IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Video Processing using Deep learning Techniques: A Systematic Literature Review

Vijeta Sharma[1,2], Manjari Gupta[2], Ajai Kumar[1] and Deepti Mishra[3]
[1]Centre for Development of Advanced Computing (C-DAC), Pune-411008, India
[2]Computer Science, DST- Center for Interdisciplinary Mathematical Sciences, Institute of Science, Banaras Hindu University, Varanasi-221005, India
[3] Department of Computer Science (IDI), NTNU - Norwegian University of Science and Technology, Gjøvik, Norway

Corresponding author: Deepti Mishra (email: deepti.mishra@ntnu.no), Manjari Gupta (e-mail: manjari@bhu.ac.in).

ABSTRACT Studies show lots of advanced research on various data types such as image, speech, and text using deep learning techniques, but nowadays, research on video processing is also an emerging field of computer vision. Several surveys are present on video processing using computer vision deep learning techniques, targeting specific functionality such as anomaly detection, crowd analysis, activity monitoring, etc. However, a combined study is still unexplored. This paper aims to present a Systematic Literature Review (SLR) on video processing using deep learning to investigate the applications, functionalities, techniques, datasets, issues, and challenges by formulating the relevant research questions (RQs). This systematic mapping includes 93 research articles from reputed databases published between 2011 and 2020. We categorize the deep learning technique for video processing as CNN, DNN, and RNN based. We observe the significant advancements in video processing between 2017 and 2020, primarily due to the advent of AlexNet, ResNet, and LSTM based deep learning techniques. The prominent fields of video processing research are observed as human action recognition, crowd anomaly detection, and behavior analysis. This SLR is a helpful guide for the researchers to explore the recent literature, available datasets, and existing deep learning techniques for video processing.

INDEX TERMS Video processing; Computer vision; Artificial intelligence; Deep learning; Human action recognition; Systematic literature review

I. INTRODUCTION

Many deep learning (DL) research works have shown successful results, primarily focusing on three data types: images, speech, and text. In addition, DL has also been successfully applied to communication signals/packets, e.g. [1], [2]. Widely used applications of these data domains are image classification, speech recognition, regression problem, pattern recognition, and text sentiment classification. Apart from these, one more fascinating data modality is video data. However, video data is also interesting for research from the perspective of its big size and dimension. Millions of video data are uploaded every day on YouTube; thus, it became a rich repository and empowered artificial intelligence (AI) research. However, video data is challenging to analyze and process because of its large file sizes and complexity despite having rich data. Research on video processing using AI gained popularity after many AI algorithms were developed for Image processing for various applications, particularly in the past ten years.

Video data is one of the popular choices of users of different platforms like Twitter, YouTube, Facebook, etc. also the fastest-growing data type nowadays.

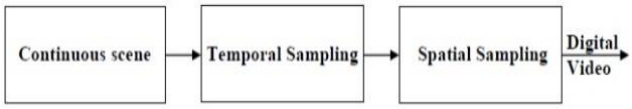

FIGURE 1. A concept of video data

We first clarify the exact meaning of video data in the computer-vision research field, which is considered in our study. Video data (or digital video data) is any sequence of time-varying images. In the video data, the picture information is digitized both spatially and temporally. The resultant pixel intensities are quantized. We can say a set of frames per second.

Figure 1 depicts the concept of video data. Most of the fundamental research of computer vision today focuses on images, focusing less on sequences of images, i.e., video

frames. However, video data provides deeper situational understanding because a series of images gives various information about the subject. For example, we can track an object through an optical flow of the sequence of images and predict its next action[3].

After explaining the abovementioned meaning of video data, we show the interpretation of video processing we considered in our study. In the context of computer vision, video processing or digital video processing is the ability to automatically analyze video, frame by frame, to detect and determine temporal and spatial features.

A. LITERATURE SURVEY

While many types of research have been carried out on video processing [4], very few studies have been systematically analyzed that focus on video processing using deep learning techniques. Instead, most of them perform surveys by targeting only specific functionality. However, a study conducted by Nayak et al. [5] shows the advancement in video anomaly detection using deep learning techniques. The authors present the various deep learning techniques for video processing to detect the anomalies such as abnormal activities- fighting, riots, traffic rule violations, stampede, and strange entities - weapons, abandoned luggage, etc. In another survey [6], researchers reported video processing for abnormal human activity recognition by leveraging the deep learning method for video processing. Borja-Borja et al. [7] surveyed state-of-the-art deep learning methods for video processing to list the group and crowd activities. The main techniques of deep learning are grouped into Convolutional Neural Network(CNN), Autoencoders (AEs), and Recurrent Neural Network (RNN). Another survey-based on anomaly detection from video data by [8] focuses deep learning approach where the author listed generative adversarial networks (GANs) along with other deep learning approaches mentioned in [5]. A significant application of video processing in computer-vision research is pedestrian detection. Brunetti et al. [9] present a review on deep learning video processing methods for pedestrian detection focusing on methods CNN, Deep Neural Network(DNN), Restricted Boltzmann Machine (RBM), and Gaussian Mixture Model. In a survey by Ciaparrone et al. [10], the author reports deep learning methods for Multiple Object Tracking (MOT) from video data. They explored the Faster R-CNN, Mask R-CNN, SSD methods of deep learning for muti-object tracking. Apart from that, the authors also listed out the YOLO series of detectors – YOLOV2. Yan et al. [11] reported a review on deep multi-view learning from videos focusing on representational deep learning methods such as conventional neural networks, deep brief networks, and multi-view auto-encoders.

Taskiran et al. [12] present a taxonomy for face recognition as an image-based and video-based method. For video-based face recognition, various recent deep learning methods were discussed by grouping as set-based method and sequence-based method. Authors of [13] present a review of the video scene parsing application of video processing using deep learning techniques. They highlight the 2D CNN, 3D CNN, Clockwork FCN, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Spatio-temporal transformer GRU (STGRU), and GAN methods of deep learning. In [14], Wang et al. surveyed salient object detection from video data using deep-learning-based methods. They mentioned the specially designed methods for salient object detection such as fully convolutional network, Spatio-temporal cascade neural network, attentive feedback network, etc.

Tong et al. [15] investigate deep learning algorithms for video processing, specifically small object detection. The authors show the gradual improvement in CNN-based methods such as R-CNN, Fast R-CNN, Deconv-R-CNN, Improved Faster R-CNN, etc., along with other deep learning approaches. Sánchez et al. [16] demonstrate a study on crowd behavior analysis or crowd anomaly detection by video processing using deep learning techniques. Authors listed convolutional RBM, Fast R-CNN, 3D CNN, PCANet, deep Gaussian Mixture Model, Convolutional AutoEncoder (CAE) with LSTM (CAE - LSTM), Spatio-temporal CNN, and GAN based approach.

Table 1 shows the summary of related surveys included in this study.

TABLE 1
SUMMARY OF THE RECENT IMPORTANT RELATED SURVEYS

Year	Reported Paper	Focus of Study	Highlights
2018	Dhiman et al. [6]	Abnormal Human Activity Recognition	Handcrafted and deep learning approaches and their complexities, 2D and 3D datasets, Applications,
2018	Borja-Borja et al. [7]	Understanding group and crowd activity	Categories deep learning methods into CNN, AEs and RNN, Datasets, Challenges
2018	Brunetti et al. [9]	Pedestrian Detection	Field of application, acquisition technology, Deep learning techniques, Datasets, classification strategies
2019	Yan et al. [13]	Video Scene Parsing	Preliminaries, Background, fundamental terminologies, deep learning methods, dataset, evaluation of techniques based on experimental results, advantage and disadvantage of methods, Challenges, and future research
2020	Ramachandra et al. [8]	Video Anomaly detection	Types of anomalies, Deep learning approach along with GANs, Evaluation of methods, Future Research

Year	Reported Paper	Focus of Study	Highlights
2020	Ciaparrone et al. [10]	Multiple Object Tracking	Stage wise identification and explanation of deep learning methods for MOT, experimental comparison on datasets, Future research
2020	Taskiran et al. [12]	Face Recognition	Categories image-based and video-based face recognition methods, Deep learning methods for video-based face recognition, the historical development of methods, datasets, future research
2020	Wang et al. [14]	Salient Object Detection	Recent development, Deep learning methods with the framework, datasets, comparison of approaches, challenges, and future research
2020	Tong et al. [15]	Small-Object Detection	Deep learning methods from 5 different aspects, performance analysis on two datasets, Future research from various perspective
2020	Sánchez et al. [16]	Crowd behavior analysis	Taxonomy based analysis of deep learning method focuses on emotional aspects of crowd behavior analysis and the need for challenging datasets in this research area
2021	Nayak et al. [5]	Video Anomaly detection	Deep Learning methods, Datasets, Computational Infrastructure, Performance, Challenges, and future research
2021	Yan et al. [11]	Multi-view Learning (MLV)	Deep learning methods, an extension of traditional methods of MLV, Applications, datasets, performance analysis, challenges, and future research
2021	This SLR study	A systematic review and combined study of video processing using deep learning techniques	Various functionalities of video processing, application, categorization of DL for video processing into CNN, RNN, DNN, and Hybrid, video datasets, challenges, future research

B. MOTIVATION

We observed that a missing part is a combined review of various up-to-date video processing functionalities using deep learning techniques in related surveys. The surveys presented above focus only on specific functionality like video anomaly detection [5], abnormal human activity recognition[6], multi-object tracking[10], behavior analysis [16]. None of the surveys collate the research done on various functionality in one survey. Therefore, we motivate to present the recent advancement of deep learning-based video processing methods for multiple functionalities such as motion detection, object detection, human action recognition, object tracking, video classification, etc., and deep learning techniques to perform these functionalities. Another motivation is that in the past ten years, so many review papers are present for deep learning for image processing[17][18]; similarly, deep learning techniques for video processing are needed.

C. CONTRIBUTION

However, there is no single survey that provides an inside-out study covering all the aspects; our contribution in this paper is as follows:

- A systematic literature review to investigate the up-to-date research in video processing using deep learning techniques
- We include 93 research papers from journals and conferences listed in top databases that show the development pattern of advanced deep learning algorithms for video processing
- This paper can be helpful for researchers, where it gives knowledge for a better understanding of the advancement of deep learning techniques for video processing after the massive success of image processing.
- Shows open challenges and future research in this field
- This is the first SLR to present the various functionalities of video processing using deep learning in one paper to the best of our knowledge.
- Our observation shows that video processing advancements using deep learning techniques are majorly between 2017 and 2020 due to the advent of very deep networks based on AlexNet, ResNet, and LSTM.
- Significant research works found towards human action recognition, crowd anomaly detection, and behavior analysis from a computer-vision perspective.

The organization of our paper is as follows. Section 2 shows the research methodology; Section 3 shows the result of our study. The result section answers the RQs formed in section 2 based on evidence reported in the literature included in this SLR. Section 4 discusses the results of our analysis. Finally, section 5 summarizes the conclusion and section 6 shows the future research direction in this field.

In the end, Appendix 'A' contains the list of a glossary, and Appendix 'B' has the list of abbreviations for the most frequent terms used in this study.

II. RESEARCH METHODOLOGY

We conduct an SLR to assess the deep learning techniques in video processing. In this paper, we follow the standard procedures of SLR, explained by the authors Chitu Okoli, Kira Schabram [19]. This method identifies, specifies, and analyzes all the publications in deep learning for video Processing to present the answer to each research question (RQ) and reveal the gaps. This methodology of literature study shows the new insights of deep learning research on video data.

A. RESEARCH QUESTION

The vital part of the systematic review is determining research questions (RQs). We prepare research Questions (RQ) to follow the review process to stay focused at the beginning of the study. It is a novel approach to investigate

the answer of listed RQs deeply. During this process of forming RQs, the following points are considered:

– The search stage must identify the significant study that addresses the RQs.
– The data-extraction stage must extract the data items needed to answer the RQs.
– The data-analysis stage must synthesize the data; thus, the RQs can be answered.

Table 2 shows the list of RQs arises in this paper.

TABLE 2
LIST OF RESEARCH QUESTIONS PREPARED FOR THIS SLR.

ID	Research Question	Motivation
RQ1	What are the applications of video processing?	To investigate the wide range of areas where video processing techniques can help to ease the automation
RQ2	What are the various functionalities of video processing in the computer vision context?	To investigate what can be done with video data.
RQ3	What are the various deep learning techniques used by computer-vision researchers for video processing?	To explore the numerous successful researches done on video processing using deep learning techniques
RQ4	What datasets have been used by the researchers for video processing?	To explore the publicly available video datasets
RQ5	What are the issues and challenges in video processing?	It presents the challenges of implementing AI algorithms on video data while dealing spatially and temporally.

Here, RQ1 deeply investigates the application areas where video processing research is highly required—followed by RQ2 categorizes the video processing research with specific functionalities that can be performed on video data. Then, RQ3 tries to find up-to-date deep learning algorithms for various functionalities. RQ4 investigates the publicly available datasets for video processing so that researchers can discover suitable datasets for their experiments. Finally, RQ5 focuses on issues during research and implementation of video data using deep learning and future direction to pursue research in this area.

B. SEARCH METHOD

We use the following search methods in a step-by-step manner:

1) CHOOSE DATABASE

Table 3 shows the list of databases with respective focus subject areas we chose to perform this SLR.

2) CHOOSE KEYWORDS

Table 4 lists the keywords we used to search the papers from online databases.

3) CHOOSE TIME RANGE

We extract a total of 593 articles from 2011 to 2020 initially. Afterward, 276 articles were excluded since the paper's work did not match our objective. In the next step, we also exclude 204 masters, doctoral or unpublished articles. Finally, we include 93 papers for the final study, which are purely conference and journal papers. Figure 2. shows the year-wise distribution of publications in terms of percentage.

4) INCLUSION AND EXCLUSION CRITERIA

We include only conference and journal papers in this study published in the English language. Therefore, we have not considered under-reviewed papers and book chapters. Table 5 shows the various criteria on which basis we include and exclude the searched papers.

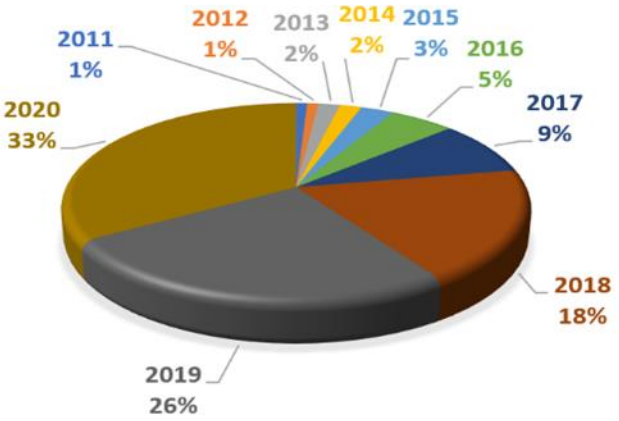

FIGURE 2. Year-wise distribution of the publication

5) QUALITY ASSESSMENT

We also assessed the quality of the selected 93 papers. The list of quality assessment questions is below.

— Is the purpose of the research meet?
– Since we exclude the unpublished papers, therefore we assess whether they were published in peer-reviewed journals?
– Is the video processing using deep learning techniques clearly explained?
– Are the accuracy and result of these researches being acceptable?
– Are these meets the standard of novel techniques of deep learning for video processing research
– Is there any systematic literature review conducted on video processing using deep learning before?
– Does the author clearly explain the purpose of using deep learning techniques on the specific dataset?

- Is there any comparative study conducted on deep learning methods for video processing?
- Does the study have an adequate average citation count per year?
- Has the author been given open access to data and code to apply these techniques by researchers further?

Finally, we reach the level to present the result of our study in answers to RQs defined in section 2.1. We point out that list of publications we considered purely to answer the RQs is between the time range 2011-2020, and few papers which are beyond time range are used only for background Study

TABLE 3
LIST OF ONLINE DATABASES AND RESPECTIVE SUBJECT AREAS *

IEEE Xplore	ACM Digital Library	Web of Science
Data Analytics and Soft Computing	Intelligent Machine	Video Data processing and video Analytics
Computing and Processing	Artificial Intelligence	Human Activity Recognition
Software Engineering	Distributed Computing	Object Detection
Information Technology	Real-Time System	Deep Learning
Computational Intelligence	High-Performance Computing	Neural Network
Cognitive Computing	Robotics	
Machine Learning	Computer Vision	
Data Science	Pattern recognition	
Business Intelligence	Image Processing	
Parallel Computing		

* Articles published in top conferences – ICML, CVPR, NeurIPS, ECCV, BMVC are also included and referenced in this paper, wherever required.

TABLE 4
KEYWORDS FOR SEARCHING IN THE ONLINE DATABASE.

S.No.	Keywords	No. of Articles
1	'Deep Learning' AND 'Video Processing'	7
2	'Deep Learning' AND 'Video Analysis'	5
3	('Artificial Intelligence' OR 'Deep Learning') AND ('Intelligent Video Analytics' OR 'Video Processing')	9
4	'Deep Learning' AND ('Human Action Recognition' OR 'Object Detection' OR 'Visual Object Tracking')	18
5	'Artificial Intelligence' OR 'Deep Learning' AND ('Video Data' OR 'Video Analysis' OR 'Video Processing')	19
6	'Video Classification' AND ('Artificial Intelligence' OR 'Deep Learning')	14
7	'Video Classification' AND ('Artificial Intelligence' OR 'Neural Network' OR 'Deep Learning')	11
8	(Object Detection AND Video Analysis) AND ('Deep Learning' OR 'Artificial Intelligence' OR 'Neural Network')	5
9	('Gait Analysis' AND 'Video Analysis') AND ('Deep Learning' OR 'Artificial Intelligence' OR 'Neural Network')	3
10	'Behavior analysis' AND 'Neural Network' AND 'Video.'	2
	Total	**93**

TABLE 5
INCLUSION AND EXCLUSION CRITERIA.

Inclusion criteria	Only conference and journal publication methodology of work shows deep learning for video processing for any functionality choose single publication if same deep learning method for same functionality found in multiple publications only digital video data processing is considered in the context of computer-vision research
Exclusion criteria	articles in other than the English language unpublished papers studies without proper validation and robust experimental research study having content other than video data processing-related research which did not use deep learning techniques.

III. RESULTS

A total of 93 peer-reviewed research papers on video processing using deep learning techniques were studied. This section observes the characteristics, methods, threats, solutions, and deep learning algorithms mentioned in the selected papers. After an intensive study, we present the results based on the research questions listed in table 1:

A. RQ 1: WHAT ARE THE APPLICATIONS OF VIDEO PROCESSING?

The current era is full of technology and rolling on the digital revolution. A wide range of video processing applications is in entertainment[20], healthcare[21], retail[22][23],traffic management[24], transport[25], home automation[26][27], flame and smoke detection[28][29][30],safety[31][32],and security[33]. These applications automatically generate the

caption of the actions happening in the video sequence. These video captions are further used to identify persons, vehicles, and other objects in the video sequence [34] and their appearance and actions [35]. The most desired application for video processing is producing actionable intelligence to help policymakers understand and respond to current situations[36][37]. Video processing applications also include education [38] and smart city video surveillance [39]. One of the popular applications of video analysis is crowd management[40]. This application helps to count the people at exit and entry points within a premise in real-time or at a periodic interval.

B. RQ 2: WHAT ARE THE VARIOUS FUNCTIONALITIES OF VIDEO PROCESSING IN THE COMPUTER VISION CONTEXT?

Video data has lots of functionalities on which researchers perform processing. For example, identifying some properties like attributes estimation, human pose estimation, person identification, motion detection can achieve when an object is detected in a video. Similarly, the task of human action recognition can perform in the video. Furthermore, some video analysis applications can process video offline, and some online, but many applications require situational awareness. Therefore, various applications of video analysis suggest the following functionalities, which performs on video data:

1) HUMAN ACTION RECOGNITION (HAR)

Human Action Recognition (HAR) is the task of identifying some actions from a video sequence[6], [41], [42].HAR is applicable in monitoring daily activities such as walking, bending, falling, climbing, sitting, etc., which is essential for activity analysis. HAR aims to identify the actions of one or more persons in the scene and gives helpful information about types of activities. HAR systems are also a part of human behavior monitoring in applications like injury detection during sports, elderly and child care, students' classroom behavior analysis, student-teacher classroom action recognition, and surveillance.

2) MOTION DETECTION

Motion detection is used to determine the presence of relevant motion in the observed scene. The objectives of motion analysis are to detect movements within frames of the video sequence, track an object's motion over time, group objects that move together, and identify the direction of motion. Specific techniques for implementing motion or movement analysis include background segmentation and differential equation models [43].

3) OBJECT DETECTION

Object detection is a technique to identify an object or entity, for example, a truck or a human in the video. In object detection tasks, visually observable objects in images of videos can be detected, localized, and recognized by computers[44], [45]. Detecting moving objects in video data has various applications in real life. In addition, the object detection in the video data helps a lot in real life, for example, to determine if there was a goal or not (in football), if a tennis ball is in/out of court (in tennis), or which athlete has finished first (in speed races), etc.

4) OBJECT RECOGNITION

Object recognition is a way of identifying the type of objects in the video sequence. When people observe something in a video, they can easily recognize the objects, scenes, and visual details. A driverless car is the best application of object recognition, modern technology now. Object recognition truly helps driverless cars distinguish a pedestrian from a street light [46] and recognize road signs, etc. It is also helpful in various applications such as robotics, industrial inspection, safety, smart city surveillance, and medical imaging.

5) OBJECT TRACKING

Object tracking or visual object tracking or video object tracking in video data is the process of tracking an object as it moves through space in a video. Object tracking divide into three different sections: initial object detection, assigning unique IDs, and tracking the objects across frames. Video object tracking is used for various applications like tracking faces and eyes for human-computer interaction, traffic control, video editing, surveillance, and security[10].

6) VIDEO CLASSIFICATION

Video classification focuses on automatically labeling videos based on video contents and frames. It is similar to image classification, in which images are classified based on the features belonging to a particular class. In the video classification task, video divides into frames (image) per second, and then a similar job of image classification performs [47].

7) BEHAVIOR ANALYSIS

Intelligent video processing, along with automatically detecting, recognizing, and tracking particular objects from image sequences, also aims to understand and describe object behaviors, detect abnormal behavior[48], hostile intent, etc. Behavior detection increases the speed and accuracy of suspicious detection and improves surveillance while reducing staff and equipment costs. The behavior detection system automatically detects suspicious behavior such as intrusion, loitering, and object abandonment based on user-defined time and location parameters. It can distinguish between humans, shadows, and moving objects.

8) GAIT ANALYSIS

Gait is the motion of human walking, whose movements can be faithfully reflected by the acceleration of the body sections. For every individual, human gait gestures a unique motion pattern. Therefore, gait analysis is a study of locomotion in both humans and animals. Coordination of several parts of the human body is watched and observed for gait analysis, such as the brain, spinal cord, nerves, muscles, bones, and joints. The study of gait analysis is widely applicable in healthcare, biometrics, sports, and many others. Gait dynamics [49] are captures using accelerometers and gyroscopes.

9) BACKGROUND SUBTRACTION

Sometimes, the interesting portion is not the background but the objects present in the foreground in a video scene. These interesting objects can be any object such as animals, humans, cars, etc. Detecting and processing a foreground object from a video is also known as background subtraction. Also, a standard method for search-space reduction and focus of attention modeling in video analysis is background subtraction techniques [50]. Eventually, it's easy to detect foreground objects if the background of a scene remains unchanged.

10) EVENT RECOGNITION

Event recognition is the technique of automatic analyses and recognizes the matching events from the video clips. Some Event recognition techniques from procedural videos are "baking a cake," "starting a vehicle," while other types of social activities like "birthday celebration," "Prayer," "Street dance." Many practical applications require identifying events, such as web video search, consumer video management, and intelligent advertising [51].

TABLE 6
FUNCTIONALITIES OF VIDEO ANALYSIS.

Functionality	References
Human action recognition	[6][41][42]
Motion detection	[43]
Object detection	[44][45]
Object recognition	[46]
Object Tracking	[10]
Video classification	[47]
Behavior analysis	[48]
Gait analysis	[49]
Background subtraction	[50]
Event recognition	[51]
Action segmentation	[52]
Scene understanding	[53]

11) ACTION SEGMENTATION

Video segmentation is a technique of dividing a video sequence into different sets of continuous frames similar to specific criteria. We observe that performing action segmentation before doing action recognition gives better recognition performance [52]. A challenging problem in human action understanding is to recognize a sequence of continuous actions, which is generally a segment. It recognizes primary actions such as jogging, jumping and sitting, etc., from a video sequence where a person's actions can be segmented into various categories. Action segmenting can be applied to different movements from the input video and recognizing the action types simultaneously.

12) SCENE UNDERSTANDING

Scene understanding is a study of scene structure (e.g., pedestrian road cross, market area, traffic on the road, waiting for the queue at the entrance, scene status (traffic light color change), scene motion patterns (cars taking U-turns). Unusual activity recognition improves with the understanding of scene patterns, tracking, and motion patterns. The increasing surveillance of massive crowds at sporting events, concerts, amusement parks, airports, and other venues motivates a growing desire to process and analyze crowd scenes, i.e., scene understanding[53].

C. RQ 3: WHAT ARE THE VARIOUS DEEP LEARNING TECHNIQUES USED BY COMPUTER-VISION RESEARCHERS FOR VIDEO PROCESSING?

AI algorithms have excellent success in video processing research. However, diversity in spatial and temporal makes video data a challenging task to recognize in the video sequence. To answer this question, we group the deep learning techniques for video processing into Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Hybrid approach.

1) CNN BASED APPROACH

In an early work, large-scale YouTube videos containing 487 sports classes were used to train a CNN model[54]. This CNN model includes a multi-resolution architecture that utilizes the local motion information in videos. In addition, it consists of a context stream for low-resolution image modeling, and further to classify videos, it contains fovea stream (for high-resolution image processing) modules. The author has also explained the three broad connectivity patterns: early fusion, late fusion, and slow fusion to extend the network's connectivity for time dimension to learn spatial-temporal features of the video data. Another work presents event detection from videos using CNN[55]. The author proposed an encoding method for spatial and temporal information using CNN and a frame descriptor to enhance the visual information. Similarly, for the event detection task, in [56], the author benefits from a pre-trained model on ImageNet to classify unusual events from the surveillance camera. This practice reduces the computational cost to train a large CNN model for video processing.

A general deep learning approach is two-stream CNN, proposed by Simonyan et al. [57]; it has two streams of CNN. In this architecture, two-stream has two separate layers, in which spatial information is stored through a single frame and another layer, using optical flow, temporal information is stored. Two-stream CNN combines regular images and optical flow images as input. To achieve high throughput, these two separate networks were combined with a late fusion technique. This video processing method has been experimented with human activity recognition tasks. To overcome the limitation of Spatio-temporal stream fusion at the softmax layer [57], another HAR approach [58] was introduced by fusing spatial and temporal networks at a convolution layer without losing performance. Other CNN-based methods developed for video processing for HAR tasks are [54], [59]. In a novel approach to video processing[60], the author proposes a MultiD-CNN framework for multimodal gesture recognition. This model combines two models; one is 3D Color-Depth Convolutional Network (3D-CDCN) and 2D Motion Representation Convolutional Network (2D-MRCN). These models mimic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3118541, IEEE Access

the architecture of deep residual networks (ResNets). The study shows the advantages of these two networks because the convolutional layers in the ResNets reduce the number of trainable parameters using the concept of weights sharing. From another's point of view, the ResNets also has few small connections that perform identity mapping and directly add the output of a particular layer to the output of later layers.

The author precisely [61] develop a deep learning algorithm to deal with large displacements in videos. They first create a matching algorithm – DeepMatch and prepare this model to match the 2D warping problem. They linked it to having a deformable SIFT descriptor grid, where all four quadrants can move independently to each other till a certain distance. A non-negative cosine similarity function uses for overlapping pixels to achieve a good scoring on possible warping. Finally, max-pool and subsample of the responses perform, generating a pyramid of features like SIFT. DeepFlow is DeepMatch combined with an energy minimization approach to generate the final optical flow for efficient video processing.

Another CNN-based method proposed by Nam et al. [62] for visual object tracking in online video data-Multi-Domain Network (MDNet).It is designed to learn the shared representation of targets from multiple and annotated video sequences for tracking, where each video is regarded as a separate domain. A different branch of domain-specific layers for binary classification has been used in this proposed method at the network's end. It shares the generic information captured from all video sequences, especially for generic representation learning. It is also observed that each domain in the MDNet is trained individually, where the shared layers are updated in each iteration. By following this method, the author has segregated the domain-independent information from domain-specific information. Using this technique, a generic feature was learned by the model for the representations of visual tracking.

DeepSORT [63] is one of the most widely used elegant object tracking CNN-based frameworks. The author used this Simple Online and Realtime Tracking (SORT) method for multiple object tracking, ahead of [62], focusing on simple, practical algorithms. The author has adopted a single conventional hypothesis tracking methodology with recursive Kalman filtering [1] and frame-by-frame data association. Therefore, this tracking scenario is defined on the eight-dimensional state space containing the bounding box center position, aspect ratio, height, and respective velocities in image coordinates. Also, a standard Kalman filter with constant velocity motion and linear observation model has been applied. The bounding points have been considered for the accurate position of the object.

To demonstrate efficient video processing for human action recognition, a multitask learning model is ActionFlowNet [64]. It trains a single stream network directly from raw pixels to jointly estimate optical flow simultaneously with the action recognition through CNN. Authors trained this model on motion information on unlabeled video clips. Also, it has more accuracy in action recognition with a large margin of 23.6% compared with the state-of-the-art CNN-based unsupervised representation learning [65]. A new CNN model for 6D object poses estimation, proposed by Xiang et al. [66], is introduced as PoseCNN. This PoseCNN estimates the 3D translation of an object by localizing its center in the image and predicting its distance from the camera. Here method used for the estimation of 3D rotation of the object is by performing the regression on quaternion representation of each frame. This method experiments for 21 objects on a large-scale 6D video dataset [2] exclusively designed for the 6D object pose estimation task.

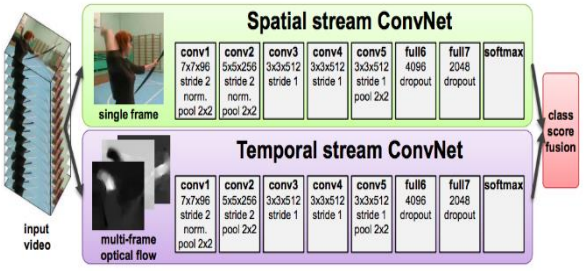

FIGURE 3. A typical two-stream CNN for video processing

Inspired by the great successes of CNNs in image recognition tasks, authors have used gait energy image (GEI), a popular image-based gait representation. The GEI was used as an input to the GEINet [67].GEINet was developed with two sequential triplets of convolution, pooling, and normalization layers, followed by two subsequent fully connected layers. From the cross-view gait recognition perspective, the authors have demonstrated that GEINet performs very well on the OU-ISIR large population dataset. In the popular video processing application for vehicle recognition, authors proposed[68] nine layers based CNN. For other video processing tasks such as background subtraction [69], object detection[70],[71], obstacle detection for self-driving car[46], anomaly detection in the crowded scene [72], [73], lane marking[74], monitoring wild animals [75], CNN based approach has been proven a good choice by the researchers.

Figure 3 shows the concept of a two-stream convolutional neural network for video processing in a Spatio-temporal manner.

2) DNN BASED APPROACH

A deep neural network is considered an advanced form of neural network with a higher level of complexity, i.e., a neural network with more than two layers.DNN based algorithms developed for video processing are capable of handling extensive higher dimension data. In [76], an object detection task is performed in the video using DNN with

[1]https://www.kalmanfilter.net/default.aspx

[2]https://rse-lab.cs.washington.edu/projects/posecnn/

SIFT [77] and tensor, which shows higher accuracy than previously developed methods. A robust deep neural network-based Multivariate Gaussian Fully Convolution Adversarial Autoencoder (MGFC-AAE) model was proposed by [78] to serve the demand of video anomaly detection and localization. In this model, the latent space representations of standard samples are trained to accord with a specific prior distribution.

Pashchenko et al. [79] use the DNN-based model for a transport system to recognize the critical situation. Amosov et al. [80] developed DNN based method for video processing for the classification probabilities for each video fragment; for normal and abnormal situations detection and recognition. Similarly, for road sign detection and lane detection tasks [81] for road analysis in automatic driving. DNN based video processing is performed by Luo et al. [82] for anomaly detection, in addition to sparse coding. This method aims to learn a dictionary to analyze various regular events with minor reconstruction errors.

3) RNN BASED APPROACH

A recurrent neural network (RNN) is also an artificial neural network that uses sequential or time-series data. This deep learning algorithm is generally used for temporal information. Since the nature of video data is Spatio-temporal, many video processing RNN based methods are developed in recent years. Long-short term memory (LSTM), Gated recurrent unit (GRU), Neural Turing Machines (NTM) are few popular variants of RNN. Among these, LSTM is one of the most widely used algorithms for video processing research.

The author of [83] proposed LSTM based 2-stage deep temporal model for group activity recognition. First, they design an LSTM model to analyze the action dynamics of individual people in a video sequence. The purpose of the second LSTM model is to aggregate person-level information for activity understanding in the entire video. In [84], Guo et al. propose GRU based approach for video processing for facial expression recognition. Authors implement facial landmark points and facial action units as input features in the training phase to effectively identify the facial regions and their components. One more GRU-based method was implemented for the video segmentation task by [85] by fully utilizing the temporal data in online videos.GRU preserves the temporal information part and maintains the spatial connectivities in the sequential frames of video data. A novel approach by [86] introduces a Multi-layer perception recurrent neural network (MLP-RNN), which is suitable for differentiating multiple objects of varying sizes. It works on the reference outline template to foreground analysis for the absence and presence of anomalies. For efficient target tracking, [87]implement LSTM based method in consideration of target motion uncertainty. An efficient target tracking problem estimates the target states from the measurements. Therefore, this method estimates the true states of the object strictly in a sequential manner for target tracking.

A big huddle in video processing is the low quality of video data. To perform the face recognition task from low-quality video frames, authors [88] propose a multi-mode Aggregation recurrent network, which obtains the discriminative features by aggregating the video frames' information. A different approach for video reconstruction is GRU-based, by [82], which is a tricky task in reconstructing each video frame and loss of temporal redundancy as a resultant. Unlike this traditional approach, authors invent a fast deep-learning GRU reconstructor that utilizes the Spatio-temporal features in a video. A vital feature pointed by the authors is GRU requires low memory.

A step ahead, In [89], authors utilize the video reconstruction techniques for estimates the scene background in videos. Authors exploit semantic segmentation to extract foreground objects, for example, person or moving vehicle, and summing background regions to reconstruct the background. Authors utilize Conditional Random Field as Recurrent Neural Networks (CRF as RNN) for the semantic segmentation to detect the areas of essential objects in each frame and map with foreground and background. In [90], the authors introduce RNN based video manipulation detection method, which shows a uniques advancement in video processing. This method shows the alteration in the new video while comparing it with pre-existing video frame by frame. This RNN based network is the fusion of ResNet and LSTM. A convolutional RNN and optical flow-based object segmentation from the video data approach are proposed by [91]. This method separates the object from the background in a video and mask the previous frames.

4) HYBRID APPROACH

The hybrid approach shows the combination of multiple deep learning methods uses for video processing. We found lots of research done using this approach. LSTM networks work based on frame-level CNN activation and combine information over time, as explained by the author [92], similar to temporal feature pooling. Compared to the vanilla recurrent neural network, LSTM has uninterrupted gradient flow, which is more accessible to back-propagate. LSTM is also more stable without gradient exploding or vanishing. The authors of [93] introduce temporal segment networks for human action recognition. The idea is to process the video data by dividing it into equal parts, extract small snippets from each part of the video, classify each snippet using a multistream CNN network and then form a consensus over the classification to output a softmax score for the whole video finally. They justify this as a way of better modeling global temporal dynamics. They find that RGB, Optical Flow, and Warped optical flow together give the best results. In [94], the authors of [93] extend their work on temporal segment networks to work with untrimmed videos and combined the classification scores from different snippets instead of simple averaging. They experimented with different weights, including those based on attention, measured using activations from the last layer of the RGB CNN.

Video processing on real-time yoga pose recognition using deep learning has been done by [95], in which a hybrid deep learning model is proposed using CNN and LSTM.In

this method, The CNN algorithm extracts various pose features, and then it exploits the features of LSTM for actual temporal predictions. In addition, a new technique called Recurrent Convolution Networks (RCN) [96] was introduced for video processing in recent years. It applies CNNs on video frames for visual understanding and then feeds the frames to RNNs for analyzing temporal information in videos. To leverage the advantage of RNN for modeling high-level temporal feature sequences, authors of [97] propose an Inflated 3D [98] and LSTM based novel model for human action recognition. First, the author pre-trains a 3D CNN model on a vast video action recognition dataset Kinetics, which improves the model's generality. Thereafter, long-short term memory(LSTM) is used to learn the high-level temporal features produced by the Kinetics-pre-trained 3D CNN model. Another contribution for HAR [96], employ the long-term recurrent convolutional network to overcome the issues of variable length input sequences. It combines convolutional layers(for visual recognition) and LSTM(for time-varying sequence learning) network. This network is also helpful for image caption generation and video description tasks. Another hybrid approach for video processing for HAR is [99], which uses 3D CNN and LSTM. In contrast, older methods of deep learning-based HAR [100] experimented with a hybrid classifier by fusion weights generated by homogeneous models arranged in a parallel architecture. A new transformer network adopts the attention mechanism in deep learning and outperforms when combined with spatio-temporal based models like CNN for human activity recognition. Girdhar et al. [101] prepose deep learning Transformer-based model combined with I3D network for HAR to collectively identify the spatiotemporal features of the person and the surrounding of the person whose action is trying to recognize. The connected network works on attention mechanisms unsupervised, primarily focusing on hands and faces, which mainly contribute to accurate human action recognition tasks.

Authors A. Hu et al. [102] presented a deep learning probabilistic model for the autonomous vehicle's video scene understanding of real-world urban scenes. This model learns features from the spatio-temporal convolutional network to predict future scene representation jointly by encoding the future state into a low-dimensional future distribution.

In [103], the authors use optical flow, CNN, LSTM, and support vector machine (SVM) for gesture recognition from video data. This approach is highly applicable for decoding the news for the deaf-mute community. The optical flow method is used to detect and process moving target objects on video. In [104], the authors propose a novel approach for adult content detection in videos, namely ACORDE (Adult Content Recognition with Deep Neural Networks). This method combines CNN as a feature extractor and LSTM for classification. [105] explain the anomaly detection in real-time videos by using optical-flow convolutional autoencoder and convolutional LSTM. It shows a better performance than the vanilla CNN or DNN based approach for anomaly detection. Automatic event detection from video data is presented in [106] by using CNN and RNN. This network efficiently recognizes soccer events from live video streaming by leveraging feature learning and deriving temporal relations through CNN and RNN.

The need to detect the most desirable objects from the dynamic video is fulfilled in [86], where salient object detection is done using a hybrid Convolutional Recurrent Neural Network (CRNN). Salient objects in the moving scene are detecting by capturing the temporal, spatial, and local constraint features with the CNN and RNN based CRNN model.

A very challenging application of the video process is healthcare, where a novel approach proposed by [107] shows the intelligent monitoring of tools used during surgery in the operation theatre. Authors bring in the notice that the tool used to record either through a microscope or an endoscope. This state-of-the-art video processing technique process each frame of the video by CNNs, and its learned outputs are sent to RNNs to fully utilize the temporal relationships between frames. Another video processing approach in medical is proposed in [21]. The author developed a multitask recurrent convolutional network with correlation loss (MTRCNet-CL) and fulfills the need for surgical tool presence detection and surgical phase recognition.

Furthermore, [108] performed video captioning to generate text descriptions of video frames using CNN and transformers by introducing a video encoder, proposal decoder, and captioning decoder.

Table 7 shows the summary of deep learning approaches for video processing.

D. RQ 4: WHAT DATASET HAVE BEEN USED BY THE RESEARCHERS FOR VIDEO PROCESSING

We found various video datasets in the literature on which researchers have shown the experiment of deep learning for video processing. Table 8 shows the details of the video dataset with applications used by the researchers.

TABLE 7
SUMMARY OF DEEP LEARNING APPROACHES FOR VIDEO PROCESSING

Algorithms	Reported Papers	Focus Area	Model Architecture	Published Year
Convolutional Neural	Sanil et al. [46]	Obstacle detection in self-driving car	CNN with IoT implemented as object detection in videos	2020
	Karpathy et al. [54]	Human Action Recognition	Multiresolution CNN- Context stream + Fovea stream	2014

Algorithms	Reported Papers	Focus Area	Model Architecture	Published Year
Network (CNN)	Xu et al. [55]	Event Detection	CNN+ encoding method + Dense trajectories	2015
	Sahoo et al. [56]	Unusual event detection	Two-stream 2D ConvNet pre-trained on ImageNet+Classifiers(SVM, K-NN, RBFN, Naive Bayes, Logistic Regression, K-means clustering)	2019
	Simonyan et al. [57]	Human Action Recognition	2 Stream ConvNet	2014
	Feichtenhofer et al. [58]	Human Action Recognition	Convolutional Two Stream (VGG 16)	2016
	Ji et al. [59]	Human Action Recognition	Three streams of CNN	2013
	Elboushaki et al. [60]	Gesture Recognition	Multi-Dimensional CNN (Convolutional Residual Networks + Convolutional Long Short-Term Memory Networks)	2020
	Nam et al. [62]	Visual Object Tracking	Pre-trained CNN with a new binary classification layer	2016
	Wojke et al. [63]	Object Tracking	CNN - Simple Online and Realtime Tracking (SORT) + appearance information	2017
	Ng et al. [64]	Human Action Recognition	CNN+Optical Flow, Pretrained on ImageNet and Sports-1M	2018
	Fernando et al. [65]	Video Representation Learning	Multi-stream convolutional neural network	2017
	Xiang et al. [66]	Pose estimation	CNN for 6D object + 3D translation+3D Rotation	2018
	Shiraga et al. [67]	Gait Recognition	Gait Energy Image Network (GEINet) - two sequential triplets of CNN + pooling and normalization layers and two fully connected layers,	2016
	Luo et al. [68]	Vehicle and Face Recognition	CNN with 9 layers	2017
	Babaee et al. [69]	Background Subtraction	CNN with parameter tunning	2018
	Lu et al. [70]	Object Detection	YOLO network based on GoogLeNet	2019
	Pérez-Hernández et al. [71]	Object Detection	CNN with binary classifier	2020
	Sabokrou et al. [72]	Anomaly Detection	Fully convolutional neural networks	2018
	Cheong et al. [73]	Crowd monitoring and Counting	Background subtraction+CNN based classifier	2019
	Y. Tian et al. [74]	Lane marking detection	Faster R-CNN + context cues	2018
	Villa et al. [75]	Video Monitoring	Deep Convolutional Neural Network, Pre-trained on ImageNet dataset	2017
	Najva and Bijoy [76]	Object Detection	DNN+SIFT+Tensor features	2016
	Li and Chang [78]	Anomaly detection	Deep neural network - Multivariate Gaussian Fully Convolution Adversarial Autoencoder (MGFC-AAE)	2019
Deep Neural Network (DNN)	Pashchenko et al. [79]	Recognizing critical situations for transport systems	DNN Classification with subsequent reinforcement	2019
	Amosov et al. [80]	Normal and abnormal situation detection and recognition	Ensemble deep neural network	2019
	Shukla et al. [81]	Road sign and lane detection	Neural network + HOG+total variation biletaral and wavelet filter	2019
	W. Luo et al. [82]	Anomaly detection	Temporally-coherent Sparse Coding + DNN	2020
Recurrent Neural	[83]	Group Activity Recognition	2-stage deep temporal model with LSTMs	2016
	Guo et al. [84]	Facial Recognition	Face-to-sequence approach with GRU	2019

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3118541, IEEE Access

Algorithms	Reported Papers	Focus Area	Model Architecture	Published Year
Network (RNN)	Siam et al. [85]	Video segmentation	Convolutional GRU for binary and semantic video segmentation	2017
	Murugesan and Thilagamani [86]	Anomaly detection	Maximally Stable Extremal Region (MSER) + Multi-layer perception recurrent neural network (MLP RNN)	2020
	Gao et al. [87]	Target Tracking	Vanilla RNN + LSTM+DeepRNN + DeepLSTM	2019
	Gong et al. [88]	Face Recognition	Multi-mode Aggregation Recurrent Network (MARN)	2019
	Savakis et al. [89]	Background estimation	Conditional Random Field as Recurrent Neural Networks (CRF as RNN)	2018
	Howard et al. [90]	Video manipulation detection	Joint Residual Network (ResNet) feature extractor + Long Short- Term Memory (LSTM) as Recurrent Residual Feature Learning Network	2019
	Kalezic et al. [91]	Video segmentation	Convolutional RNN+optical flow	2020
	Mur et al. [109]	Video construction	Convolutional GRU	2020
	Jin et al. [21]	video analysis	Multi-task recurrent convolutional network with correlation loss (MTRCNet-CL) + LSTM	2019
	Ng et al. [92]	Human action recognition	CNN(ALexNet+ GoogleNet) + Two-Stream LSTM	2015
	L. Wang et al. [93]	Human action recognition	Temporal Segment Network (Spatial ConvNet + Temporal ConvNet)	2016
	L. Wang et al. [94]	Human action recognition	Temporal Segment Network	2019
	Yadav et al. [95]	Human action recognition (Yoga recognition)	CNN+LSTM	2019
Hybrid	J. Donahue et al. [96]	Human action recognition	CNN+LSTM	2017
	Wang et al. [97]	Human action recognition	Inflated 3D + LSTM	2019
	Arif et al. [99]	Human action recognition	3D-CNN and LSTM	2019
	Ijjina and Mohan [100]	Human action recognition	fusion of homogeneous convolutional neural network (CNN) classifiers	2016
	Girdhar et al. [101]	Human action recognition	Deep learning -Transformer	2019
	Hu et al. [102]	Scene understanding	Deep learning – Probabilistic method	2020
	Xiaoxue et al. [103]	Gesture recognition	Optical flow+CNN+LSTM+PReLU(activation function)	2019
	Wehrmann et al. [104]	Content detection	CNN+LSTM	2018
	Duman and Erdem [105]	Anomaly detection	Optical flow+Convolutional Auto Encoder + LSTM	2019
	Jiang et al. [106]	Event detection	CNN + LSTM	2017
	Hajj et al. [107]	Object detection (monitoring tool detection)	CNN + RNN	2018
	Zhou et al. [108]	Video captioning	Deep learning - Transformer	2018

TABLE 8
VIDEO DATASETS USED IN VARIOUS VIDEO PROCESSING TECHNIQUES.

Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	Application	Release year	Ref.
UCF-101	101	13320	Realistic action videos, an extension of the	YouTube	Human Activity	2012	[110]

Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	Application	Release year	Ref.
			UCF50 data set, which has 50 action categories		Recognition (HAR)		
HMDB51	51	6849	Each class containing a minimum of 101 clips	Movies, public databases, and YouTube		2011	[111]
ActivityNet	200	20000	Videos for Human Activity Understanding	Web-Search		2015	[112]
Kinetics-400, 600, 700	400, 600, 700	300000, 500000, 65000	YouTube video URLs dataset	YouTube		2017, 2018, 2019	[113]–[115]
YouTube-8M	1000	8000000	YouTube video URLs dataset	YouTube	Video classification, HAR	2016	[116]
UCSD	Binary flag per frame (Anomaly present or not)	200	Manually recorded through stationary camera mounted at an elevation, overlooking pedestrian walkways. Anomalies are - bikers, skaters, small carts, wheelchair	Manual Recording	Anomaly detection, localization	2014	[117]
UMN	Binary (Normal and abnormal behavior)	11	Crowd Escape Panic, Videos are of a normal starting section and an abnormal ending section	Manually Recorded through head-mounted camera	Crowd Behavior Detection	2009	[118]
Avenue	-	37	A challenging dataset with where training videos capture normal situations whereas, Testing videos include both normal and abnormal events.	videos are captured in CUHK campus avenue	Abnormal Event Detection	2013	[119]
VOT2013-2018	16		Sequences are from a large pool of sequences using a methodology based on clustering visual features of objects.	Annotation by VOT committee	Visual object Tracking	2013-2018	[120]–[125]
UCF Sports	10		Sports dataset consists of a set of actions collected from various sports	broadcast television channels such as the BBC and ESPN	Sports action recognition	2014	[126]
BAHAVE	5	4	125 instances of people were marked up for a total of 83545 bounding boxes		Multi-person behavior classification	2010	[127]
Multimodal Human Action Database (MHAD)	11	660	Recording of T-pose for each subject to use as skeleton extraction; and the background data	Manual recording	HAR (depth camera-based)	2015	[128]
KITTI Vision benchmark	-		Recording at rural areas, up to 15 cars, and 30	Manual recording with standard	Object Tracking	2012	[129]

Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	Application	Release year	Ref.
			pedestrians are visible per image.	station wagon with two high-resolution color and grayscale video cameras			
IJB-S	-	202	Videos are of interest to law enforcement and national security communities	Recorded manually at Department of Defense (DoD) training facility, U.S.A		2014	[130]
YTF	-	3425	database of face videos designed for studying the problem of unconstrained face recognition in videos	YouTube	Face recognition	2011	[131]
PaSC	-	2802	This database is part of the Point and Shoot Face Recognition Challenge (PaSC) to spur advancement in the face and person recognition	Manually Recorded		2013	[132]
DAVIS		3455	This database is now part of the RobMOTS Challenge	Manually Recorded	Video segmentation	2016	[133]

E. RQ 5: WHAT ARE THE ISSUES AND CHALLENGES OF VIDEO PROCESSING USING DEEP LEARNING?

Although there has been significant progress over the past few years, there are still many challenges in applying deep learning techniques to video processing and develop models for real-life application. Various challenges exist as a huddle in this research area, such as:

1) POOR QUALITY OF VIDEOS

Poor-quality videos captured through live cameras installed at long-distance create severe occlusions, and it exists in many scenarios of the video surveillance system. Public gatherings and overcrowded places such as religious events, airport arrival, and departure terminals are significant points where occlusions happen frequently. Apart from surveillance cameras installed in high areas cannot capture high-quality videos like present video datasets in which the target person is apparent and obvious. Due to the long distance of cameras, the subject is relatively small, making it challenging to process. The relatively low quality of those long-distance videos further increases the difficulty.

2) TRACKING AND LOCATING OF MULTISUBJECT

In real-world tracking, any single object from multiple moving objects in the video is complex. The main challenge of tracking is the target motion uncertainty due to the tracker's unavailability of an accurate dynamic model [87]. In particular, no surety of the transition function and the complex calculation of the densities between time series.

3) DYNAMIC BACKGROUNDS

Most of the real-world applications capture complex and evolving backgrounds through the surveillance camera. As a result, these types of videos are recorded in various dynamic backgrounds. Also, real-time video scenes certainly have illumination variance, occlusions, and changing viewpoints, which makes it very difficult for video processing in such complex and various dynamic situations.

4) LACK OF DATASET

A large number of video processing datasets is also desirable to experiment using various DL techniques. Such as the action recognition task on JHMDB was proven too challenging because of its data annotation method. It achieves inaccurate performance in the past research. So apart from data collection, proper annotation is also vital in video datasets. The lack of a properly formed video dataset is still a challenge in video processing research. Video datasets from various domains are also highly desirable. Studies have focused that the availability of video data is also a major issue. While few are publicly available but many data sets are still not available for open research. (Training, Testing, and inferencing).

5) LACK OF COMPUTATION POWER

Besides methodology breakthroughs and available big training data, the recent success for video processing is also due to advances in hardware. Researchers faced few challenges of unavailability of enough computing resources for large-scale video data processing. Since deep learning

algorithms need specialized processing hardware called GPU[134], highly data-intensive and compute-intensive computing machines are required.

IV. DISCUSSION

A. SUMMARY OF REVIEWED STUDIES

We observe a rapid advancement in video processing using deep learning techniques between 2017 and 2020 compared to earlier research between 2011 and 2017. As per the study, CNN has outperformed in most video processing functionalities-video classification, scene labeling, and scene understanding, whereas RNN based approaches are best proven for visual object tracking or long-term temporal relationships. The performance is much improved in Temporal Segment Networks combined with LSTMs than vanilla CNN or RNN. When comparing LSTM with CNN, most researchers have concluded that both the algorithms perform well and fit appropriately for the Video Classification task. In addition, CNN has also performed well for scene labeling using a parametric model to learn discriminative features and classifiers. In the case of DNN, it significantly improves action recognition accuracy by a large margin than CNN-based unsupervised representation learning methods trained without large-scale external data and additional optical flow input. Without pretraining on large external labeled datasets, the models trained with large labeled datasets such as ImageNet and Sports-1M achieve more considerable accuracy.

In contrast, vanilla RNN faces a short-term memory problem due to the vanishing gradient problem; therefore, LSTM has been proven to better perform in the Spatio-temporal nature of video processing. Furthermore, research shows that a deep-learning GRU reconstructor is fast and requires low memory, unlike traditional approaches. The various hybrid approach, where a combination of CNN and LSTM is employed, has shown tremendous improvement in the network architecture and handling of Spatio-temporal feature and long-term learning of patterns. It has been observed that most of the research aimed to design the algorithms with significant speed-up without loss of accuracy.

The video processing deep learning techniques are also advancing due to the advent of various video datasets in multiple domains – UCF 101, UMN, UCSD, Avenue, etc. listed in table 9. However, the diversity and nature of datasets make the algorithm learn close to real-time features in a controlled environment.

Various challenges have also been faced by researchers for developing deep learning algorithms for video processing. A big huddle lacks openly available data and costly hardware, which requires training, testing, and inferencing. Apart from occlusion, the poor camera video quality in a real-time environment makes the deep-learning models challenging to perform well.

B. THREATS TO VALIDITY

This section discusses the possible threats that might have affected our systematic literature review and how we alleviated them. Validity is the degree to which the results estimate what they are supposed to do.

1) THREATS TO INTERNAL VALIDITY

The fundamental threat to internal validity is the literature we collected for our study. We found limited research papers- 93 out of 593, in which researchers use actual deep learning techniques for video processing in the context of computer vision. Furthermore, the few studies that show a novel approach with an experimental dataset, either those methods or datasets, are in arXiv. In contrast, our study was bound to include only peer-reviewed articles published in journals/conferences indexed in reputed databases.

2) THREATS TO EXTERNAL VALIDITY

External validity limits the ability to generalize the results beyond our study. We mentioned the accuracy of deep learning methods on datasets as reported by the original papers by authors. We did not perform any experimental research to re-calculate the results. Hence the generalization of deep learning techniques for video processing is shown as reported in its original research papers.

3) THREATS TO CONCLUSION VALIDITY

Conclusion validity is the degree to show the reasonability of the relationship between data and conclusions. Since the restricted access of other reputed databases like Scopus, Springer, and Wiley are beyond our research work. Therefore, we could not retrieve the literature published in these databases. Undoubtedly the inaccessibility of literature listed in these databases mitigates us to conclude the final result on the advancement of video processing using deep learning techniques. Deep learning methods for video processing may be a lot more than we include in this study. Hence the conclusion based on only 93 literatures extracted from WoS, ACM, and IEEE databases may not be adequately present the advancement of deep learning techniques for video processing between 2011 and 2020. This limitation also impacts our study.

V. CONCLUSION

In this paper, we have presented a systematic literature review of the deep learning techniques for video processing in the context of computer vision. We included 93 research papers published in the peer-reviewed journal/ conference indexed in WoS, ACM, and IEEE Xplore between 2011 and 2020. We present the SLR by forming the RQs and systematically answering them in terms of various applications and functionalities of video processing and deep learning techniques, datasets, and challenges. Finally, we conclude the few main points of our study:

— Deep learning techniques can now boost video understanding, video classification, video analysis, action recognition, and pose recognition.
— The advent of pivotal research in AlexNet and ImageNet for image processing gave a clear direction to perform

analysis in video processing. Therefore, more literature was found on this topic between 2017 to 2020.

— We found significant work on video processing using deep learning for human action recognition, behavior analysis, and crowd anomaly detection.

VI. FUTURE RESEARCH

Since deep learning techniques are suitable for handling large-scale video data, they can process and analyze millions of data captured from the distributed sensors. There are many active research topics in future directions regarding such data, such as threat identification, multi-person identification, multi-object tracking, scene labeling, etc. It has been observed that some topics like action recognition, video classification, and object tracking got enough research. However, surprisingly we did not find many research articles on scene labeling, scene understanding, video analysis from moving cameras, and cluttered backgrounds. However, despite remarkable progress, the advances achieved so far do not meet high accuracy standards and the correct realization of video processing in some areas, such as video surveillance in low light, partially captured areas, Gait Recognition, etc. A large number of video datasets should also be freely available in the future. However, lots of deep research is required, along with colossal computation power such as tensor core-based GPU for training huge neural networks.

APPENDIX A

- *Artificial Intelligence:* Artificial Intelligence is a field of computer science where algorithms are designed to make the machine capable of performing tasks intelligently without being explicitly instructed
- *Computer Vision:* Computer vision is a field of artificial intelligence (AI) that enables machines to derive meaningful information from digital images, videos, and other visual inputs and take actions based on that information.
- *Deep Learning:* Deep learning is an AI algorithm designed by using neural network architectures that contain many layers, sometimes called deep layers.
- *Video Processing:* In the context of computer vision, video processing or digital video processing is the ability to automatically analyze video, frame by frame, to detect and determine temporal and spatial features.
- *Systematic literature Review:* A systematic literature review (SLR) identifies, selects, and critically appraises research to answer a formulated question.
- *Convolutional Neural Network (CNN):* A convolutional neural network is formed by stacking different layers that transform the input; convolutional layer, pooling layer, activation layer, and fully connected layer.
- *Deep Neural Network (DNN):* A neural network with some level of complexity, usually at least two layers, qualifies as a deep neural network or, say, the deeper form of neural network.
- *Recurrent Neural Network (RNN):* A recurrent Neural Network is a type of neural network where the output from the previous step are fed as input to the current step
- *Long Short Term Memory (LSTM):* LSTM is a type of RNN where the information flows through a mechanism known as cell states. This way, LSTMs can selectively remember or forget things.
- *Transformers:* A transformer is a deep learning algorithm model that adopts the attention mechanism, thus differentially weighing the significance of each part of the input data.
- *Probabilistic Model:* A probabilistic model predicts the probability distribution over a set of classes, rather than only outputting the most likely class as output.
- *Benchmark Video Datasets:* Benchmark video datasets are adequately prepared, annotated, validated, and proven to be accurate compared with other datasets.
- *Peer-Reviewed Articles:* The articles, reviewed and critiqued by the author's peers who are experts in the same subject area.
- *High-Performance Computers or Supercomputers:* One of the best-known types of HPC solutions is the supercomputer. A supercomputer contains thousands of compute nodes that work together to complete one or more tasks. This is called parallel processing. It's similar to having thousands of PCs networked together, combining compute power to complete tasks faster.

APPENDIX B

- DL: Deep Learning
- HAR: Human Activity Recognition
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- DNN: Deep Neural Network
- LSTM: Long-Short Term Memory
- GRU: Gated Recurrent Unit
- RCNN: Recurrent Convolutional Neural Network
- MTRCNet-CL: Multi-task Recurrent Convolutional Network with Correlation Loss
- YOLO: You Only Look Once
- SIFT: Scale-Invariant Feature Transform
- HOG: Histogram of Oriented Gradients
- MSER: Maximally Stable Extremal Region
- MLP-RNN: Multi-Layer Perception-Recurrent Neural Network

REFERENCES

[1] T. O'Shea and J. Hoydis, "An Introduction to Deep Learning for the Physical Layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, 2017, doi: 10.1109/TCCN.2017.2758370.

[2] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges," *IEEE Trans. Netw. Serv. Manag.*, vol. 16, no. 2, pp. 445–458, 2019, doi: 10.1109/TNSM.2019.2899085.

[3] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, "Exploring Behavior Analysis in Video Surveillance Applications," *Int. J. Comput. Appl.*, vol. 93, no. 14, pp. 22–32, 2014, doi: 10.5120/16283-6045.

[4] Y. K. Han and Y. B. Choi, "Human Action Recognition based on LSTM Model using Smartphone Sensor," *Int. Conf. Ubiquitous Futur. Networks, ICUFN*, vol. 2019-July, pp. 748–750, 2019, doi: 10.1109/ICUFN.2019.8806065.

[5] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis. Comput.*, vol. 106, p. 104078, 2021, doi: 10.1016/j.imavis.2020.104078.

[6] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, no. June 2018, pp. 21–45, 2019, doi: 10.1016/j.engappai.2018.08.014.

[7] L. F. Borja-Borja, M. Saval-Calvo, and J. Azorin-Lopez, "A Short Review of Deep Learning Methods for Understanding Group and Crowd Activities," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, 2018, doi: 10.1109/IJCNN.2018.8489692.

[8] B. Ramachandra, M. Jones, and R. R. Vatsavai, "A Survey of Single-Scene Video Anomaly Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2020, doi: 10.1109/TPAMI.2020.3040591.

[9] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, 2018, doi: 10.1016/j.neucom.2018.01.092.

[10] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020, doi: 10.1016/j.neucom.2019.11.023.

[11] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021, doi: 10.1016/j.neucom.2021.03.090.

[12] M. Taskiran, N. Kahraman, and C. E. Erdem, "Face recognition: Past, present and future (a review)," *Digit. Signal Process. A Rev. J.*, vol. 106, p. 102809, 2020, doi: 10.1016/j.dsp.2020.102809.

[13] X. Yan *et al.*, "Video scene parsing: An overview of deep learning methods and datasets," *Comput. Vis. Image Underst.*, vol. 201, no. November 2019, p. 103077, 2020, doi: 10.1016/j.cviu.2020.103077.

[14] Q. Wang, L. Zhang, Y. Li, and K. Kpalma, "Overview of deep-learning based methods for salient object detection in videos," *Pattern Recognit.*, vol. 104, p. 107340, 2020, doi: 10.1016/j.patcog.2020.107340.

[15] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image Vis. Comput.*, vol. 97, p. 103910, 2020, doi: 10.1016/j.imavis.2020.103910.

[16] F. Luque Sánchez, I. Hupont, S. Tabik, and F. Herrera, "Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects," *Inf. Fusion*, vol. 64, no. July, pp. 318–335, 2020, doi: 10.1016/j.inffus.2020.07.008.

[17] L. Jiao and J. Zhao, "A Survey on the New Generation of Deep Learning in Image Processing," *IEEE Access*, vol. 7, pp. 172231–172263, 2019, doi: 10.1109/ACCESS.2019.2956508.

[18] H. L. MD. Zakir Hossain,Ferdous Sohel,Mohd Fairuz Shiratuddin, "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.

[19] C. Okoli and K. Schabram, "A Guide to Conducting a Systematic Literature Review of Information Systems Research," *SSRN Electron. J.*, vol. 10, no. 2010, 2012, doi: 10.2139/ssrn.1954824.

[20] M. Khan, M. A. Tahir, and Z. Ahmed, "Detection of Violent Content in Cartoon Videos Using Multimedia Content Detection Techniques," *Proc. 21st Int. Multi Top. Conf. INMIC 2018*, 2018, doi: 10.1109/INMIC.2018.8595563.

[21] Y. Jin *et al.*, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Med. Image Anal.*, vol. 59, p. 101572, 2020, doi: 10.1016/j.media.2019.101572.

[22] A. H. Ferrández Vicente J., Álvarez-Sánchez J., de la Paz López F., Toledo-Moreo F., "Identification of Loitering Human Behaviour in Video Surveillance Environments," 2015, [Online]. Available: https://doi.org/10.1007/978-3-319-18914-7_54.

[23] T. T. Zin, P. Tin, H. Hama, and T. Toriu, "Unattended object intelligent analyzer for consumer video surveillance," *IEEE Trans. Consum. Electron.*, vol. 57, no. 2, pp. 549–557, 2011, doi: 10.1109/TCE.2011.5955191.

[24] V. Mandal, A. R. Mussah, P. Jin, and Y. Adu-gyamfi, "sustainability Artificial Intelligence-Enabled Tra ffi c Monitoring System," *MPDI*, 2020.

[25] E. Vorakitolan, J. P. Havlicek, R. D. Barnes, and A. R. Stevenson, "Simple, effective rate control for video distribution in heterogeneous intelligent transportation system networks," *Proc. IEEE Southwest Symp. Image Anal. Interpret.*, pp. 37–40, 2014, doi: 10.1109/SSIAI.2014.6806023.

[26] T. Sultana and K. A. Wahid, "IoT-Guard: Event-Driven Fog-Based Video Surveillance System for Real-Time Security Management," *IEEE Access*, vol. 7, pp. 134881–134894, 2019, doi: 10.1109/ACCESS.2019.2941978.

[27] R. Majeed, N. A. Abdullah, I. Ashraf, Y. Bin Zikria, M. F. Mushtaq, and M. Umer, "An Intelligent, Secure, and Smart Home Automation System," *Sci. Program.*, vol. 2020, 2020, doi: 10.1155/2020/4579291.

[28] S. Saponara, A. Elhanashi, and A. Gagliardi, "Real-time video fire/smoke detection based on CNN in antifire surveillance systems," *J. Real-Time Image Process.*, vol. 18, no. 3, pp. 889–900, 2020, doi: 10.1007/s11554-020-01044-0.

[29] F. Gong *et al.*, "A real-time fire detection method from video with multifeature fusion," *Comput. Intell. Neurosci.*, vol. 2019, 2019, doi: 10.1155/2019/1939171.

[30] Y. Valikhujaev, A. Abdusalomov, and Y. Im Cho, "Automatic fire and smoke detection method for surveillance systems

based on dilated cnns," *Atmosphere (Basel).*, vol. 11, no. 11, pp. 1–15, 2020, doi: 10.3390/atmos11111241.

[31] Q. Zhang, H. Sun, X. Wu, and H. Zhong, "Edge video analytics for public safety: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1675–1696, 2019, doi: 10.1109/JPROC.2019.2925910.

[32] Iulia Lefteret al., "Automated safety control by video cameras," 2012.

[33] J. Panchashila and P. Malathi, "Implementation of tripwire system using video surveillance for railway platform security," *Proc. - 2014 IEEE Int. Conf. Adv. Commun. Comput. Technol. ICACACT 2014*, pp. 1–4, 2015, doi: 10.1109/EIC.2015.7230743.

[34] M. Mazumdar, V. Sarasvathi, and A. Kumar, "Object recognition in videos by sequential frame extraction using convolutional neural networks and fully connected neural networks," *2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput. ICECDS 2017*, pp. 1485–1488, 2018, doi: 10.1109/ICECDS.2017.8389692.

[35] A. W. Senior, Y. Tian, and M. Lu, "Interactive motion analysis for video surveillance and long term scene monitoring," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6468 LNCS, no. PART1, pp. 164–174, 2011, doi: 10.1007/978-3-642-22822-3_17.

[36] Z. Xu, S. Sinha, S. Shah Harshil, and U. Ramachandran, "Space-time vehicle tracking at the edge of the network," *Proc. Annu. Int. Conf. Mob. Comput. Networking, MOBICOM*, pp. 15–20, 2019, doi: 10.1145/3349614.3356025.

[37] K. Yu *et al.*, "Design and Performance Evaluation of an AI-Based W-Band Suspicious Object Detection System for Moving Persons in the IoT Paradigm," *IEEE Access*, vol. 8, pp. 81378–81393, 2020, doi: 10.1109/ACCESS.2020.2991225.

[38] R. Fu, T. Wu, Z. Luo, F. Duan, X. Qiao, and P. Guo, "Learning Behavior Analysis in Classroom Based on Deep Learning," *10th Int. Conf. Intell. Control Inf. Process. ICICIP 2019*, pp. 206–212, 2019, doi: 10.1109/ICICIP47338.2019.9012177.

[39] L. Tian, H. Wang, Y. Zhou, and C. Peng, "Video big data in smart city: Background construction and optimization for surveillance video processing," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 1371–1382, 2018, doi: 10.1016/j.future.2017.12.065.

[40] G. Dai, "Deep learning method for citywide crowd flows prediction," *Proc. - IEEE Int. Conf. Mob. Data Manag.*, vol. 2019-June, no. Mdm, pp. 373–374, 2019, doi: 10.1109/MDM.2019.00-25.

[41] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 843–852, 2015.

[42] C. Di Huang, C. Y. Wang, and J. C. Wang, "Human action recognition system for elderly and children care using three stream ConvNet," *Proc. 2015 Int. Conf. Orange Technol. ICOT 2015*, pp. 5–9, 2016, doi: 10.1109/ICOT.2015.7498476.

[43] G. Botella and C. García, "Real-time motion estimation for image and video processing applications," *J. Real-Time Image Process.*, vol. 11, no. 4, pp. 625–631, 2016, doi: 10.1007/s11554-014-0478-y.

[44] J. Haapala, "Recurrent neural networks for object detection in video sequences," p. 58, 2017.

[45] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[46] N. Sanil, P. A. N. Venkat, V. Rakesh, R. Mallapur, and M. R. Ahmed, "Deep Learning Techniques for Obstacle Detection and Avoidance in Driverless Cars," *2020 Int. Conf. Artif. Intell. Signal Process. AISP 2020*, pp. 1–4, 2020, doi: 10.1109/AISP48273.2020.9073155.

[47] A. Burney and T. Q. Syed, "Crowd Video Classification Using Convolutional Neural Networks," *Proc. - 14th Int. Conf. Front. Inf. Technol. FIT 2016*, pp. 247–251, 2017, doi: 10.1109/FIT.2016.052.

[48] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," *2017 14th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2017*, 2017, doi: 10.1109/AVSS.2017.8078482.

[49] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li, "Deep Learning-Based Gait Recognition Using Smartphones in the Wild," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3197–3212, 2020, doi: 10.1109/TIFS.2020.2985628.

[50] A. Shahbaz, J. Hariyono, and K. H. Jo, "Evaluation of background subtraction algorithms for video surveillance," *2015 Front. Comput. Vision, FCV 2015*, 2015, doi: 10.1109/FCV.2015.7103699.

[51] Y. G. Jiang, S. Bhattacharya, S. F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimed. Inf. Retr.*, vol. 2, no. 2, pp. 73–101, 2013, doi: 10.1007/s13735-012-0024-2.

[52] C. Peng, S. L. Lo, J. Huang, and A. C. Tsoi, "Human Action Segmentation Based on a Streaming Uniform Entropy Slice Method," *IEEE Access*, vol. 6, pp. 16958–16971, 2018, doi: 10.1109/ACCESS.2017.2788943.

[53] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: A survey," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 13, no. 2, pp. 1–23, 2017, doi: 10.1145/3052930.

[54] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classification with convolutional neural networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 1725–1732, 2014, doi: 10.1109/CVPR.2014.223.

[55] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 1798–1807, 2015, doi: 10.1109/CVPR.2015.7298789.

[56] S. R. Sahoo, R. Dash, R. K. Mahapatra, and B. Sahu, "Unusual event detection in surveillance video using transfer learning,"

Proc. - 2019 Int. Conf. Inf. Technol. ICIT 2019, pp. 319–324, 2019, doi: 10.1109/ICIT48102.2019.00063.

[57] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Inf. Process. Syst.*, vol. 1, no. January, pp. 568–576, 2014.

[58] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, no. i, pp. 1933–1941, 2016, doi: 10.1109/CVPR.2016.213.

[59] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.

[60] A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences," *Expert Syst. Appl.*, vol. 139, p. 112829, 2020, doi: 10.1016/j.eswa.2019.112829.

[61] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," *Proc. IEEE Int. Conf. Comput. Vis.*, no. Section 2, pp. 1385–1392, 2013, doi: 10.1109/ICCV.2013.175.

[62] H. Nam and B. Han, "Learning Multi-domain Convolutional Neural Networks for Visual Tracking," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 4293–4302, 2016, doi: 10.1109/CVPR.2016.465.

[63] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," *Proc. - Int. Conf. Image Process. ICIP*, vol. 2017-Septe, pp. 3645–3649, 2018, doi: 10.1109/ICIP.2017.8296962.

[64] J. Y. H. Ng, J. Choi, J. Neumann, and L. S. Davis, "ActionFlowNet: Learning motion representation for action recognition," *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018*, vol. 2018-Janua, pp. 1616–1624, 2018, doi: 10.1109/WACV.2018.00179.

[65] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 5729–5738, 2017, doi: 10.1109/CVPR.2017.607.

[66] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," 2018, doi: 10.15607/rss.2018.xiv.019.

[67] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," *2016 Int. Conf. Biometrics, ICB 2016*, 2016, doi: 10.1109/ICB.2016.7550060.

[68] X. Luo, R. Shen, J. Hu, J. Deng, L. Hu, and Q. Guan, "A Deep Convolution Neural Network Model for Vehicle Recognition and Face Recognition," *Procedia Comput. Sci.*, vol. 107, no. Icict, pp. 715–720, 2017, doi: 10.1016/j.procs.2017.03.153.

[69] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, 2018, doi: 10.1016/j.patcog.2017.09.040.

[70] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, "A real-time object detection algorithm for video," *Comput. Electr. Eng.*, vol. 77, pp. 398–408, 2019, doi: 10.1016/j.compeleceng.2019.05.009.

[71] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos, H. Fujita, and F. Herrera, "Object Detection Binary Classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," *Knowledge-Based Syst.*, vol. 194, p. 105590, 2020, doi: 10.1016/j.knosys.2020.105590.

[72] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Underst.*, vol. 172, no. January, pp. 88–97, 2018, doi: 10.1016/j.cviu.2018.02.006.

[73] K. H. Cheong *et al.*, "Practical Automated Video Analytics for Crowd Monitoring and Counting," *IEEE Access*, vol. 7, pp. 183252–183261, 2019, doi: 10.1109/ACCESS.2019.2958255.

[74] Y. Tian *et al.*, "Lane marking detection via deep convolutional neural network," *Neurocomputing*, vol. 280, pp. 46–55, 2018, doi: 10.1016/j.neucom.2017.09.098.

[75] A. Gomez Villa, A. Salazar, and F. Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *Ecol. Inform.*, vol. 41, no. November 2016, pp. 24–32, 2017, doi: 10.1016/j.ecoinf.2017.07.004.

[76] N. Najva and K. E. Bijoy, "SIFT and Tensor Based Object Detection and Classification in Videos Using Deep Neural Networks," *Procedia Comput. Sci.*, vol. 93, no. September, pp. 351–358, 2016, doi: 10.1016/j.procs.2016.07.220.

[77] D. G. Lowe, "Object recognition from local scale-invariant features," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, pp. 1150–1157, 1999, doi: 10.1109/iccv.1999.790410.

[78] N. Li and F. Chang, "Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019, doi: 10.1016/j.neucom.2019.08.044.

[79] F. F. Pashchenko, O. S. Amosov, S. G. Amosova, Y. S. Ivanov, and S. V. Zhiganov, "Deep neural network method of recognizing the critical situations for transport systems by video images," *Procedia Comput. Sci.*, vol. 151, no. 2018, pp. 675–682, 2019, doi: 10.1016/j.procs.2019.04.090.

[80] O. S. Amosov, S. G. Amosova, Y. S. Ivanov, and S. V. Zhiganov, "Using the ensemble of deep neural networks for normal and abnormal situations detection and recognition in the continuous video stream of the security system," *Procedia Comput. Sci.*, vol. 150, pp. 532–539, 2019, doi: 10.1016/j.procs.2019.02.089.

[81] U. Shukla, A. Mishra, S. G. Jasmine, V. Vaidehi, and S. Ganesan, "A Deep Neural Network Framework for Road Side Analysis and Lane Detection," *Procedia Comput. Sci.*, vol. 165, pp. 252–258, 2019, doi: 10.1016/j.procs.2020.01.081.

[82] W. Luo *et al.*, "Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1070–1084, 2021, doi: 10.1109/TPAMI.2019.2944377.

[83] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 1971–1980, 2016, doi: 10.1109/CVPR.2016.217.

[84] J. M. Guo, P. C. Huang, and L. Y. Chang, "A hybrid facial expression recognition system based on recurrent neural network," *2019 16th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2019*, 2019, doi: 10.1109/AVSS.2019.8909888.

[85] N. R. Mennatullah Siam, Sepehr Valipour, Martin Jagersand, "CONVOLUTIONAL GATED RECURRENT NETWORKS FOR VIDEO SEGMENTATION Mennatullah Siam * , Sepehr Valipour *, Martin Jagersand , Nilanjan Ray University of Alberta," pp. 3090–3094, 2017.

[86] M. Murugesan and S. Thilagamani, "Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network," *Microprocess. Microsyst.*, vol. 79, no. September, p. 103303, 2020, doi: 10.1016/j.micpro.2020.103303.

[87] C. Gao, J. Yan, S. Zhou, P. K. Varshney, and H. Liu, "Long short-term memory-based deep recurrent neural networks for target tracking," *Inf. Sci. (Ny).*, vol. 502, pp. 279–296, 2019, doi: 10.1016/j.ins.2019.06.039.

[88] S. Gong, Y. Shi, and A. Jain, "Low quality video face recognition: Multi-mode aggregation recurrent network (MARN)," *Proc. - 2019 Int. Conf. Comput. Vis. Work. ICCVW 2019*, pp. 1027–1035, 2019, doi: 10.1109/ICCVW.2019.00132.

[89] A. Savakis and A. M. Shringarpure, "Semantic Background Estimation in Video Sequences," *2018 5th Int. Conf. Signal Process. Integr. Networks, SPIN 2018*, pp. 597–601, 2018, doi: 10.1109/SPIN.2018.8474279.

[90] M. J. Howard, A. S. Williamson, and N. Norouzi, "Video manipulation detection via recurrent residual feature learning networks," *Glob. 2019 - 7th IEEE Glob. Conf. Signal Inf. Process. Proc.*, 2019, doi: 10.1109/GlobalSIP45357.2019.8969458.

[91] M. Kalezic, P. Sekulic, and S. Kovacevic, "Video Object Segmentation using Optical Flow and Recurrent Neural Networks," *2020 9th Mediterr. Conf. Embed. Comput. MECO 2020*, pp. 8–11, 2020, doi: 10.1109/MECO49872.2020.9134313.

[92] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4694–4702, 2015, doi: 10.1109/CVPR.2015.7299101.

[93] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9912 LNCS, pp. 20–36, 2016, doi: 10.1007/978-3-319-46484-8_2.

[94] L. Wang *et al.*, "Temporal Segment Networks for Action Recognition in Videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, 2019, doi: 10.1109/TPAMI.2018.2868668.

[95] S. K. Yadav, A. Singh, A. Gupta, and J. L. Raheja, "Real-time Yoga recognition using deep learning," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 9349–9361, 2019, doi: 10.1007/s00521-019-04232-7.

[96] J. Donahue *et al.*, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, 2017, doi: 10.1109/TPAMI.2016.2599174.

[97] X. Wang, Z. Miao, R. Zhang, and S. Hao, "I3D-LSTM: A New Model for Human Action Recognition," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 3, 2019, doi: 10.1088/1757-899X/569/3/032035.

[98] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4724–4733, 2017, doi: 10.1109/CVPR.2017.502.

[99] S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, "3D-CNN-based fused feature maps with LSTM applied to action recognition," *Futur. Internet*, vol. 11, no. 2, 2019, doi: 10.3390/fi11020042.

[100] E. P. Ijjina and C. Krishna Mohan, "Hybrid deep neural network model for human action recognition," *Appl. Soft Comput. J.*, vol. 46, pp. 936–952, 2016, doi: 10.1016/j.asoc.2015.08.025.

[101] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 244–253, 2019, doi: 10.1109/CVPR.2019.00033.

[102] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, "Probabilistic Future Prediction for Video Scene Understanding," 2020.

[103] D. Xiaoxue, X. Chao, and Y. Quya, "A video gesture processing method based on convolution and long short-term memory network," *2019 IEEE 4th Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2019*, pp. 383–388, 2019, doi: 10.1109/ICCCBDA.2019.8725619.

[104] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," *Neurocomputing*, vol. 272, pp. 432–438, 2018, doi: 10.1016/j.neucom.2017.07.012.

[105] E. Duman and O. A. Erdem, "Anomaly Detection in Videos Using Optical Flow and Convolutional Autoencoder," *IEEE Access*, vol. 7, pp. 183914–183923, 2019, doi: 10.1109/ACCESS.2019.2960654.

[106] H. Jiang, Y. Lu, and J. Xue, "Automatic soccer video event detection based on a deep neural network combined CNN and RNN," *Proc. - 2016 IEEE 28th Int. Conf. Tools with Artif. Intell. ICTAI 2016*, pp. 490–494, 2017, doi: 10.1109/ICTAI.2016.78.

[107] H. Al Hajj, M. Lamard, P. H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," *Med. Image Anal.*, vol. 47, pp. 203–218, 2018, doi: 10.1016/j.media.2018.05.001.

[108] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-End Dense Video Captioning with Masked Transformer," *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 8739–8748, 2018, doi: 10.1109/CVPR.2018.00911.

[109] A. L. Mur, F. Peyrin, and N. Ducros, "Recurrent Neural Networks for Compressive Video Reconstruction," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2020-April, pp. 1651–1654, 2020, doi: 10.1109/ISBI45749.2020.9098327.

[110] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," no. November, 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.

[111] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition," 2011.

[112] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 961–970, 2015, doi: 10.1109/CVPR.2015.7298698.

[113] W. Kay *et al.*, "The Kinetics human action video dataset," *arXiv*, 2017.

[114] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv*. 2018, [Online]. Available: http://activity-net.org/challenges/2018/evaluation.html.

[115] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv*, 2019.

[116] S. Abu-El-Haija *et al.*, "YouTube-8M: A Large-Scale Video Classification Benchmark," 2016, [Online]. Available: http://arxiv.org/abs/1609.08675.

[117] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, 2014, doi: 10.1109/TPAMI.2013.111.

[118] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, vol. 2009 IEEE, no. 1, pp. 935–942, 2009, doi: 10.1109/CVPRW.2009.5206641.

[119] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2720–2727, 2013, doi: 10.1109/ICCV.2013.338.

[120] M. Kristan *et al.*, "The visual object tracking VOT2013 challenge results," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 98–111, 2013, doi: 10.1109/ICCVW.2013.20.

[121] [121] R. C. Agapito L., Bronstein M., "The Visual Object Tracking VOT2014 Challenge Results," *Lect. Notes Comput. Sci.*, 2015, [Online]. Available: http://votchallenge.net.

[122] M. Kristan *et al.*, "The Visual Object Tracking VOT2015 Challenge Results," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015-Febru, pp. 564–586, 2015, doi: 10.1109/ICCVW.2015.79.

[123] J. H. Hua G., "The Visual Object Tracking VOT2016 Challenge Results," *Lect. Notes Comput. Sci.*, [Online]. Available: http://votchallenge.net.

[124] M. Kristan *et al.*, "The Visual Object Tracking VOT2017 Challenge Results," *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-Janua, pp. 1949–1972, 2017, doi: 10.1109/ICCVW.2017.230.

[125] R. S. Leal-Taixé L., "The Sixth Visual Object Tracking VOT2018 Challenge Results.," *Lect. Notes Comput. Sci.*, vol. 11129, 2018, [Online]. Available: http://votchallenge.net.

[126] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," *Adv. Comput. Vis. Pattern Recognit.*, vol. 71, pp. 181–208, 2014, doi: 10.1007/978-3-319-09396-3_9.

[127] S. Blunsden and R. B. Fisher, "The BEHAVE video dataset: ground truthed video for multi-person behavior classification - https://homepages.inf.ed.ac.uk/rbf/BEHAVE/," *Ann. BMVA*, vol. 4, no. 4, pp. 1–11, 2010.

[128] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," *Proc. - Int. Conf. Image Process. ICIP*, vol. 2015-Decem, pp. 168–172, 2015, doi: 10.1109/ICIP.2015.7350781.

[129] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3354–3361, 2012, doi: 10.1109/CVPR.2012.6248074.

[130] N. D. Kalka *et al.*, "IJB – S : IARPA Janus Surveillance Video Benchmark ∗," 2014.

[131] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 529–534, 2011, doi: 10.1109/CVPR.2011.5995566.

[132] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," *IEEE 6th Int. Conf. Biometrics Theory, Appl. Syst. BTAS 2013*, 2013, doi: 10.1109/BTAS.2013.6712704.

[133] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 724–732, 2016, doi: 10.1109/CVPR.2016.85.

[134] L. Huang, W. Xu, S. Liu, V. Pandey, and N. R. Juri, "Enabling versatile analysis of large scale traffic video data with deep learning and HiveQL," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 1153–1162, 2017, doi: 10.1109/BigData.2017.8258041.