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

Analysis of Deep Neural Networks For Human Activity Recognition in Videos – A Systematic Literature Review

HADIQA AMAN ULLAH[1], SUKUMAR LETCHMUNAN[2], M. SULTAN ZIA[1], UMAIR MUNEER BUTT[1,2], FADRATUL HAFINAZ HASSAN[2]

[1]Faculty of Computer Science and Information Technology, The University of Lahore, Gujrat Campus, Lahore 50700, Pakistan
[2]School of Computer Sciences, Universiti Sains Malaysia George Town 11800, Malaysia

Corresponding author: Umair Muneer Butt, Sukumar Letchmunan (e-mail: umair@student.usm.my, sukumar@usm.my).

This work was supported by financial support from Universiti Sains Malaysia (USM) under FRGS grant number FRGS/1/2020/TK03/USM/02/1 and the School of Computer Sciences USM for their support.

ABSTRACT From the past few decades, Human activity recognition (HAR) is one of the vital research areas in computer vision in which much research is ongoing. The researcher's focus is shifting towards this area due to its vast range of real-life applications to assist in daily living. Therefore, it is necessary to validate its performance on standard benchmark datasets and state-of-the-art systems before applying it in real-life applications. The primary objective of this Systematic Literature Review (SLR) is to collect existing research on video-based human activity recognition, summarize, and analyze the state-of-the-art deep learning architectures regarding various methodologies, challenges, and issues. The top five scientific databases (such as ACM, IEEE, ScienceDirect, SpringerLink, and Taylor & Francis) are accessed to accompany this systematic study by summarizing 70 different research articles on human activity recognition after critical review. Human activity recognition in videos is a challenging problem due to its diverse and complex nature. For accurate video classification, extraction of both spatial and temporal features from video sequences is essential. Therefore, this SLR focuses on reviewing the recent advancements in stratified self-deriving feature-based deep learning architectures. Furthermore, it explores various deep learning techniques available for HAR, challenges researchers to face to build a robust model, and state-of-the-art datasets used for evaluation. This SLR intends to provide a baseline for video-based human activity recognition research while emphasizing several challenges regarding human activity recognition accuracy in video sequences using deep neural architectures.

INDEX TERMS Activity Recognition, Deep Learning, Computer Vision, SLR, Spatio-temporal features

I. INTRODUCTION

Distinguishing between various physical activities executed by humans to accomplish their daily living tasks refers to Human activity recognition (HAR). A HAR system can identify subject activities to provide authorities with valuable information to perform specific actions [1]. A variety of sensors are available for recording activities, including a variety of physiological activity sensors, ambient sensors, infrared motion detectors and magnetic sensors [2], RADAR [3], acoustic sensors [4], Echo, everyday objects, video cameras. Video-based HAR systems are popular due to their numerous real-life applications, but they also pose multiple privacy and environmental restrictions in smart environments.

Distinguishing between various physical activities executed by humans to accomplish their daily living tasks refers to Human activity recognition (HAR). A HAR system can identify subject activities to provide authorities with valuable information to perform specific actions [1]. A variety of sensors are available for recording activities, including a variety of physiological activity sensors, ambient sensors, infrared motion detectors and magnetic sensors [2], RADAR [3], acoustic sensors [4], Echo, everyday objects and video cameras. Video-based HAR systems are popular due to their numerous real-life applications, but they also pose multiple privacy and environmental restrictions in smart environments.

The objective of the HAR system is to identify real-life human activities and categorize them. Human activities

are highly complicated and diverse, which makes accurate activity recognition a challenge in computer vision.

Earlier studies in HAR systems consider activity recognition as a typical pattern identification problem [5]. Early HAR techniques were based on Support Vector Machine (SVM) and Hidden Markov models (HMM). Later research in this field has moved towards machine learning. The traditional techniques in machine learning, also known as shallow learning, involves heuristically driven feature extraction from data that mainly relies on human expert knowledge for a particular domain, limiting the architecture designed for one environment to surpass the problem of another area [6]. With the evolution of deep learning, handcrafted approximations are replaced as deep learning allows direct feature extraction from data, hence does not require any expert knowledge or optimal features selection.

Moreover, in traditional handcrafted-based techniques, correct classification is entirely dependent on accurate feature extraction. While in learning-based approaches, end-to-end neural network architectures are trained directly from unprocessed data like pixels to classification. Deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNN) are highly effective in learning complex activities due to their characteristics of local dependency and scale invariance [5].

A. HAR APPLICATIONS

A variety of sensors are available for recording activities, including a variety of physiological activity sensors, ambient sensors, infrared motion detectors and magnetic sensors [7], RADAR [8], acoustic sensors [9], Echo, everyday objects, video cameras. Video-based HAR systems are popular due to their numerous real-life applications, but they also pose multiple privacy and environmental restrictions in smart environments.

The task of assigning labels to actions is a topic of interest due to its number of applications in various fields. Applications of individual and group activity recognition comprise several areas like surveillance, medical, sports, entertainment industry, gaming, robotics, video indexing, video annotation, etc,. [10] . With the increasing usage of surveillance cameras, Network-based surveillance systems provide cooperative, instant observing which increases human throughput and performance [11]. Content-based video study and intuitive labeling of video clips lead towards improved searching [12]. In human-computer interaction, activity recognition can accommodate towards better natural language understanding that can help us in creating computers with improved speech recognition [13]. Finally, home care technologies can be developed with the ability to identify activities of daily living, decreasing the charges and burdens of care-giving with enhanced care and self-sufficiency in old age [14].

Smart home technology controls the incorporated lighting, heating, electrical, and all domestic components, as well as it can also recognize the activity of all home residents. Vision-based human activity recognition in smart homes has become a significant issue in terms of developing the next generation technologies which can improve healthcare and security of smart homes.

By taking advantage of automatic feature extraction and using large-scale datasets of deep learning methods special CNN architecture for extracting features following a sequence of convolutional and pooling layers has been proposed [15]. In this regard, a real human activity video dataset DMLSmartAcions has been used [16].

B. PROBLEM STATEMENT

Currently, there is much research exists in the domain of HAR. To improve accuracy on activity classification, understanding various preprocessing techniques, feature extraction and selection, fusion, and classification is an essential and complicated task. For every model developed for activity recognition, a dataset is needed for its validation. The empirical process of finding the most appropriate dataset and methods to overcome the challenges for improving HAR performance requires ample time from scholars at the beginning of their research. According to our observation, no study yet available on HAR, which inspects and reviews the latest video-based human activity recognition developments using deep neural architectures. This research aims to systematically analyze the recent literature and document evidence to answer our proposed research questions. For that reason, in this artefact, the latest HAR approaches, datasets, and developments for modelling and validation from the past five years are systematically examined and summarized to find the answers to the following Research Questions (RQ):

- RQ1: What are the various deep learning techniques currently supporting the discrimination of various human activities in video sequences?
- RQ2: What are the potential challenges emphasized in current studies to design a robust HAR architecture?
- RQ3: Which datasets are most familiar to the academic community regarding HAR in videos?

C. RESEARCH CONTRIBUTION

A Systematic Literature Review (SLR) is carried out to select and analyze the 70 research studies published from January 2015 to May 2020, focusing on deep learning architectures for modelling video data to answer the above-mentioned RQ's. This research is the first systematic literature review in video-based human activity recognition using deep neural frameworks. This study aims to provide recent advancements in human activity recognition. The scope of this research is only limited to video-based human activity recognition approaches implemented using deep learning. This paper does not consider RGB-D sensors or wearable sensors like actuators because they are not readily available in the real world and propose solutions specific to an environment. Moreover, upon studying literature we found an SLR on sensor-based activity recognition [17]. Besides, this study only considers standard publicly available, widely used

datasets for research. The main contribution of this research observed regarding this research field is:

- This research analytically explores and reviews the recent advances in deep learning architectures for HAR in video sequences. TA description and analysis of each technique used to model Spatio-temporal feature representation of video data are identified. This is the first effort to systematically analyze and summarize video-based human activity recognition developments using deep neural networks comprehensively to the best of our knowledge.
- This study identifies the most recognized datasets used to evaluate recognition accuracy in HAR, assessing the types of activities, subjects involved, background and situation where the data have been collected, type of camera, the duration of the capture of clips. Researchers/practitioners can benefit from this type of study to select a suitable dataset for evaluation as per their requirements.
- Finally, this study will also highpoints the substantial research gaps regarding video-based HAR, where enhancements are required to model and evaluate diverse human activities in real-life settings.

This SLR has been organized in the following sections: Section II describes related literature research, Section III explains the proposed methodology used for this systematic literature review. In Section IV, the bibliometric analysis is presented and, Section V presents the characterization and analysis for different techniques, challenges and their solutions, and validation datasets. Section VII presents the discussion about results analysis and, finally, Sections VIII presents the conclusions and future work, respectively.

II. RELATED WORK

The top five scientific databases, i.e., ACM, IEEE, ScienceDirect, SpringerLink, and Taylor & Francis were searched using the search string (Deep) AND (Neural Network) AND (Architecture OR Framework) AND (Human Activity Recognition OR Human Action Recognition OR Human Motion Recognition) AND (Video). A detailed search has been conducted but failed to find any SLR during 2015-2020 that focuses on Spatio-temporal activity recognition in videos using deep learning. Besides, we found 15 survey papers during the search process.

In computer vision, pose estimation refers to infer the gesture/position of a person/object of interest in an image or video. The traditional way to estimate pose is followed by identifying, locating, and tracking the various critical points on a given object or person. For humans, body joints like the elbow or knee are considered vital points. Human pose estimation is applicable in tacking and measuring human movement. Zhang et al. [18] publish a human pose estimation survey in 2015. This survey paper compares human pose estimation for colour images and depth data and discusses the limitations of developed approaches in both. This survey also discusses human pose estimation in object and action context. According to this survey, object detection and action recognition can serve as a mutual context for human pose estimation. In 2020, Chen et al. [19] analyzed 2D and 3D human pose estimation problems from monocular pictures or video sequences based on deep learning. Based on particular tasks, this paper summarizes papers published in 2014 and classifies them into four types: 2D single person pose estimation, 2D multi-person pose estimation, 3D single person pose estimation, and 3D multi-person pose estimation. This paper also highlights the issues and challenges, state-of-the-art approaches, standard datasets available for evaluation, and performance comparison based on evaluation metrics. After analyzing depth cameras, Li et al. [20] present a survey for 3D hand-pose estimation methods that include model-driven, data-driven, and hybrid methods. They also highlighted the limitations and standard benchmark datasets for 3D human pose estimation.

A comprehensive and updated survey for the newly designed methods presents a taxonomy to discriminate spatial and temporal representation to analyze 3D static and dynamic human data, considering many applications for the body and face [21]. For 3D body recognition, the framework usually follows, removing noise through preprocessing, lower-level description representations, and at large contraposition between handcrafted and learned features to capture spatial information and temporal dimension coherently. This paper categorizes spatial modelling as geometric, volumetric, topological, and marks based. For temporal modelling following dynamic methods are organized as trajectory-based, matrix-based, Markov Models, and Recurrent neural networks. This survey also discusses the limitations of each representation and provides exciting research directions.

Herath et al. [22] published an article in 2017 and reviewed several aspects of action recognition solutions. Starting with different definitions of actions, they provide a comparative analysis of handcrafted and deep learning based action recognition methods. Moreover, they discuss holistic representations/ global representation-based solutions and local features extraction (including Interest point detection, Local descriptors, and aggregations for local features). They also categorize deep learning architectures for action recognition, such as spatiotemporal, multiple streams, deep generative, and temporal coherency networks. This paper discusses nine video-based datasets and their complexity describing their contents relative to reality. It discusses thirty-one methods and their performance on seven challenging datasets to highlight improvements required for future researchers.

Zhang et al. [23] discussed action recognition in traffic context. This paper briefly named the traditional approaches used for activity recognition, including improved dense trajectories, HOF, HOG, MBH, HOF features and Fisher Vector encoding, and SVM classifier. For deep learning-based literature, this paper cited 13 papers and briefly highlights the developments of Two-Stream, C3D, and RNN based architectures. This paper does not discuss any dataset.

Khurana et al. [24] present a comparative analysis of

deep learning methods for HAR in video surveillance. They included feature extraction techniques, type of network used, methodology, dataset, classifier, and accuracy of 13 research articles published up to 2017. They consider those papers that utilized different kinds of datasets based on RGB videos, smartphone sensors like gyroscope, accelerometer, depth cameras, and live streaming videos.

Presti and Cascia [25] argued on the challenges and techniques regarding 3D skeleton and multimodal based action categorization, depth maps collection, and reviews recent skeleton body pose estimation approaches. This work describes the most common data preprocessing practices for biometric variances and several methods to deal with temporal variation. This paper categorizes 3D skeleton data features as joint-based descriptors, mined joint-based, and dynamic-based representations. This article also compares performance in publicly available benchmark datasets and proposes future research work.

Zhu et al. [26] presented a survey and focuses on representation techniques published from 2010 to 2016. This paper compares various aspects of feature representations and highlights their advantages and disadvantages. Primarily, Handcrafted features are categorized as spatial-temporal, volume-based, depth image, trajectory, and global approaches. Learning-based methods are reviewed for non-neural network-based systems, genetic programming, dictionary action representations, and neural network techniques. Besides, static frames, handcrafted features, and frame transformations are taken as input to networks like 3D CNN and hybrid models.

Han et al. [27] published a survey in 2017 and reviewed 171 3D skeleton representations from 150 papers published in five years. This paper briefly describes skeleton acquisition devices such as motion capture systems, time-of-flight sensors, and structures-light cameras and construction methods based on a single image and multiple images. This survey categorizes 3D skeleton-based human representations from four perspectives, including information modality, representation encoding, structure, and transition as displacement-based, orientation-based, raw joint based and multimodal presentations. This paper also classified encoding methods into three groups of concatenation-based, statistics-based, and bag-of-words features encoding. Dhillon et al. [28] reviewed the activity recognition trends with deep learning models. This paper examined various models based on two-stream networks, C3D and RNN, used for activity recognition. This survey considers activities presented as sequences of RGB camera images, i.e., videos, depth maps, and skeleton joints. It summarizes and comparatively analyzes 17 prominent techniques and highlights their advantages. This paper also presents a quantitative analysis by comparing the accuracy of state-of-the-art.

RGB-D data consists of three modalities: RGB, depth, and skeleton. Wang et al. [29] provide comprehensive coverage of challenges and deep learning-based methods for RGB-D motion recognition for the past five years. This paper defined a taxonomy organizing two groups, i.e., segmented and continuous/online motion recognition. It also analyzes different approaches and categorizes them based on modalities characteristics into four groups: RGB-based, depth-based, skeleton-based, and RGB+D-based. Additionally, several standard benchmark datasets are also surveyed. This survey paper's main applications of interest are gesture recognition, activity recognition, and interaction recognition. This survey also provides performance comparison, pros, and cons of these approaches from the perspective of Spatio-temporal encoding, and at the end, offers future research directions.

Stergiou et al. [30] analyzed human-to-human interaction in videos. Video data pose many challenges due to recording settings inconsistency, the subject's physical appearance, interaction duration, and interaction performance. Liu et al. [31] surveyed human action and interaction recognition approaches based on handcrafted and learning-based features based on the input modalities: depth-based, skeleton-based, and hybrid-based methods along with performance comparison. This survey also highlights practical challenges such as viewpoint and biometric variation, occlusion, various execution rates, and online adaption and their solutions in human activity analysis, along with potential future directions. The data-driven home automation development methods based on action detection using sensors have been analyzed [32], [33]. This paper proposes an integrated, personalized system that can create a customized dataset based on user preferences and feedback for target homes using both survey and transfer learning methods.

The scope of existing literature is significant, but with the advancement of research, we need survey papers that are more organized, specific towards the problem, data required for evaluation, application domain, and represent state-of-the-art solutions. A comparison of existing survey papers is given in Table 1. This research is the first systematic literature review in video-based human activity recognition using deep neural frameworks. This study aims to provide recent advancements in human activity recognition. Furthermore, this study organizes the literature by summarizing state-of-the-art papers published during the last five years. All of these characteristics make this study interesting for future researchers.

III. RESEARCH METHOD

Extensive research is being done in the field of HAR, ranging from shallow learning to deep learning. The key aim of this study is to inspect which approaches are working well over others. The scope of this study is to learn the impact of deep neural architecture for Spatio-temporal feature extraction for improved activity classification. To present the practical challenges of this domain and compare different datasets in the literature for architecture learning. This study is conducted by following the state-of-the-art guidelines [34]–[38]. This SLR on human activity recognition is the first attempt from 2015 to 2020 based on RGB visual sequences and deep neural architectures to the best of the author's knowledge.

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.2021.3110610, IEEE Access

TABLE 1: Classification based comparative analysis of state-of-the-art studies with the proposed SLR

Study	Objective	Methodology Analysis	Dataset Analysis	Challenges	Future Direction
[18]	Human Pose Estimation along with Object detection and action recognition context	✓	✓	✓	X
[20]	3D Hand Pose Estimation	✓	X	✓	✓
[19]	Deep learning based Human Pose Estimation	✓	✓	✓	X
[22]	Action Recognition	✓	X	✓	X
[23]	Action recognition in a traffic context	✓	X	X	✓
[24]	Deep learning based Human Activity Recognition in Video Surveillance	✓	X	X	X
[21]	3D Human Recognition	✓	X	✓	X
[25]	3D skeleton based Human Action Classification	✓	✓	✓	X
[26]	Human Action Recognition	✓	X	✓	X
[27]	3D skeletal human representation	✓	X	✓	✓
[28]	Deep learning approaches study for Human Activity Recognition	✓	X	X	X
[29]	RGB-D based Human Motion Recognition with deep learning	✓	X	✓	✓
[30]	Human-human interaction for Human Activity Recognition	✓	✓	✓	X
[31]	RGB-D based human action and interaction analysis	✓	X	✓	✓
[32]	Human Activity Recognition in Smart Homes	✓	X	X	✓
Proposed	Video-based Human Activity Recognition with deep learning	✓	✓	✓	✓

A. INFORMATION SOURCES AND SEARCH PROCESS:

Five databases are manually consulted to organize this study. The selected databases are:

- ACM
- IEEE
- ScienceDirect
- Springer
- Taylor & Francis

The search procedure is followed using two kinds of operators, i.e., AND, OR. The steps used for this search process are:

- Step 1: Research Question Identification
- Step 2: Title Definition for the problem domain
- Step 3: Find synonymous
- Step 4: Use of Boolean operators for searching
- Step 5: Database and parameters selection for articles searching
- Step 6: Search strings refinement
- Step 7: Papers selection according to selection and rejection criteria.

B. RESEARCH QUESTION:

Identification of common factors and methods that can influence the performance of HAR is essential. Therefore, the following research questions are constructed to find desire answers:

- RQ1: What are the various deep learning techniques currently supporting the discrimination of various human activities in video sequences?
  - What deep learning techniques currently exist since 2015 to model Spatio-temporal information to support activity recognition in videos?
  - What evaluation measures have been used in literature for validating those models/approaches?
- RQ2: What are the potential challenges emphasized in current studies to design a robust HAR architecture?
  - How the video data challenges for human activity recognition are identified from selected studies.
  - How state-of-the-art systems deal with these challenges.
- RQ3: Which datasets are most familiar to the academic community regarding HAR in videos?
  - In which scenario and context, the videos have been gathered, background environment, the number of subjects involved?
  - What are the critical characteristics of these captured activities, type of activities, clips duration?
  - What are the typical challenges represented by these datasets?

C. ARTICLES SELECTION

Selection and rejection criteria are defined logically to find the answers to our selected research questions. The selection and rejection rules are summarized below:

- The study must be an original research paper instead of a review/survey paper.
- Articles published from 2015 to 2020 are considered only to provide state-of-the-art insights for HAR. Research is conducted on papers published up to May 2020.
- Research studies are only considered from ACM, IEEE, ScienceDirect, Springer, Taylor& Francis, our selected well-reputed scientific repositories to perform this SLR.
- Only those research papers are considered that experiment with data captured using a single RGB camera to make this SLR provide insights for real-life applications.

TABLE 2: Results obtained after limitations applied on each search database.

Database	Limits	Returned Papers
ACM	Jan 2015-May 2020, Journals, Research Article	1076
IEEE	2015-2020, Conferences, Journals	105
Science Direct	2015-2020, Research Articles, Journals(Neurocomputing, Pattern Recognition, Journal of Visual Communication and Image Representation, Computer Vision and Image Understanding, Image and Vision Computing)	914
Springer	2015-2020, Article, English, Computer Science, Image Processing and Computer Vision	115
Taylor & Francis	2015-2020, Computer Science	74
Total		2284

- Only those research papers are considered where the author makes use of deep learning-based methods for action categorization.

D. SEARCH STRING:

Search strings are defined to find relevant articles from the literature. The resultant search string is as follows: **Analysis**: Question **OR** Reasoning **OR** Research **OR** Search **OR** Study **OR** Survey **AND Deep**: Broad **OR** Large-Scale **OR** Wide **AND Neural**: Interconnected **OR** Semantic **OR** Visual **AND Networks**: Net **OR** System **OR** Web **AND Architectures**: Framework **OR** Layout **OR** Structure **AND Human**: Body **OR** Child **OR** Individual **AND Activity**: Act **OR** Action **OR** Change **OR** Exercise **OR** Flow **OR** Motion **OR** Movement **OR** Operation **OR** Process **OR** Response **AND Recognition**: Detection **OR** Memory **OR** Perception **OR** Recall **OR** Understanding **AND Classification**: Organization **OR** Systematization **AND Videos**: Broadcast **OR** Program **OR** Recorded **OR** Taped

Only those terms are considered that can maximize relevant search outcomes. Only those research papers are considered that follow our research objective. We used a custom range of publications from 2015 to May 2020.

E. STRING REFINEMENT:

After string formation, the next step is to refine our search results returned from defined search repositories. The purpose of determining a string and searching through it is to find potential papers for this SLR. If the search string returns irrelevant or very few articles, it requires fine-tuning. This study also refined the search string after analyzing the results returned from the initial search string. We performed five iterations and analyzed the search results on each database before finalizing the search string. The search string refinement progression is shown in Fig. 1. The papers returned after applying filters with the final selected search string is shown in Table 2. Science Direct has a specific limitation that it does not support more than 8 Boolean operators.

F. STUDY SELECTION

The search process results in 2284 papers. Data abstraction and summarization template are developed to extract relevant materials for this SLR. In our first step, bibliographic information. In the first step, bibliographic information such as title and publication information is extracted and analyzed. In the first phase, 2073 papers were excluded based on title and abstract analysis. In the second step, the abstract and conclusion of filtered research papers are thoroughly analyzed to understand the problem addressed. It helps to shortlist relevant studies according to our scope, resulting in the exclusion of 82 more articles. In the third step, each paper's core details, such as proposed methodology, implementation details, and data required for validation, are identified. In this final stage, 129 articles are analyzed based on full text to fulfil our selection and rejection criteria, resulting in excluding other 59 papers, and 70 papers are selected for our study. Finally, we performed a detailed analysis to get RQ's answers given in the Introduction (Section I). The complete process of study selection is shown in Fig. 2.

Distinguished databases such as ACM, IEEE, ScienceDirect, Springer, Taylor & Francis are selected for our search to confirm the reliable outcome of this SLR because they are acknowledged worldwide. We try to consider only the latest articles, as shown in Fig. 3. Furthermore, we filtered studies that perform proper validation for their architectures. Table 3 summarizes the results of the database concerning the publications. For ACM, Science Direct, and Springer, we apply the filter to select only journals articles only as shown in Table 2, due to the many search results obtained and relevance to our research questions. Table 3 depicts the returned result of study selection with the count in each journal category. Type represents whether the selected research paper belongs to a conference or journal. References for each selected article is provided. The total number of journal conference chosen papers are given in the last column.

TABLE 3: Selected articles in each database.

Database	Type	References	Total Papers
ACM	Journal Conference	[39], [40] Nil	2 0
IEEE	Journal Conference	[41]–[45] [46]–[72]	5 27
ScienceDirect	Journal Conference	[73], [73]–[99] Nil	27 0
Springer	Journal Conference	[100]–[106] Nil	7 0
Taylor&Francis	Journal Conference	[107], [108] Nil	2 0

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.2021.3110610, IEEE Access

FIGURE 1: String refinement process and the results obtained for each search string in selected databases.

IV. BIBLIOMETRIC ANALYSIS

Bibliometrics is the use of statistical evaluation to analyze published books and scientific articles. It is used in the survey paper as an effective way to measure the influence of publication in the scientific community. After recording the bibliometric variables of 70 publications, they were quantified based on publication year, scientific database, and paper citations. The probability distribution of selected studies published per year is presented in Fig. 3. It clearly shows that this study demonstrated state-of-the-art as most of the papers are from 2019. And it also shows that with each passing year, research focuses more on deep learning techniques as the cope of this SLR is only limited to deep learning approaches. Fig. 4 shows the selection of our articles for the search database collected from each year.

The bibliometric analysis quantitatively assesses the academic quality using statistical methods such as citation rates. The sum of citations for our selected articles for each year is shown in Fig. 5. Fig. 6 compares the sum of the citations for each publication chosen from its published database with the sum of publications selected from that database with a line chart, the upper line showing the sum of citations and the lower line highlighting the count of papers. Fig. 7 illustrates the average citations of our selected articles concerning the published Database.

V. CHARACTERIZATION AND ANALYSIS

This section discusses the extracted results from the selected studies to answer the research questions after a detailed analysis. According to the relevant articles included in this study, this SLR tried to find answer of defined research questions.

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.2021.3110610, IEEE Access

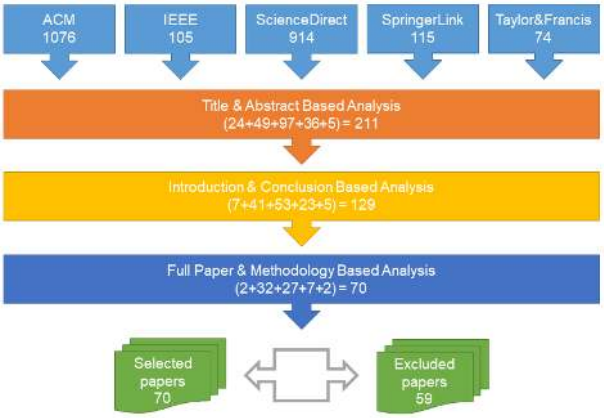

FIGURE 2: Study selection process summarization.

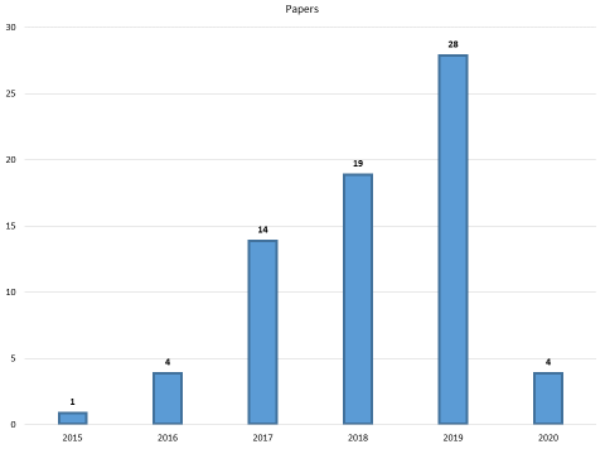

FIGURE 3: The trend in the number of publications by year.

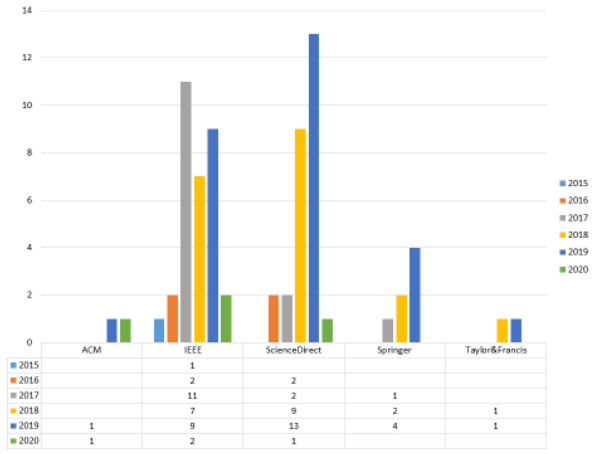

FIGURE 4: Detail distribution of selected articles from each repository according to its published year.

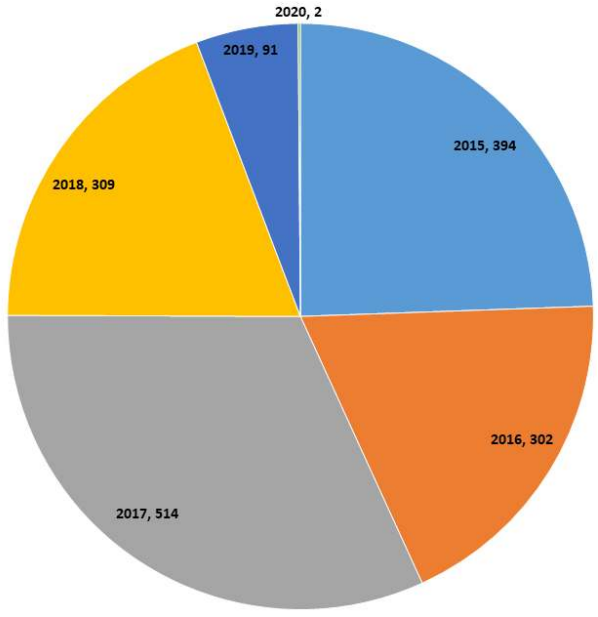

FIGURE 5: Sum of citations per year.

A. SPATIO-TEMPORAL HUMAN ACTIVITY RECOGNITION TECHNIQUES (RQ1(A))

The scope of this study is to explore deep neural networks for human activity recognition. Upon exploring 70 selected studies, we categorize deep neural architecture into eight types. These categories and their description is mentioned in table 4. This literature aims to provide new researchers with the crucial support for a better understanding of the recent approaches currently progressing for video-based human activity recognition. This paper highlights the most prominent current practices for researchers and beginners. This section discusses 70 state-of-the-art deep learning techniques. However, it is difficult to precisely answer which approaches are superior because every study tries to work on different parameters. We will answer that question by first analyzing each study's methodology and then highlighting the techniques that are used most frequently and have been more successful than others. Architecture details are extracted from each paper according to their architecture type as single or hybrid are organized and presented in Table 5 and 6 respectively. Fig. 8 presents the distribution for the type of architectures used by different studies. This figure depicts that to capture Spatio-temporal information, researchers focus more on 3D convolutional neural networks and two-stream CNN.

But all selected papers do not use only one architecture at a time. Most of the studies use a hybrid approach, which means using more than one architecture type in their systems. Fig. 9 shows the hybrid approaches in which we can see that authors have to combine even three methods in one architecture to take advantage of all of them.

B. EVALUATION MEASURES USED FOR VALIDATION (RQ1(B))

Most of the selected studies used accuracy as their performance metric, other evaluation measures used are recall and precision as presented in Table 7 and 8.

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.2021.3110610, IEEE Access

TABLE 4: Architecture type description.

Name	Description
CNN	A Convolutional Neural Network (ConvNet or CNN) takes structured arrays of data processing in a deep neural network. If, multiple layers stacked on top of each other in a sequence formed a multilayer feed-forward neural network. CNN layers are of four types: convolutional layer, pooling layer, fully-connected layer, and activation layer.
RNN	A recurrent Neural Network (RNN) is different from a traditional neural network because the output from the previous step is fed as input to the current step. In contrast, in a conventional feed-forward neural network, all the inputs and outputs are independent. RNN is designed to operate across sequences of vectors. The hidden state allows RNN to remember information about a series within a network using directed cycles.
LSTM	Long Short-Term Memory (LSTM) extends the memory of RNN by introducing an input gate, an output gate, and a forget gate in a single LSTM unit to capture long-term dependency.
BiLSTM	Bidirectional long short-term memory can learn long term dependencies in both directions between time steps of time series or sequence data. A BiLSTM takes an input sequence and calculates both a hidden forward sequence and a backward hidden sequence and later concatenates the forward and backward outputs into a single encoded vector.
C3D	In a 3D Convolution Neural Network (3D CNN or C3D or 3D ConvNet), the kernel moves through three dimensions of data (height, width, and depth) and produce a 3D activation map.
3D-DNN	In a 3D deep neural network (3D DNN), 3D arrays are input to the kernel representing height width and number of frames.
Two-Stream CNN	Two parallel streams are used to capture various features in a two-stream convolutional neural network (Two-stream CNN or two-stream ConvNets). Usually, spatial and temporal characteristics and later perform class score fusion or feature fusion to improve video classification tasks.
Multi-Stream CNN	An architecture having more than two streams are considered as multi-stream CNN.

TABLE 5: Spatio-temporal human activity analysis of individual classification techniques.

Study	Architecture Type	Network Input	Base Network	Classification
HAR techniques analysis for C3D architecture				
[46], [63], [70], [72], [79], [105]	C3D	RGB Frames	NA	Softmax
[45]	C3D	ResNet101, ResNet152, DenseNet169	Softmax	
[76], [102]	C3D	RGB Frames DT	NA	SoftMax
[95], [97]	C3D	RGB Frames	ResNet, ResNeXt	Softmax
[99]	C3D	RGB Frames	ResNet-50, I3D	Softmax
HAR techniques analysis for CNN and 3D DNN architectures				
[39]	CNN	RGB Frames Optical Flow Images	TSN, VGG16, DenseNet161	Average Score
[51], [100]	CNN	RGB Frames Optical flow images	VGG-16, ResNet-152	SoftMax
[55], [69]	CNN	RGB Frames VGGNet-21	SVM	NA
[74]	CNN	Action bank features	NA	NA
[90], [96]	CNN	RGB Frames ResNet-34	SVM	NA
[93], [108]	CNN	RGB Frames IDT	AlexNet	SoftMax + SVM
[104]	CNN	Binary space-time map (BSTMs)	NA	SoftMax
[53]	3D DNN	RGB Frames	GoogLeNet	SoftMax
HAR techniques analysis for Two-stream and Multi-stream architectures				
[47], [67]	Two-Stream CNN	RGB Frames Optical Flow Images	Merge predictions for context & saliency nets	Softmax
[49], [52], [94], [101]	Two-Stream CNN	RGB Frames Optical flow Images	VGG-16	SoftMax
[84]	Two-Stream CNN	RGB Frames Optical Flow Images	InceptionV1	SoftMax
[89]	Two-Stream CNN	RGB Frames WMHI	VGG-16, Flow-net	SoftMax
[98]	Two-Stream CNN	RGB Frames Optical Flow Images	TSN, DenseNet	SoftMax
[48]	Multi-Stream CNN	RGB Frames, Optical, Dyanamic Flow Images	ResNeXt-50 ResNeXt-101	SoftMax
[61]	Multi-Stream CNN	RGB Frames Optical Flow Images, Visual Rhythm	ResNet152, InceptionV3	SoftMax
[83], [85], [102]	Multi-Stream CNN	RGB Frames Optical Flow Images	SMAID, VGG16, ResNet152	SVM
HAR techniques for RNN, LSTM and BLSTM architectures				
[77], [92]	RNN	RGB Frames	Inception V1, V3	Softmax
[81]	RNN	DT	NA	Softmax
[81]	BLSTM	RGB Frames	NA	Softmax

C. POTENTIAL CHALLENGES TO BUILD A ROBUST MODEL FOR VIDEO-BASED HAR (RQ2)

The following challenges are identified from all the research articles that are essential to build a robust HAR model as shown in Fig. 10.

- Objective factors in visual appearance include jitter, camera motion, complex background, background mo-

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.2021.3110610, IEEE Access

TABLE 6: Spatio-temporal human activity analysis of hybrid classification techniques.

Study	Architecture Type	Network Input	Base Network	Classification
[50]	CNN + LSTM	Human Trajectories	VGG-16	Energy layer for estimating the energy of predictions using SoftMax layer.
[40], [56]	CNN + LSTM	RGB Frames Optical Flow Images	VGG-16	SoftMax
[59]	CNN + LSTM	RGB Frames	MobileNet	Softmax
[87]	CNN + LSTM	RGB Frames	GoogLeNet	SoftMax
[107]	CNN + LSTM	Dense Trajectory Motion Map (DTM)	NA	SoftMax
[44], [65], [80], [88]	Two-Stream CNN + C3D	RGB Frames, Optical Flow Images	NA	Softmax+SVM
[54], [58]	Two-Stream CNN + LSTM	RGB Frames Optical Flow Images	VGG-16	SoftMax
[75]	Two-Stream CNN + LSTM	RGB Frames, Optical Flow Images, iDT	GoogLeNet	SoftMax + SVM
[66], [68]	CNN + C3D	RGB Frames	DenseNet	Softmax
[41]	C3D + LSTM	Saliency aware clips Frames	NA	Softmax
[42]	C3D + LSTM	RGB Frames	VGGNet	Softmax
[62]	Multi-Stream CNN + C3D	RGB Frames of Human pose, Hand	NA	Softmax
[71]	Multi-Stream CNN + C3D	RGB Frames of Human pose, Hand	NA	Softmax
[57]	Multi-Stream CNN + LSTM	RGB Frames Optical, Flow Images, Trajectory Texture Image	GoogLeNet	SoftMax
[86]	Multi-Stream CNN + LSTM	RGB Frames Optical Flow Images	AlexNet	SoftMax
[43]	Two-Stream CNN + RNN	RGB Frames Optical Flow Images	GoogleNet	Softmax
[82]	Multi-Stream CNN + RNN	RGB Frames Optical Flow Images	TSN	SoftMax
[60]	Two-Stream CNN + C3D + RNN	RGB Frames	VGG-16	Softmax
[73], [91]	Two-Stream CNN + C3D + RNN	RGB Optical Flow Images	NA	Softmax
[78]	Two-Stream CNN + C3D + LSTM	RGB Frames, Optical Flow Images, MT, Video Segmentation	AlexNet	Softmax
[106]	Multi-Stream CNN + LSTM + C3D	RGB Frames, SemI, WOF, SemOF	NA	Softmax

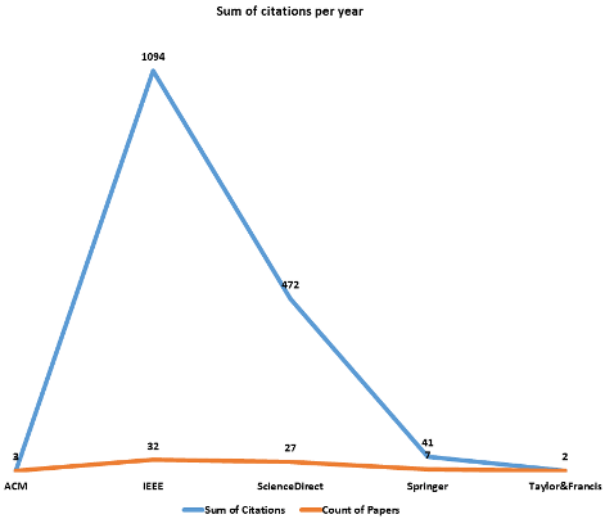

FIGURE 6: Citations vs article for each database.

tion, viewpoint variation, occlusions, dynamic background clutter, illumination, deformation, motion blur, video defocus, environmental and recording settings variations, and scale. To overcome these factors, different interaction strategies are required for appearance and motion streams along different hierarchies. S-TPNet [96] introduces a Spatio-temporal module to use essential multi-scale information into a hierarchical frame-level feature to merge high, mid, and low-level feature representations. Two-stream CNN [47] proposed calculating homography between two consecutive frames without human detection to remove global camera motion. Given the estimated homography due to camera motion, background motion can be canceled out from the warped optical flow. The transformation of pixels in a video clip can be observed as a blend of two types of movement: the global background motion and local foreground motion. The author introduces a weber motion history image (WMHI) which significantly reduces unwanted background noise [89]. This approach works in three stages: extracting pose information, extracting appearance features through VGG-16 and motion features from WMHI through flow-net, and later concatenating these features and feeding them into an artificial neural network for action classification. Using a variance-based algorithm, MFA [39] detects and approximates foreground motion from background motion considering background motion is typically produced by camera motion and is far slower and smoother than the

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.2021.3110610, IEEE Access

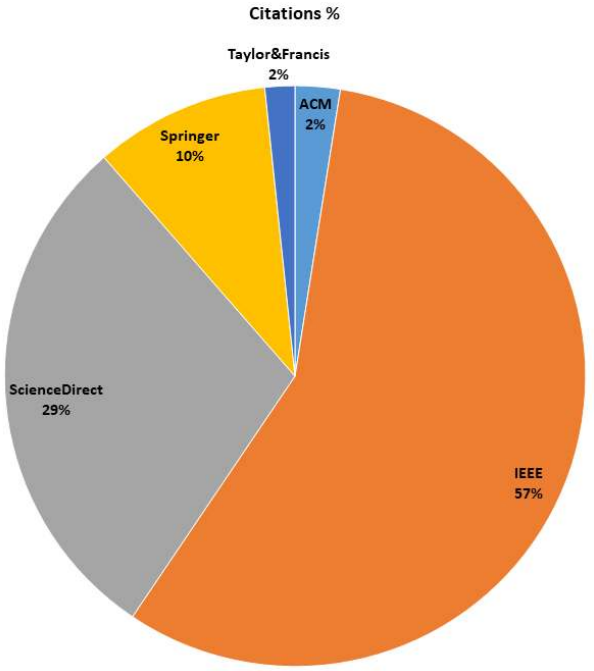

FIGURE 7: Citations per search repository.

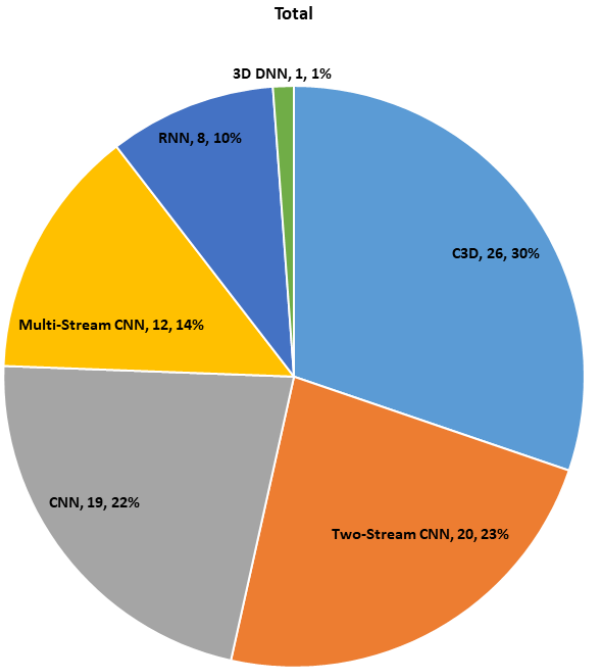

FIGURE 8: Publications by the architecture type.

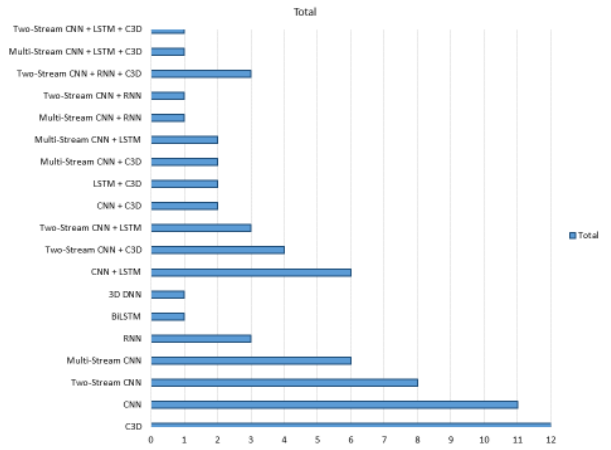

FIGURE 9: Publications by architecture type (including hybrid ones).

TABLE 7: Evaluation measures used for HAR techniques.

Performance Measures	Description
Accuracy, Multi-class Classification Accuracy (MCA), MAcc	Accuracy is the measure of how often the classifier is correct. Accuracy is the ratio of the number of correct predictions divided by the total number of predictions. $$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$
Precision	Precision is the measure of how often the model predicts yes when there is an actual yes. Precision is the ratio of true positives and total positives predicted. $$P = \frac{TP}{TP + FP} \quad (2)$$
Recall	A recall is a measure of how many times the model predicts yes when there is an actual yes. A Recall is essentially the ratio of true positives to all the positives in the ground truth. $$P = \frac{TP}{TP + FN} \quad (3)$$
Mean per-class Accuracy (MPCA)	Mean Per Class Accuracy (in Multi-class Classification only) is the average of the accuracy of each class in multi-class data set. It calculates the accuracy for each class, then takes the mean of that.
Mean Average Precision (mAP)	The average precision (AP) is a way to summarize the precision-recall curve into a single value representing the average of all precisions. To calculate the mAP, start by calculating the AP for each class. The mean of the APs for all classes is the mAP.

foreground and can be approximated independently of time.

- Subjective factors in appearance variation are caused by the complex relationship between participants, perspective, pose, contextual information, and behavior type diversity. It recognizes realistic, diverse group activities, human interactions, human-object interaction, complex and co-occurring individual actions, and object appearance, including hard-to-detect objects. Co-existing activities of individual humans referred to as group activities, which are challenging to understand due to complicated relationships between humans. Concerning the local motion, the Discriminative group context feature

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.2021.3110610, IEEE Access

TABLE 8: Performance metrics used for validation.

Performnace Metric	Study
Accuracy	[39]–[61], [63]–[72], [74]–[82], [84]–[100], [102]–[108]
Multiclass Classification Accuracy, Mean Perclass Accuracy	[49], [50]
Mean Average Precision	[62], [83], [101], [102]
Recall, Precision	[69], [85]
Mean Accuracy	[73]

(DGCF) [81] studies influential sub-events that create misperception within different group activities. Another research on multi-person activity recognition proposes using a two-level attention mechanism, including individual and scene levels using two-stage GRUs [77].

- Intra-class and inter-class activity variation is another major challenge. Motion dynamics, including motion speed, action length variation, leads towards an intra-class variation. For understanding temporal sequences in an unconstraint environment and temporal inference, extracting and representing high-level motion patterns, modeling image sequence dynamics for the progression in the appearance is vital. A sequential deep trajectory descriptor (SDTD) [57] is proposed that represents relative action by extracting dense trajectories from multiple consecutive frames and then anticipate them onto a two-dimensional plane called Trajectory Texture image. By extraction of informative global features at inter-frame and intra-frame levels, a Spatio-temporal deformable 3D convolution module with an Attention mechanism (STDA) [103] can capture long-term dependencies. In real-world human activity, data participants vary in activity lengths; the GRU model [81] can handle arbitrary length of video clips and can learn temporal changes in a sequence.
- Due to the neighboring frames' redundancy, treating every video clip frame as input is time-consuming and inefficient in video activity data. Traditional techniques to solve this problem are to use random or uniformly sampled frames as input. However, these methods' issue is that different classes' actions may not have the same motion distribution or temporal extents. Moreover, the action clips from the same type can also have different performing speeds. Besides, in different action samples, the performance may happen across different temporal scales. Based on IDT, Two-stream CNN [52] proposes video dynamics mining schemes to extract temporal information by calculating the motion intensity of each frame of a video sequence and developing action data features. It introduces an n-skip optical flow extraction method at different temporal scales. The video comprises frames that lead to frame redundancy challenges, focusing on discriminative foreground targets and temporal attention while focusing selectively on the informative frames and order-aware frame-level representations. In a single temporal scan of a video, AdaScan [49] can reject most unnecessary frames by using an algorithm that can predict each frame's discriminative importance and subsequently pools comparative and descriptive frames. Two-stream attention [54] introduces a novel temporal attention model that dynamically adjusts per-frame features weights of a video through the iterations of RNN, referring to the importance of the current frame.
- For using deep learning for human activity recognition, deep neural networks possess a large number of parameters with high computational costs to process video data. Therefore to implement a video understanding method on end terminals in the real-world is a challenge. For implication on terminal computers, DEEPEYE [40] is a deeply tensor-compressed video comprehension neural architecture that extracts 8-bit quantized features and uses structured, tenderized time-series features. This is done using tensor decomposition to simplify low-rank matrix to high-dimensional data (tensors), resulting in a higher compression ratio. To reduce computation power and provide faster annotation [51] introduces an algorithm to distribute feature representations extracted from frames and labels statistically is assigned without further processing if the confidence value is calculated from the log-likelihood ratio is significant.
- Video data requires optimal, efficient and complete, discriminative and compact video representation. By utilizing the concept of rank pooling video, Bilen et al. [48] proposed a method to summarize video dynamics to a dynamic image that encodes temporal video representation in addition to appearance features. It helps neural architectures pre-trained on image data to be extended to video data. Hierarchical rank pooling networks [100] encode a video sequence at multiple hierarchies using rank pooling-based temporal pooling to acquire discriminative video dynamics. CATNet [99] constructs video with two-level attention, i.e. local and global level. The extraction of segmented features learns local features in a self-attention manner by learning multiple attention weights for each segment. It integrates them into multiple global representations by the multi-head attention module. To construct the final global representation, it models the relations between these global features. SemI [106] is an improved video representation to model action-motion dynamics, obtained by applying localized sparse segmentation using global clustering before approximate rank pooling. It summarizes the motion characteristics in single or multiple images and overlays a static background from the window onto the subsequent segmented frames.
- Sometimes massive learning framework leads to an overfitting problem. For trajectory data, DGCF [81] introduces an augmentation method that works by locating subjects in the frames and shifting them to randomly eight directions to reduce the overfitting problem. The

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.2021.3110610, IEEE Access

data augmentation method is used to keep the semantic labels of the data.

- Small datasets limit observation capacity. To overcome the limited training data and RGB-data only [64], it introduces an end-to-end framework based on BLSTM. A variety of data augmentation techniques are used to increase the training size while avoiding overfitting, including Dynamic Frame Dropout, regularization techniques, and the vanishing gradient issue.
- Open-set action recognition refers to train a model train with incomplete knowledge of the data at training time so that at the testing phase, it can classify unknown categories. To comply open-set, a collection of informative feature representations that describe actions are defined in an action dictionary [103]. After extracting Spatio-temporal features representation via K-means clustering, the matching similarity between sample representation and action dictionary components is calculated for action recognition. Based on Euclidean distance for assigning a new class, if no class is present, it generates a new action class in the assembled dictionary if required.

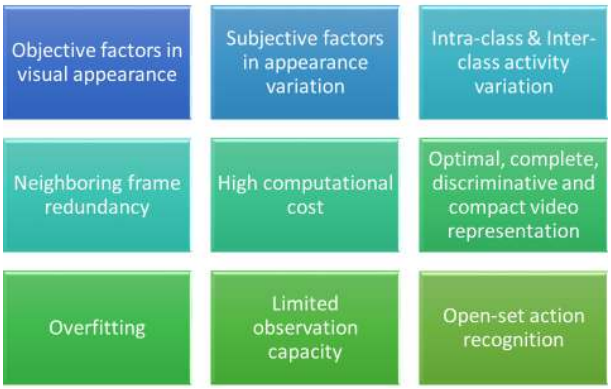

FIGURE 10: Categorization of Various challenges posed by all 70 articles.

D. RECOGNIZED DATASETS FOR THE EVALUATION OF HUMAN ACTIVITY RECOGNITION TECHNIQUES (RQ3)

The following 20 datasets have been used in the evaluation process for our selected articles. Characteristics of these datasets are mentioned in the Table 9 and count of each dataset in various studies is shown in Fig. 11.

VI. DISCUSSION

This section highlights the significance of an architecture type suitable for a particular dataset. Moreover, the most prominent techniques in HAR in terms of performance are discussed thoroughly. We critically analyze the various state-of-the-art techniques and reported datasets. In some scenarios, we cannot compare the methodology and evaluation measure of selected studies because they may have addressed the same problem, but their contexts and purposes are different. Furthermore, this SLR addresses the technique selection

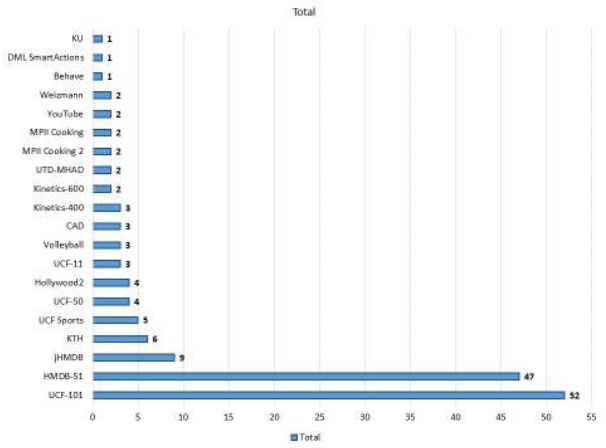

FIGURE 11: Count of each dataset used in selected studies for evaluation.

problem in a particular scenario on a specific dataset to assist researchers in selecting architecture and evaluation processes. The results obtained by critically analyzing the selected studies are shown in Table 10.

Fig. 12 to Fig. 26 compares best performing architectures on different datasets. After extracting features through GoogLeNet, 3D DNN feed them as 3-dimensional characteristics, i.e. height × width × number of frames and achieves the state of the art results on KTH and UCF Sports [53]. The maximum performance of various networks on KTH and UCF Sports is compared in Fig. 12 and Fig. 13, which shows that it also achieves high performance with LSTM based networks.

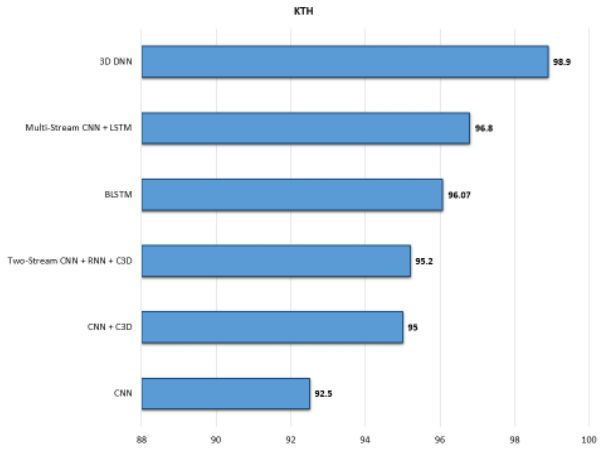

FIGURE 12: Maximum performance of each network on KTH.

Weber Motion History Image (WMHI) is a new motion estimation technique for pose based HAR to reduce unwanted background motion. WMHI outperforms available motion techniques of optical flow and motion history image and provides state-of-the-art results on the MPII Cooking dataset and jHMDB [89] as compared in Fig. 14 and Fig. 15.

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.2021.3110610, IEEE Access

TABLE 9: Spatio-temporal human activity classification techniques analysis.

Dataset	Description	Challenges	No. of videos	Duration	No. of classes	Video per class
KTH [109]	Captured using 25 subjects in four different scenarios (outdoor, with scale variation, with different clothes, and indoor)	Variation in scale, frequency, and velocity of the pattern. Homogeneous backgrounds with a static camera. Complex non-stationary background, intra-class similarity	2391	4s in average	6	10
Weizmann [110]	Outdoor video recording on still backgrounds	Partial occlusions, non-rigid deformations, imperfections in the extracted silhouettes, and high irregularities	90	NA	10	9
UCF-Sports [111]	Television broadcasts for sports from channels, e.g., BBC, ESPN (780 *480)	NA	150	6.39s	10	6-22
Hollywood2 [112]	Clips taken 69 Hollywood movies divided into 33 training and 36 testing clips labeled based on a movie script.	Physical properties of scenes, high correlation	3669	NA	12	33-146
UCF-11 [113]	Collected from YouTube, TV Broadcasts, and personal videos.	Variations in camera motion, object appearance, and pose.	1160	NA	11	100
YouTube [113]	YouTube and personal video collections.	A mix of steady and shaky cameras with variations.	1168	NA	11	100
Behave [114]	Multi-person behavioral interactions (2 to 5 people interactions)	NA	4	NA	10	
HMDB-51 [115]	Digitized movies, public databases, other videos available on the internet, Five types of actions are captured.	Camera viewpoint and motion variations.	6766	2-3s	51	101+
UCF-101 [116]	Web videos of five types of actions: Human action, Body-Motion, Human Interaction, Music, and Sports.	Web videos recorded in unconstrained environments .	13320	2-10s	101	100+
MPII Cooking [117]	Continuous Recordings in realistic settings with 12 subjects.	Low inter-class variability and high intra-class variability.	44	3-41min	65	NA
CAD [118]	Atomic activities, interactive activities, collective activities.	Tracking multiple people and estimating their collective activities	32	NA	6	NA
jHMDB [119]	Annotation of human Joints for the HMDB dataset.	Variation in poses, human sizes, camera motions, and visibility.	NA	NA	21	36-55
KU [81]	Group activity recognition. Daily outdoor scenes.	NA	88	NA	5	NA
UCF-50 [120]	Web videos.	Random camera motion, poor lighting conditions, clutter, scale variation, appearance, viewpoints, and action of interest.	5000+	NA	50	100+
Volleyball [121]	YouTube volleyball videos. Group activity recognition.	Seven player action labels, six-team activity labels.	NA	15	NA	13
Kinetics-400 [122]	YouTube videos about Person Actions.	camera motion/shake, illumination/background variations.	306245	10s	400	400-1150
Kinetics-600 [123]	YouTube videos about Person Actions.	camera motion/shake, illumination/background variations.	495547	10s	600	600
UTD-MHAD [124]	Four temporally synchronized data modalities captured from 8 subjects using a Kinect camera .	Intra-class variations due to subjects action.	861	NA	27	NA
MPII Cooking 2 [125]	Fine-grained activities and composite activities Video recordings of 30 human subjects.	Low inter-class variability and high intra-class variability.	273	1-41min	67/59	NA
DML SmartActions [126]	Include a variety of objects, scenes, and actions for smart home applications.	Temporal correlation between different actions.	932	NA	12	NA

Multi-steam I3D network extracts RGB video frames, optical flow, human pose, and hand features and feed them to I3D network and achieve mean average precision up to 54.1 on MPII Cooking 2 dataset [62]. The discriminative

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.2021.3110610, IEEE Access

TABLE 10: Superior techniques for video-based human activity recognition.

Study	Architecture Type	Dataset	Result
[53]	3D DNN	KTH	98.9
[62]	Multi-Stream CNN + C3D	MPII Cooking2	54.1
	CNN	Hollywood2	76.7
[100]			
[96]	CNN	Kinetics-600	79.8
[86]	Multi-Stream CNN + LSTM	Kinetics-400	80.94
[89]	Two-Stream CNN	MPII Cooking	83.8
[81]	RNN	KU	84.51
[71]	Multi-Stream CNN + C3D	HMDB-51	87.7
[69]	CNN + C3D	DML SmartActions	87.77
[89]	Two-Stream CNN	jHMDB	88.91
[89]	Two-Stream CNN	sub-JHMDB	90.48
[64]	BLSTM	UTD-MHAD	90.95
[77]	RNN	Volleyball	91.2
[81]	RNN	CAD	91.25
[58]	Two-Stream CNN + LSTM	UCF-11	94.6
[59]	CNN + LSTM	YouTube	95.8
[81]	RNN	Behave	98.4
	CNN + LSTM + C3D	UCF-101	99.1
[106]			
[74]	CNN	UCF-50	99.98
	CNN	Weizmann	100
[104]			
[53]	3D DNN	UCF Sports	100

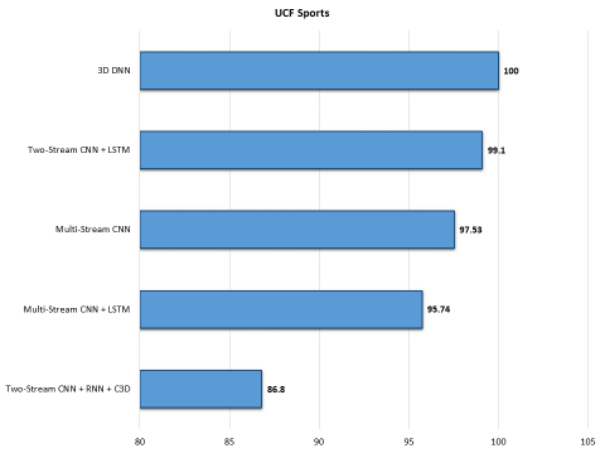

FIGURE 13: Maximum performance achieved by each network on UCF Sports.

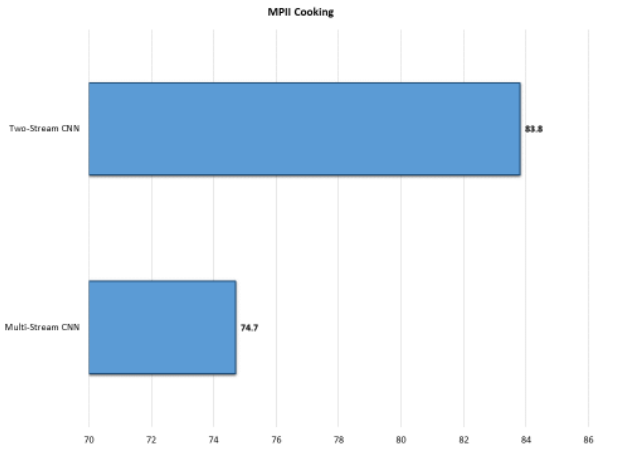

FIGURE 14: Maximum performance of two-stream and multi-stream CNN on MPII Cooking dataset.

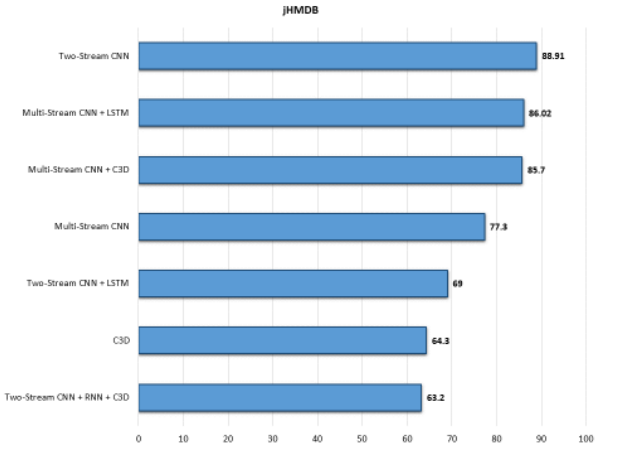

FIGURE 15: Comparison of different architecture types on JHMDB.

hierarchical rank pooling network [100] involves an effective temporal pooling layer that can be applied over any CNN architecture. For effective end-to-end learning of informative dynamics, the rich frame-based video representations achieve accuracy up to 76.7 for the Hollywood2 dataset, as shown in Fig. 16. Three stream CNN [86] achieves state-of-the-art performance on Kinetics-400 by following four main steps. First, by considering appearance, motion representations and extract salient patches. Second, obtain abstract features using three-stream CNN. Third, consider the temporal relations of frame-level descriptors, integrate them and input them to an RNN. Fourth, fuse the classification score of three streams and predict the final class label. S-TPNet [96] proposed a Spatio-temporal module that integrates multi-level information of CNNs into a hierarchical frame-level representation. The spatial pyramid module extracts multi-scale appearance features of video frames. The temporal pyramid module exploits several frame poolings to get different time-grained representations of snippets. By accumulating snippet-relations for comprehensive prediction, it achieves accuracy up to 80.94 on Kinetics-600. S-TPNet [96] proposed a Spatio-temporal module that integrates multi-level features of CNNs into a hierarchical frame-level representation. The spatial pyramid module extracts multi-scale appearance features of video frames, while the temporal pyramid module exploits several frame poolings to get different time-grained representations of snippets. By integrating snippet-relations for summarizing predictions, it achieves accuracy up to 80.94 on Kinetics-600. Fig. 17 shows that apart from multi-stream CNN and LSTM, CNN and 3D convolutional networks also perform well on the Kinetics dataset.

Trajectory-based features [71] are extracted through the Discriminative group context feature to discriminate the ac-

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.2021.3110610, IEEE Access

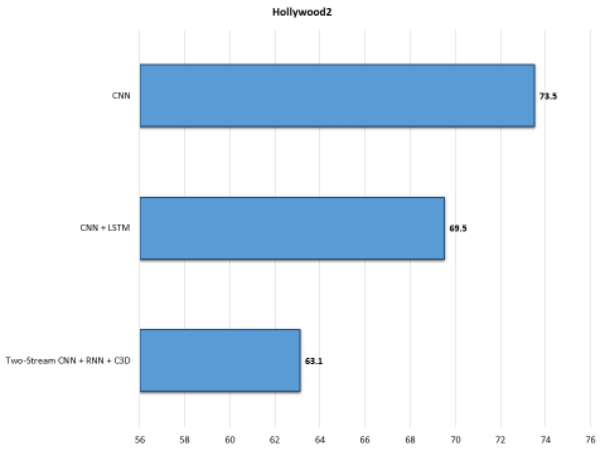

FIGURE 16: Comparison of top-performing architectures on Hollywood2 dataset.

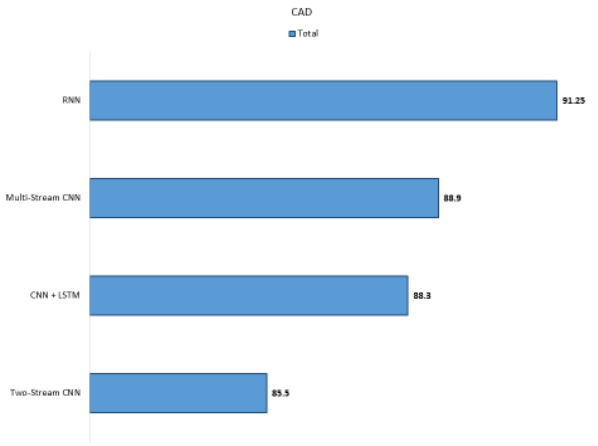

FIGURE 18: Comparison of maximum performance on Collective Activity dataset.

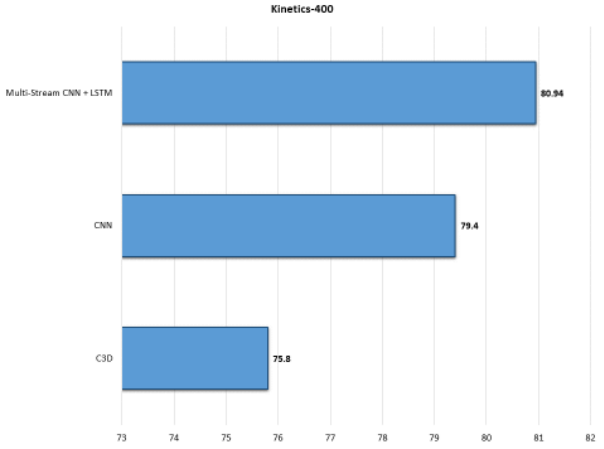

FIGURE 17: Maximum performance of each architecture type on Kinetics-400 dataset.

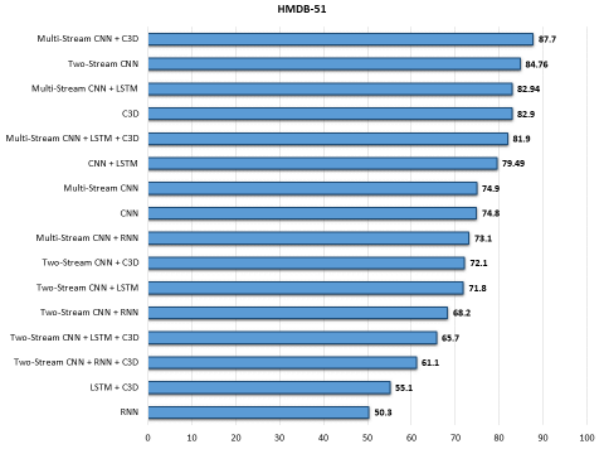

FIGURE 19: Comparison of maximum performance on HMDB-51.

tivities. After that, individual activity and sub-event based representations are merged, and RNN architecture concludes what happens in a scene by capturing the changes over time in a sequence. This approach achieves state-of-the-art performance on KU, Behave, and collective activity dataset [81], as shown in Fig. 18.

3D ResNet takes motion patches and scene patches extracted from RGB images as input to attain patch features. Based on the 2D features of RGB images and 3D features of valued patches, video descriptors are generated. Experiments show that our 2D + 3D multi-stream framework exploiting valued patches outperforms previous fusion methods and achieves accuracy up to 87.7 on the HMDB-51 dataset [71] as shown in Fig. 19.

State-of-the-art results on the DMLSmartActions dataset are obtained using a CNN-based approach for extracting features following a sequence of convolutional and pooling layers [69]. To achieve superior performance on RGB-only data, skeleton key-points are extracted from RGB videos and feed to a deep BLSTM [64]. The MLP framework consists of five consecutive BLSTM layers with dropout to regularize the model training and achieve the state-of-the-art result on UTD-MHAD as shown in Fig. 20.

Lihua et al. [77] propose a two-level attention-based interaction model. It works by first considering interaction at the individual level, i.e. pose based interaction among individuals and later at the scene level to exploit high-level activity while updating their states at each time step. The scene-level attention mechanism based on a pooling strategy explores various levels of interactions. Therefore, it achieves state-of-the-art results on the Volleyball dataset for multi-person activity recognition, as shown in Fig. 21.

Amin et al. [59] introduce a framework for HAR in which continuous video streams are first divided into essential shots. Based on human saliency features, important shots are selected using the proposed CNN. For optical flow to represent temporal features FlowNet2 is utilized. Finally, a multilayer LSTM is used for learning long-term sequences in temporal

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.2021.3110610, IEEE Access

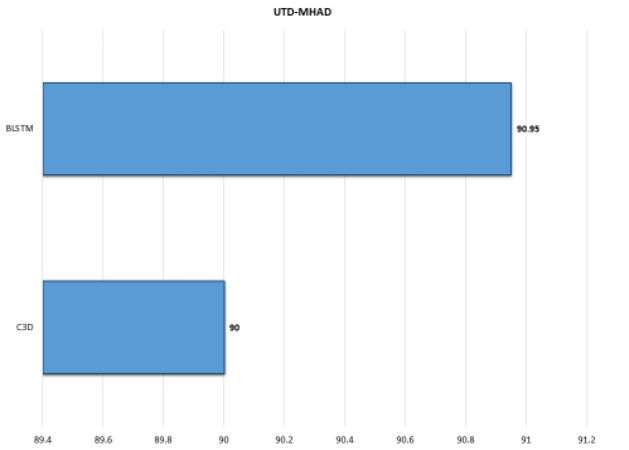

FIGURE 20: Comparison of maximum performance on UTD-MHAD.

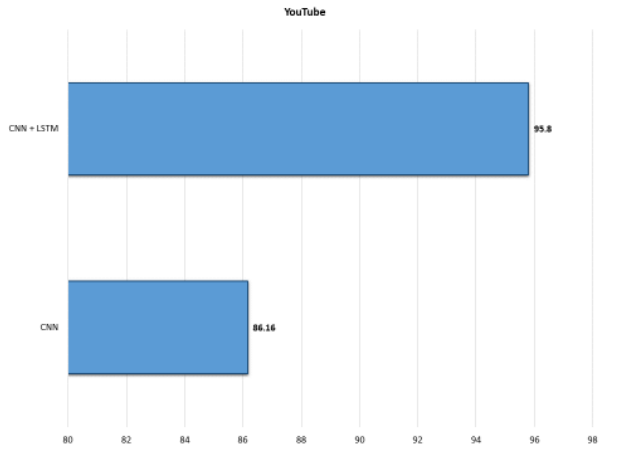

FIGURE 22: Comparison of maximum performance on YouTube dataset.

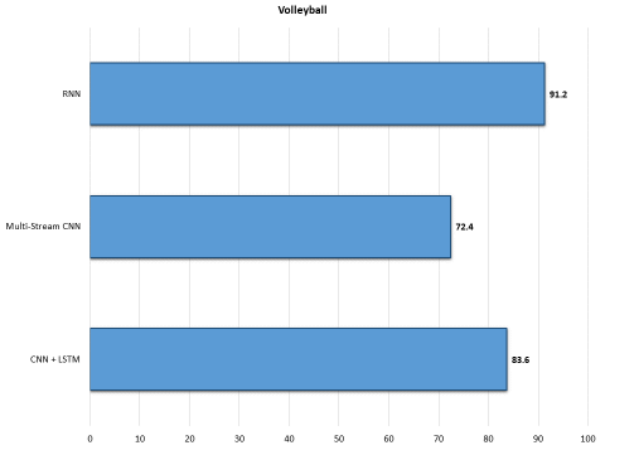

FIGURE 21: Comparison of maximum performance on Volleyball dataset.

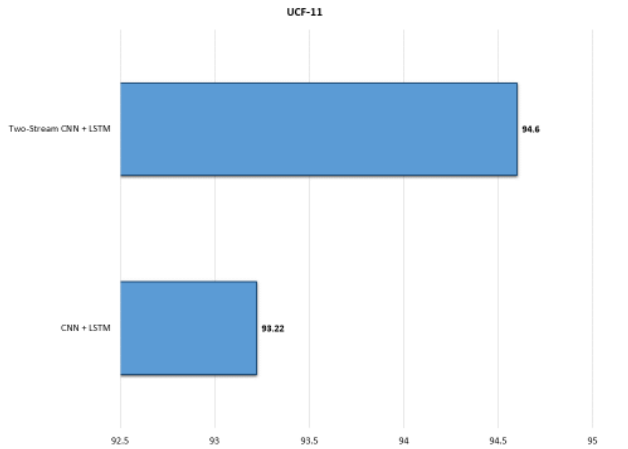

FIGURE 23: Comparison of maximum performance on UCF-11 dataset.

direction. Fig. 22 shows that this approach achieves state-of-the-art result on the YouTube dataset.

Harshala et al. [58] integrate spatiotemporal features extracted through CNN and LSTM. They map them effectively to act as an attention mechanism to concentrate the LSTM towards informational patches of the convolutional feature space and achieve accuracy up to 94.6 on UCF-11, as shown in Fig. 23.

Earnest et al. [74] improve classification performance by modelling a CNN classifier as a GA-chromosome and by integrating genetic algorithms with CNN . They also explore different weight initializations using its global search capability and use gradient descent algorithm local search capability to find a closer to global-optimum and verified its improved classification performance on UCF-50 as shown in Fig. 24.

SR-LSTM [106] achieve comparable performance on UCF-101 by proposing a semantic image to improve the representation of video action-motion dynamics. Their approach first divide videos into overlapping frames and then

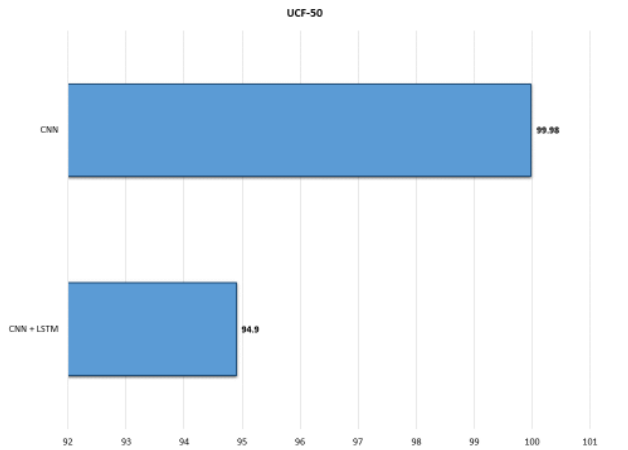

FIGURE 24: Comparison of maximum performance on UCF-50 dataset.

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.2021.3110610, IEEE Access

perform segmentation with localized sparse segmentation using global clustering (LSSGC) and shows highest results in comparison as shown in Fig. 25. To handle the sequential modelling of data, they also propose the sequential combination of Inception-ResNetv2 and LSTM to leverage the temporal variance incorporated in SemI.

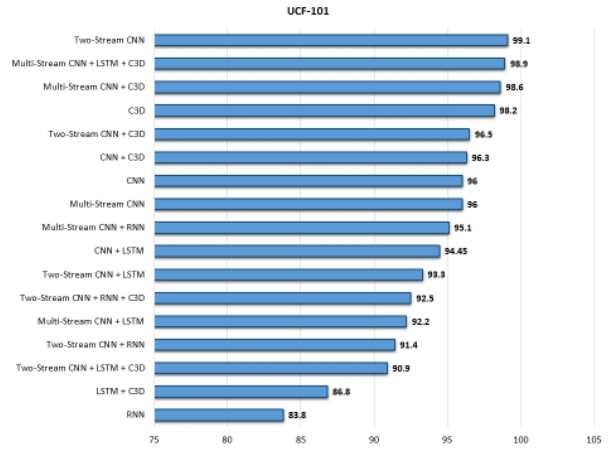

FIGURE 25: Comparison of maximum performance on UCF-101 dataset.

A binary space-time map (BSTM) [104] summarizes human activity within a defined time interval obtained by detecting and tracking the human body after background subtraction. The human body is segmented in each frame by thresholding the images using an optimum threshold. Finally, they extract Binary space-time map (BSTM) information of the human silhouettes in a lap of time. Classification is performed using CNN on BSTMs to achieve ideal results on Weizmann, as shown in Fig. 26.

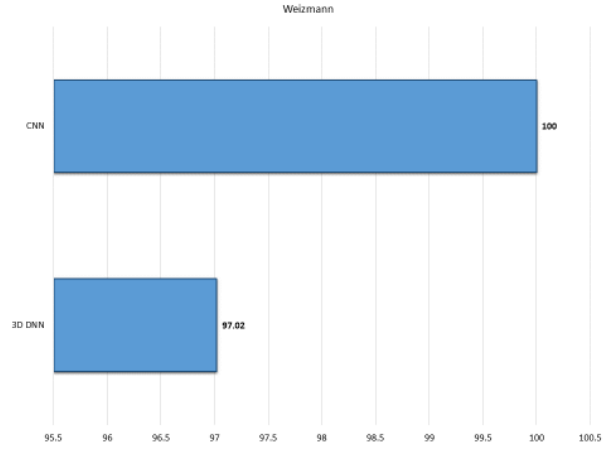

FIGURE 26: Comparison of maximum performance on Weizmann dataset.

VII. CONCLUSION

The primary objective of this SLR is to provide a baseline for beginners to assist in human activity recognition research. In HAR, activities are captured using various sensors, but this research is limited to video-based activity recognition due to its broad applicability in the real world. Moreover, this study focused on identifying suitable deep neural architectures for HAR by exploiting the state-of-the-art deep learning paradigm. This SLR is conducted by following the guidelines from papers [34]–[36], [38], [60]. After considering the selection and rejection criteria, 70 articles are selected from Jan 2015 to May 2020 from five scientific databases, including ACM, IEEE, Science Direct, Springer, Taylor & Francis. The development of research in this area is increasing continuously with each passing year, as shown in Fig. 3. Most relevant papers are found in IEEE and Science Direct, as shown in Fig. 4. All selected articles are critically analyzed to answer the proposed research questions.

To answer RQ1, all studies are investigated to understand and summarize the approaches used to extract Spatio-temporal information that correctly categorizes activities from every video clip. In most articles, features are extracted to represent motion in video data before inputting to neural architecture for final classification. Apart from RGB Images, Optical Flow Frames, Improve Dene Trajectories, Dynamic Image, Saliency aware clips frames, Trajectory Texture, Visual Rhythm, Human Pose, Motion patches, Discriminative scene patches, Action bank features, WMHI, SMAID, Binary space-time map, and Semantic image are also used as input to networks. Eight types of networks are categorized after studying each paper's architecture, i.e., CNN, RNN, LSTM, BiLSTM, C3D, 3D DNN, Two-Stream CNN, and Multi-Stream CNN. Some articles used more than one architecture type to extract information, as shown in Fig. 9. Most of the papers use existing popular networks for image data to extract features from video sequences such as AlexNet, VGG, DenseNet, ResNet, ResNeXt, GoogLeNet, Inception and MobileNet. 3D convolutional networks shown in Fig. 8 are identified as the most widely used deep learning architecture because they can capture both spatial and temporal information by moving along horizontal direction and depth. The scope of these results is only limited to our selected articles. But two-stream CNN and C3D are popular choices for researchers to work for HAR.

For RQ2, challenges discussed in each paper and proposed solutions are listed in Section V. Common challenges faced in HAR are objective and subjective factors in visual appearance due to different recording settings and human behaviors. Intra-class and inter-class diversity due to complexities involved in real-life human tasks, selecting informative frames from each clip to make the network more efficient. Optimizing numerous deep neural architectures parameters, extracting discriminative and compact video representations, avoiding overfitting, learning with small datasets, and open-set action recognition are included.

To answer RQ3, datasets used for evaluation in each paper are listed, and its characteristics are analyzed from its original published article and organized in Table 9. Particularly, we record each dataset source, its challenges and video infor-

mation like total videos, classes, clips duration and video per class. UCF-101 and HMDB-51 are the most widely used datasets for evaluation in observation to this study, as shown in Fig. 11. Table 10 summarizes the superior architectures with maximum results achieved for each dataset. Each dataset varies, and therefore architectures with maximum performance on that dataset also vary. This SLR concluded that human activity recognition requires further research to overcome identified challenges with more optimized deep neural networks.

For future work, we first proposed to focus on open-set action recognition to provide a global solution independent of datasets to make it suitable for the real world. Second, transfer learning can be exploited with existing architecture knowledge for more robust feature extraction. Third, more hybrid approaches should be presented to optimize and enhance the deep learning paradigm to take full advantage of different architecture types, as shown in Fig. 27.

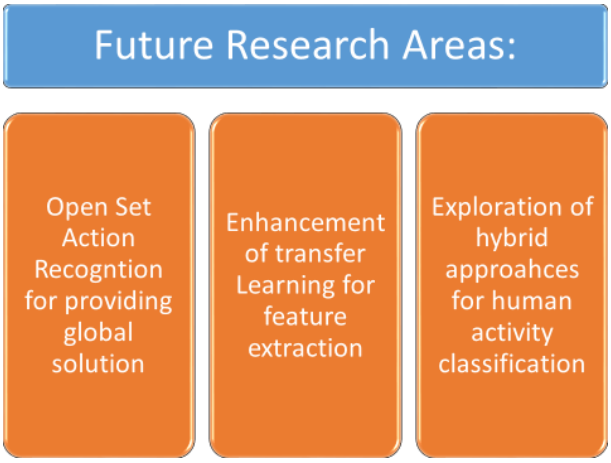

FIGURE 27: Potential research areas for future work.

REFERENCES

[1] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—a survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1254, 2018.

[2] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7882–7891.

[3] M. A. A. H. Khan, R. Kukkapalli, P. Waradpande, S. Kulandaivel, N. Banerjee, N. Roy, and R. Robucci, "Ram: Radar-based activity monitor," in IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications. IEEE, 2016, pp. 1–9.

[4] M. A. A. H. Khan, H. S. Hossain, and N. Roy, "Infrastructure-less occupancy detection and semantic localization in smart environments," in proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services on 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2015, pp. 51–60.

[5] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," Pattern Recognition Letters, vol. 119, pp. 3–11, 2019.

[6] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in Ijcai, vol. 15. Buenos Aires, Argentina, 2015, pp. 3995–4001.

[7] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7882–7891.

[8] M. A. A. H. Khan, R. Kukkapalli, P. Waradpande, S. Kulandaivel, N. Banerjee, N. Roy, and R. Robucci, "Ram: Radar-based activity monitor," in IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications. IEEE, 2016, pp. 1–9.

[9] M. A. A. H. Khan, H. S. Hossain, and N. Roy, "Infrastructure-less occupancy detection and semantic localization in smart environments," in proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services on 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2015, pp. 51–60.

[10] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Human activity recognition: Various paradigms," in 2008 international conference on control, automation and systems. IEEE, 2008, pp. 1896–1901.

[11] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in 2008 IEEE international symposium on circuits and systems. IEEE, 2008, pp. 2737–2740.

[12] S.-F. Chang, "The holy grail of content-based media analysis," IEEE MultiMedia, vol. 9, no. 2, pp. 6–10, 2002.

[13] J. Choi, Y.-i. Cho, T. Han, and H. S. Yang, "A view-based real-time human action recognition system as an interface for human computer interaction," in International Conference on Virtual Systems and Multimedia. Springer, 2007, pp. 112–120.

[14] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video based technology for ambient assisted living: A review of the literature," Journal of Ambient Intelligence and Smart Environments, vol. 3, no. 3, pp. 253–269, 2011.

[15] H. D. Mehr and H. Polat, "Human activity recognition in smart home with deep learning approach," in 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG). IEEE, 2019, pp. 149–153.

[16] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, and V. C. Leung, "Non-intrusive human activity monitoring in a smart home environment," in 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013). IEEE, 2013, pp. 606–610.

[17] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition–a systematic review of literature," IEEE Access, vol. 6, pp. 59 192–59 210, 2018.

[18] H.-B. Zhang, Q. Lei, B.-N. Zhong, J.-X. Du, and J. Peng, "A survey on human pose estimation," Intelligent Automation & Soft Computing, vol. 22, no. 3, pp. 483–489, 2016.

[19] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," Computer Vision and Image Understanding, vol. 192, p. 102897, 2020.

[20] R. Li, Z. Liu, and J. Tan, "A survey on 3d hand pose estimation: Cameras, methods, and datasets," Pattern Recognition, vol. 93, pp. 251–272, 2019.

[21] S. Berretti, M. Daoudi, P. Turaga, and A. Basu, "Representation, analysis, and recognition of 3d humans: A survey," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 14, no. 1s, pp. 1–36, 2018.

[22] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image and vision computing, vol. 60, pp. 4–21, 2017.

[23] H. Zhang, "The literature review of action recognition in traffic context," Journal of Visual Communication and Image Representation, vol. 58, pp. 63–66, 2019.

[24] R. Khurana and A. K. S. Kushwaha, "Deep learning approaches for human activity recognition in video surveillance-a survey," in 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018, pp. 542–544.

[25] L. L. Presti and M. La Cascia, "3d skeleton-based human action classification: A survey," Pattern Recognition, vol. 53, pp. 130–147, 2016.

[26] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," Image and Vision Computing, vol. 55, pp. 42–52, 2016.

[27] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," Computer Vision and Image Understanding, vol. 158, pp. 85–105, 2017.

[28] J. K. Dhillon, A. K. S. Kushwaha et al., "A recent survey for human activity recoginition based on deep learning approach," in 2017 fourth in-

ternational conference on image information processing (ICIIP). IEEE, 2017, pp. 1–6.

[29] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," Computer Vision and Image Understanding, vol. 171, pp. 118–139, 2018.

[30] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," Computer Vision and Image Understanding, vol. 188, p. 102799, 2019.

[31] B. Liu, H. Cai, Z. Ju, and H. Liu, "Rgb-d sensing based human action and interaction analysis: A survey," Pattern Recognition, vol. 94, pp. 1–12, 2019.

[32] S. M. Ali, J. C. Augusto, and D. Windridge, "A survey of user-centred approaches for smart home transfer learning and new user home automation adaptation," Applied Artificial Intelligence, vol. 33, no. 8, pp. 747–774, 2019.

[33] S. Letchmunan, F. H. Hassan, S. Zia, and A. Baqir, "Detecting video surveillance using vgg19 convolutional neural networks."

[34] A. Kofod-Petersen, "How to do a structured literature review in computer science," Ver. 0.1. October, vol. 1, 2012.

[35] S. Götz, "Supporting systematic literature reviews in computer science: the systematic literature review toolkit," in Proceedings of the 21st ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, 2018, pp. 22–26.

[36] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi, "Spatio-temporal crime hotspot detection and prediction: A systematic literature review," IEEE Access, vol. 8, pp. 166 553–166 574, 2020.

[37] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[38] F. Weidt and R. Silva, "Systematic literature review in computer science-a practical guide," Relatórios Técnicos Do DCC/UFJF, vol. 1, 2016.

[39] J. Zhang, H. Hu, and X. Lu, "Moving foreground-aware visual attention and key volume mining for human action recognition," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 15, no. 3, pp. 1–16, 2019.

[40] Y. Cheng, G. Li, N. Wong, H.-B. Chen, and H. Yu, "Deepeye: A deeply tensor-compressed neural network for video comprehension on terminal devices," ACM Transactions on Embedded Computing Systems (TECS), vol. 19, no. 3, pp. 1–25, 2020.

[41] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 510–514, 2016.

[42] Y. Gao, X. Xiang, N. Xiong, B. Huang, H. J. Lee, R. Alrifai, X. Jiang, and Z. Fang, "Human action monitoring for healthcare based on deep learning," Ieee Access, vol. 6, pp. 52 277–52 285, 2018.

[43] S. Yu, L. Xie, L. Liu, and D. Xia, "Learning long-term temporal features with deep neural networks for human action recognition," IEEE Access, vol. 8, pp. 1840–1850, 2019.

[44] Y. Wan, Z. Yu, Y. Wang, and X. Li, "Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features," IEEE Access, vol. 8, pp. 85 284–85 293, 2020.

[45] Y. Huang, Y. Guo, and C. Gao, "Efficient parallel inflated 3d convolution architecture for action recognition," IEEE Access, vol. 8, pp. 45 753–45 765, 2020.

[46] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4597–4605.

[47] Q.-Q. Chen, F. Liu, X. Li, B.-D. Liu, and Y.-J. Zhang, "Saliency-context two-stream convnets for action recognition," in 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016, pp. 3076–3080.

[48] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pp. 2799–2813, 2017.

[49] A. Kar, N. Rai, K. Sikka, and G. Sharma, "Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3376–3385.

[50] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5523–5531.

[51] H. Fang, J. Thiyagalingam, N. Bessis, and E. Edirisinghe, "Fast and reliable human action recognition in video sequences by sequential analysis," in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 3973–3977.

[52] Y. Liu, Q. Wu, and L. Tang, "Frame-skip convolutional neural networks for action recognition," in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017, pp. 573–578.

[53] F. Al-Azzo, C. Bao, A. M. Taqi, M. Milanova, and N. Ghassan, "Human actions recognition based on 3d deep neural network," in 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT). IEEE, 2017, pp. 240–246.

[54] Z. Liu, Y. Tian, and Z. Wang, "Improving human action recognitionby temporal attention," in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 870–874.

[55] L. Zhang, Y. Feng, X. Xiang, and X. Zhen, "Realistic human action recognition: When cnns meet lds," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 1622–1626.

[56] H. Yang, C. Yuan, J. Xing, and W. Hu, "Scnn: Sequential convolutional neural network for human action recognition in videos," in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 355–359.

[57] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1510–1520, 2017.

[58] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 177–186.

[59] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer lstm," IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp. 9692–9702, 2018.

[60] X. Liu, T. You, X. Ma, and H. Kuang, "An optimization model for human activity recognition inspired by information on human-object interaction," in 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, 2018, pp. 519–523.

[61] D. T. Concha, H. D. A. Maia, H. Pedrini, H. Tacon, A. D. S. Brito, H. D. L. Chaves, and M. B. Vieira, "Multi-stream convolutional neural networks for action recognition in video sequences based on adaptive visual rhythms," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018, pp. 473–480.

[62] J. You, P. Shi, and X. Bao, "Multi-stream i3d network for fine-grained action recognition," in 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2018, pp. 611–614.

[63] C. Cheng, P. Lv, and B. Su, "Spatiotemporal pyramid pooling in 3d convolutional neural networks for action recognition," in 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 3468–3472.

[64] K. Sarker, M. Masoud, S. Belkasim, and S. Ji, "Towards robust human activity recognition from rgb video stream with limited labeled data," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018, pp. 145–151.

[65] M. Li, Y. Qi, J. Yang, Y. Zhang, J. Ren, and H. Du, "3d convolutional two-stream network for action recognition in videos," in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019, pp. 1697–1701.

[66] D. Wu, J. Chen, N. Sharma, S. Pan, G. Long, and M. Blumenstein, "Adversarial action data augmentation for similar gesture action recognition," in 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019, pp. 1–8.

[67] M. A. Jalal, W. Aftab, R. K. Moore, and L. Mihaylova, "Dual stream spatio-temporal motion fusion with self-attention for action recognition," in 2019 22th International Conference on Information Fusion (FUSION). IEEE, 2019, pp. 1–7.

[68] A. Roy and D. Mishra, "Ecnn: Activity recognition using ensembled convolutional neural networks," in TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019, pp. 757–760.

[69] H. D. Mehr and H. Polat, "Human activity recognition in smart home with deep learning approach," in 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG). IEEE, 2019, pp. 149–153.

[70] Z. Hu and E.-J. Lee, "Human motion recognition based on improved 3-dimensional convolutional neural network," in 2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE). IEEE, 2019, pp. 154–156.

[71] W. Luo, C. Zhang, W. Liu, J. Wu, and W. Lin, "Improving action recognition with valued patches exploiting," in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, 2019, pp. 181–188.

[72] W. McNally, A. Wong, and J. McPhee, "Star-net: Action recognition using spatio-temporal activation reprojection," in 2019 16th Conference on Computer and Robot Vision (CRV). IEEE, 2019, pp. 49–56.

[73] M. Xin, H. Zhang, H. Wang, M. Sun, and D. Yuan, "Arch: Adaptive recurrent-convolutional hybrid networks for long-term action recognition," Neurocomputing, vol. 178, pp. 87–102, 2016.

[74] E. P. Ijjina and K. M. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," Pattern recognition, vol. 59, pp. 199–212, 2016.

[75] S. Yu, Y. Cheng, L. Xie, Z. Luo, M. Huang, and S. Li, "A novel recurrent hybrid network for feature fusion in action recognition," Journal of Visual Communication and Image Representation, vol. 49, pp. 192–203, 2017.

[76] T. V. Nguyen and B. Mirza, "Dual-layer kernel extreme learning machine for action recognition," Neurocomputing, vol. 260, pp. 123–130, 2017.

[77] L. Lu, H. Di, Y. Lu, L. Zhang, and S. Wang, "A two-level attention-based interaction model for multi-person activity recognition," Neurocomputing, vol. 322, pp. 195–205, 2018.

[78] Y. Yuan, Y. Zhao, and Q. Wang, "Action recognition using spatial-optical data organization and sequential learning framework," Neurocomputing, vol. 315, pp. 221–233, 2018.

[79] Y. Sun, X. Wu, W. Yu, and F. Yu, "Action recognition with motion map 3d network," Neurocomputing, vol. 297, pp. 33–39, 2018.

[80] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3d convolutional neural networks for action recognition," Pattern recognition, vol. 85, pp. 1–12, 2019.

[81] P.-S. Kim, D.-G. Lee, and S.-W. Lee, "Discriminative context learning with gated recurrent unit for group activity recognition," Pattern Recognition, vol. 76, pp. 149–161, 2018.

[82] F. He, F. Liu, R. Yao, and G. Lin, "Local fusion networks with chained residual pooling for video action recognition," Image and Vision Computing, vol. 81, pp. 34–41, 2019.

[83] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," Pattern Recognition, vol. 79, pp. 32–43, 2018.

[84] Z. Zhu, H. Ji, W. Zhang, and Y. Xu, "Rank pooling dynamic network: Learning end-to-end dynamic characteristic for action recognition," Neurocomputing, vol. 317, pp. 101–109, 2018.

[85] M. Ma, N. Marturi, Y. Li, A. Leonardis, and R. Stolkin, "Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos," Pattern Recognition, vol. 76, pp. 506–521, 2018.

[86] V. Adeli, E. Fazl-Ersi, and A. Harati, "A component-based video content representation for action recognition," Image and Vision Computing, vol. 90, p. 103805, 2019.

[87] Y. Quan, Y. Chen, R. Xu, and H. Ji, "Attention with structure regularization for action recognition," Computer Vision and Image Understanding, vol. 187, p. 102794, 2019.

[88] M. Tong, M. Zhao, Y. Chen, and H. Wang, "D3-lnd: A two-stream framework with discriminant deep descriptor, linear cmdt and nonlinear kcmdt descriptors for action recognition," Neurocomputing, vol. 325, pp. 90–100, 2019.

[89] S. Chaudhary and S. Murala, "Deep network for human action recognition using weber motion," Neurocomputing, vol. 367, pp. 207–216, 2019.

[90] G. An, Z. Zheng, D. Wu, and W. Zhou, "Deep spectral feature pyramid in the frequency domain for long-term action recognition," Journal of Visual Communication and Image Representation, vol. 64, p. 102650, 2019.

[91] Y. Ye, X. Yang, and Y. Tian, "Discovering spatio-temporal action tubes," Journal of Visual Communication and Image Representation, vol. 58, pp. 515–524, 2019.

[92] Z. Zhu, H. Ji, and W. Zhang, "Nonlinear gated channels networks for action recognition," Neurocomputing, vol. 386, pp. 325–332, 2020.

[93] P. Wang, L. Liu, C. Shen, and H. T. Shen, "Order-aware convolutional pooling for video based action recognition," Pattern Recognition, vol. 91, pp. 357–365, 2019.

[94] C. Zalluhoglu and N. Ikizler-Cinbis, "Region based multi-stream convolutional neural networks for collective activity recognition," Journal of Visual Communication and Image Representation, vol. 60, pp. 170–179, 2019.

[95] Z. Liao, H. Hu, J. Zhang, and C. Yin, "Residual attention unit for action recognition," Computer Vision and Image Understanding, vol. 189, p. 102821, 2019.

[96] Z. Zheng, G. An, D. Wu, and Q. Ruan, "Spatial-temporal pyramid based convolutional neural network for action recognition," Neurocomputing, vol. 358, pp. 446–455, 2019.

[97] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3d convnets with attention for action recognition," Pattern Recognition, vol. 98, p. 107037, 2020.

[98] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition," Pattern Recognition, vol. 92, pp. 13–24, 2019.

[99] J. Wang, X. Peng, and Y. Qiao, "Cascade multi-head attention networks for action recognition," Computer Vision and Image Understanding, vol. 192, p. 102898, 2020.

[100] B. Fernando and S. Gould, "Discriminatively learned hierarchical rank pooling networks," International Journal of Computer Vision, vol. 124, no. 3, pp. 335–355, 2017.

[101] H. Zhang, M. Xin, S. Wang, Y. Yang, L. Zhang, and H. Wang, "End-to-end temporal attention extraction and human action recognition," Machine Vision and Applications, vol. 29, no. 7, pp. 1127–1142, 2018.

[102] A. Cherian and S. Gould, "Second-order temporal pooling for action recognition," International Journal of Computer Vision, vol. 127, no. 4, pp. 340–362, 2019.

[103] J. Yu, D. Y. Kim, Y. Yoon, and M. Jeon, "Action matching network: open-set action recognition using spatio-temporal representation matching," The Visual Computer, pp. 1–15, 2019.

[104] A. Khelalef, F. Ababsa, and N. Benoudjit, "An efficient human activity recognition technique based on deep learning," Pattern Recognition and Image Analysis, vol. 29, no. 4, pp. 702–715, 2019.

[105] Y. Zhu and G. Liu, "Fine-grained action recognition using multi-view attentions," The Visual Computer, vol. 36, no. 9, pp. 1771–1781, 2020.

[106] S. A. Khowaja and S.-L. Lee, "Semantic image networks for human action recognition," International Journal of Computer Vision, vol. 128, no. 2, pp. 393–419, 2020.

[107] S. Arif, T. Ul-Hassan, F. Hussain, J. Wang, and Z. Fei, "Video representation by dense trajectories motion map applied to human activity recognition," International Journal of Computers and Applications, vol. 42, no. 5, pp. 474–484, 2020.

[108] K. Wang, T. Wang, L. Liu, and C. Yuan, "Human behaviour recognition and monitoring based on deep convolutional neural networks," Behaviour & Information Technology, pp. 1–12, 2019.

[109] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., vol. 3. IEEE, 2004, pp. 32–36.

[110] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 12, pp. 2247–2253, 2007.

[111] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008, pp. 1–8.

[112] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2929–2936.

[113] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 1996–2003.

[114] S. Blunsden and R. Fisher, "The behave video dataset: ground truthed video for multi-person behavior classification," Annals of the BMVA, vol. 4, no. 1-12, p. 4, 2010.

[115] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563.

[116] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," Center for Research in Computer Vision, vol. 2, no. 11, 2012.

[117] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 1194–1201.

[118] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in European Conference on Computer Vision. Springer, 2012, pp. 215–230.

[119] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 3192–3199.
[120] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," Machine vision and applications, vol. 24, no. 5, pp. 971–981, 2013.
[121] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1971–1980.
[122] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.
[123] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," arXiv preprint arXiv:1808.01340, 2018.
[124] 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," in 2015 IEEE International conference on image processing (ICIP). IEEE, 2015, pp. 168–172.
[125] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele, "Recognizing fine-grained and composite activities using hand-centric features and script data," International Journal of Computer Vision, vol. 119, no. 3, pp. 346–373, 2016.
[126] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, and V. C. Leung, "Non-intrusive human activity monitoring in a smart home environment," in 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013). IEEE, 2013, pp. 606–610.

HADIQA AMAN ULLAH received her BS (Hons.) degree in Information Technology from Punjab University in 2017. She is currently pursuing her MS (CS) degree from The University of Lahore. She has previously served as a lecturer in the field of Computer Science and IT and has one journal publication before. Her research interests are Data Science, Machine learning, and Computer vision.

UMAIR MUNEER BUTT is currently a Ph.D. student in the School of Computer Sciences at Universiti Sains Malaysia (USM), Malaysia. He received his BS (CS) degree from GIFT University, Pakistan, in 2012 and MS (CS) degree from the National University of Sciences and Technology (NUST), Pakistan, in 2016. He has more than six years of teaching and research experience in data mining, Machine learning, Data Science, and Image processing. He has served as a Research Associate for more than five years and worked on different real-world applications. Recently, he secured a Fundamental Research Grant Scheme (FRGS) from the Malaysian government for crime predictions. He has authored several journal, conferences, and book chapter in well-reputed journals during his career. His current research interests are Data Science, Data Mining, and Machine learning.

SUKUMAR LETCHMUNAN completed his PhD in Computer Science at the University of Strathclyde (UK) in 2013. Since then he is a Senior Lecturer in the School of Computer Sciences, University Sains Malaysia (USM).

He has been a tutor, technical trainer and served as lecturer and course coordinator at private college and private university prior to his PhD studies. His research interest are in Software Engineering, Software Metrics in Web applications, software cost estimation, service-oriented software engineering and agile project management.

M. SULTAN ZIA received the M.S.-C.S. and Ph.D. degrees in computer science from the FAST National University of Computer and Emerging Sciences, Islamabad, in 2008 and 2016 respectively, and the M.C.S. degree from International Islamic University, Islamabad, in 2005. He has 15 years of total teaching/research experience at university level and 4.5 years of industrial (software development) experience. He is currently an Associate Professor and the Head of the Department of Computer Science and IT, The University of Lahore, Chenab Campus. His research interest includes the IoT, machine learning, and computer vision.

DR. FADRATUL HAFINAZ HASSAN is a senior lecturer at the School of Computer Sciences, Universiti Sains Malaysia. She obtained her Doctor of Philosophy (Ph.D) degree in Computer Science (CS) in 2013 from the School of Information Systems, Computing and Mathematics, Brunel University, West London. Her research includes the area of Artificial Intelligence (AI) for pedestrian simulation and spatial layout optimization. She has co-authored over 30 publications and secured 10 research grants; 5 as principal investigators and 5 grants as co-investigators. Currently her research involved studying pedestrian simulation models in the urban planning domain with the University of Sydney School of Architecture, Design and Planning.