106] S. Alharbi, P. Rodriguez, R. Maharaja, P. Iyer, N. Bose, and Z. Ye, "FOCUS: A fog computing-based security system for the Internet of Things," in *Consumer Communications & Networking Conference (CCNC), 2018 15th IEEE Annual*, 2018, pp. 1-5.
[107] A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog computing for the internet of things: Security and privacy issues," *IEEE Internet Computing*, vol. 21, pp. 34-42, 2017.
[108] T. Wang, W. Zhang, C. Ye, J. Wei, H. Zhong, and T. Huang, "FD4C: Automatic Fault Diagnosis Framework for Web Applications in Cloud Computing," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, pp. 61-75, 2016.
[109] M. U. Yaseen, A. Anjum, O. Rana, and N. Antonopoulos, "Deep Learning Hyper-Parameter Optimization for Video Analytics in Clouds," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1-12, 2018.
[110] F. J. Baldan, S. Ramirez-Gallego, C. Bergmeir, F. Herrera, and J. M. Benitez-Sanchez, "A Forecasting Methodology for Workload Forecasting in Cloud Systems," *IEEE Transactions on Cloud Computing*, pp. 1-1, 2018.
[111] N. Rahim, J. Ahmad, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Privacy-preserving image retrieval for mobile devices with deep features on the cloud," *Computer Communications*, vol. 127, pp. 75-85, 2018/09/01/ 2018.
[112] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on*, 2009, pp. 44-51.
[113] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," *ETSI white paper*, vol. 11, pp. 1-16, 2015.
[114] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," pp. 48-55.
[115] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges," *IEEE Communications Magazine*, vol. 55, pp. 54-61, 2017.
[116] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. of the Sixth International Conference on Advances in Future Internet*, 2014, pp. 48-55.
[117] X. Liu and S. Ramakrishnan, "Efficient available bandwidth usage in transmission of compressed video data," ed: Google Patents, 2007.
[118] J. Wu, B. Cheng, M. Wang, and J. Chen, "Energy-Efficient Bandwidth Aggregation for Delay-Constrained Video Over Heterogeneous Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 30-49, 2017.
[119] Y.-H. Kao, B. Krishnamachari, M.-R. Ra, and F. Bai, "Hermes: Latency optimal task assignment for resource-constrained mobile computing," *IEEE Transactions on Mobile Computing*, vol. 16, pp. 3056-3069, 2017.