A Comparative Study of Various Human Activity Recognition Approaches

Dhruv Goel¹ and Rahul Pradhan²

GLA University, Mathura

¹dhruvgoel297@gmail.com, ²rahul.pradhan@glau.ac.in

Abstract. Human Activity Recognition (HAR) is a vast and exciting topic for researchers and students. HAR aims to recognize activities by observing the actions of subjects and surrounding conditions. This topic also has many significant and futuristic applications and a basis of many automated tasks like 24*7 security surveillance, healthcare, laws regulations, automatic vehicle controls, game controls by human motion detection, basically human-computer interaction. This survey paper focuses on reviewing other research papers on sensing technologies used in HAR. This paper has covered distinct research in which researchers collect data from smartphones; some use a surveillance camera system to get video clips. Most of the researchers used videos to train their systems to recognize human activities collected from YouTubes and other video sources. Several sensor-based approaches have also covered in this survey paper to study and predict human activities, such as accelerometer, gyroscope, and many more. Some of the papers also used technologies like a Convolutional neural network (CNN) with spatiotemporal three-dimensional (3D) kernels for model development and then using to integrate it with OpenCV. There are also work done for Alzheimer's patient in the Healthcare sector, used for their better performance in day-to-day tasks. We will analyze the research using both classic and less commonly known classifiers on distinct datasets available on the UCI Machine Learning Repository. We describe each researcher's approaches, compare the technologies used, and conclude the adequate technology for Human Activity Recognition. Every research will be discussed in detail in this survey paper to get a brief knowledge of activity recognition.

1. Introduction

Our research topic, i.e., Human Activity Recognition (HAR) is a very vast and exciting topic for research and a mode of attraction for students. There are already many kinds of research done in this domain of Computer Vision and AI. These researches also help automate applications when implemented for real-world performance, like long-distance broadcasts, video surveillance, human-computer interaction (HCI), and many more in the healthcare sector. However, it is also having its difficulty phase in terms of sensor motions and their installations; background activities also affect the model outcomes, variability in the activities played by different humans, variation in scale, viewpoint, and partial occlusion. As the thinking capabilities advance and hardware configurations for HAR system upgrades, research approaches are continually increasing. In this survey paper, much research on sensing technologies, video analysis for activities, and computer vision techniques used in HAR are reviewed. The most commonly used techniques for the HAR system, irrespective of computational powers or activities classification algorithms, are considered.

As we approach any separation process, we should consider two questions: "What action?" about the recognition problem, and "Where is the video?" with a local problem. When trying to identify human activity, one has to determine the person's kinetic state so that the computer can see the activity. The
system can quickly determine some activities which happen in daily life, such as "walking," "running," "yoga," and many more. While in comparison with daily activities, activities like "peeling apple" are more complex to determine. They should be divided into more specific activities to get recognized by the system. The HAR system should first recognize the objects to better recognize the activity as it provides more information about the ongoing event [1].

HAR system has an inherent hierarchical structure, which indicates three different levels (figure 1). First, core technologies are used step by step; the first step is object segmentation followed by feature extraction from each frame of the video, and then using algorithms will help classify the human activities. The first level is then needed to integrate with some system to observe the results of algorithms and video processing in which the second layer comes into the picture. The HAR system comprises three categories of activities namely, recognition of one-on-one work, social interaction and crowd behaviour, and unusual activity recognition. Finally, the top layer, i.e., application implementation of the HAR system, can be performed in any places, majorly crowded areas for better law regulations like surveillance environments, entertainment environments, and healthcare systems.

Human activities also follow a hierarchical structure having activities divided into three categories. First is action primitive; this consists of complex activities. The second category comprises action/activities, and then finally, complex integration is a third category that refers to human activities involving people or objects over two people. In this survey paper, we will also combine these three categories. The first verb refers to activities that involve parts of the human body, such as the arms, legs, hands, upper limbs, and many more [2]. This category includes activities like "stretching the upper body," "rotating hands," and many others. The second category, i.e., Actions and activities, refers to the movement of the whole body made in the first few actions in a series of moments and performed by one person in the frame. This category includes activities such as walking, eating, upstairs, yoga, and many others. Interactions refer to the activities performed by more than one person, such as arm wrestling, and these types of activities are sometimes very complex to recognize.
Also, for real world deployment of HAR, efficient video analysis is very important, both in online and on the edge mode of video recognition. To get efficient results we need to use more advance technologies like deep learning [86, 87, 78, 88, 93, 89, 92]. Video understanding is more difficult as compare to image recognitions, as videos include temporal models. For e.g., for distinguishing between forward and backward walking, reversing the sequence will give different results, so temporal modelling is essential. Many previous approaches use 2D CNN [84, 77, 87, 92]. In recent approaches 3D is being a most promising techniques in video recognition as they consider both spatial and temporal features, but we have to compromise on computational cost and deployment, that’s why it cannot be applied in online video recognition. We proposed a method in consideration of all the constraints i.e., Temporal Shift Module (TSM) (figure 2).

We tried the shift strategy used in image recognition for video recognition too, but we face two major issues: (1) not efficient: shift operation is conceptually zero FLOP but it is having data movement. And the extra data movement cost too much which is non-negligible and will result in increase of latency. (2) not accurate: too many shifting of channels will lead to performance degradation as it will destroy the spatial modelling capability. To tackle with these problems, we have come up with two solutions, (1) instead of moving all the channels, we will be moving some of its portion turn by turn, which will reduce the data movement cost. (figure 3a) (2) Use of residual branch, which will preserve the activation of the current frame, and will not destroy 2D CNN spatial modelling capability.

Many researchers have already concluded the HAR system in three ways of representation, i.e., Global, Local, and Depth-based. Researchers of [3] use image descriptors as space-time shapes, a clear example of Global representation. The evolution comes with space-time interest point (STIP), proposed in [4], an example of local representation that uses HOG (Histogram of Oriented Gradients) and HOF (Histogram of optical flow) as the image descriptors for informative interest points. Depth-based representation becomes a new research topic with advancements in cameras with an RGB camera, different sensors like accelerometer, gyroscope, and many others.

Machine learning classification techniques like DTW (Dynamic Time Warping) were designed for speech recognition; also Hidden Markov Model (HMM) was one of them; these algorithms were not
explicitly designed for HAR [5, 6]. Also, many deep learning techniques were developed for a large amount of image classification [7]. With the increase in crime rates, another research area is proposed, i.e., Human tracking to keep surveillance over abnormal behaviours.

The rest of the paper covers the following sections: in section 2, we will discuss the researches covered in some previous surveys, then further in section 3, we will cover some common challenges faced in human activity recognition and how to overcome them. In section 4, we presented human activities' categorization as local and global representations in RGB videos and depth image approaches. Section 5 reviewed the various HAR methods and classification techniques and analyzed the pros and cons of different approaches used for HAR. In section 6, we reviewed different human tracking methods and discussed some futuristic research topics in HAR. Finally, at last, we will conclude our survey in the last section.

2. Previous Surveys and Taxonomy

Numerous research published are covered area of human activity recognition. [8] divided the study into 2D (with exact and external shapes) and 3D methods. In [9], introduced a new tax code focusing on mobility, followed by a deep sensors multiview cameras, and the acquisition of activities recognition. In line with the spirit of the previous tax [10] the successive category of management has been promoted. A study by [11] focused on high-performance perceptions and raised four times as much tax, including the beginning of the human movement, tracking, positioning, and visualization.

[12] proposed research differentiating between "actions" and "activities," in this research author categorizes activities according to their activity’s complexities. [13] also categorize human activity methods in their approach in two categories, i.e., "top-down" and "bottom-up." In the same way, [14] also divided the methods of human activity recognition into tree-based taxonomy into two categories, "single layer" and "hierarchical" methods. Researchers like the proposed methods [15, 16] by modeling 3D data. Work recognition with representation of the human body in 3D modelling is more accurate and beneficial for analysis than 2D functions seen in photographs, as deep cameras are used to detect the most powerful features of connected organs and joints in the human body. [17] The recently proposed study focuses on methods that exploit 3D depth data from two phases of HAR 3D stereo systems and motion capture systems. Motion imaging systems have also played their role in the skeletal muscles described using deep sensors.

[18] suggested the HAR method from standing images as all known methods were used for video processing, and most HAR systems are proposed to be known for face and measurement techniques. As [19] proposed computer-to-human communication [52,53], the focus was on facial expressions, placement measurement process, and speech recognition. Research is proposed on state recognition methods that include non-verbal means, such as facial expressions and facial expressions. [21] have studied modern HAR techniques, and have gone through many open computer problems and how they can enhance the efficiency of the human-computer system. [22] presented a review of modern observational perspectives using critical and perceptual autonomous states and provided a list of data-related data ideas related to human speech. [23] suggested analyzing a variety of non-verbal forms (i.e., visual and auditory cues) behavioural recognition patterns and data details of automated agreements and disagreements. Such social values can play an important role in analyzing social behaviour, which is the key to public participation. Finally, a complete technical analysis of human behaviour recognition from a data perspective and information presented in [24].

2D CNN methods with two types of streamed data i.e., spatial and temporal are more efficient than 3D approaches [77,91], as Temporal Segment Networks [87] use some segments of frames to extract features. But this is not much effective in more complex temporal frames and videos. 3D CNN can efficiently learn spatial-temporal frames. As they are costly in terms of computation and deployment is also difficult. 3D CNN is also having more hyper parameters than 2D CNNs, that make it more chances for over-fitting. Our TSM approach is also just like 3D CNN with spatial-temporal modelling ability and computational abilities of 2D CNNs. As discussed above, we can use 3D CNN method for direct temporal modelling. An alternative way to perform temporal modelling is with the help of 2D CNN and post-hoc fusion [83, 80, 92, 79]. Use of LSTM to aggregate with 2D CNN features also proposed in previous studies [90, 79, 85, 81, 82].

4
Transfer Learning is used to transfer useful information for better classification like weights of the model [94], features, objects of the classifiers. A most in-demand application of transfer learning is used in CNN to predict the diseases by classification from images [95, 96]. Transfer learning is also used in transferring images features to ECG machine for disease prediction [97]. Among all of these applications and methods one important thing is to know what parameters to transfer prior implementing final model.

3. Challenges in Human Activity Recognition (HAR)

3.1 Intraclass Variation and Interclass Similarity
There is no grammar and a firm definition of human activity, unlike speech recognition. This creates double confusion. On the other hand, the same function can vary from topic to topic, leading to intraclass diversity. Play, speed and energy also increase meeting spaces. On the other hand, different tasks may present similar situations (e.g., using laptops and reading). These functions are called interclass simulations, which are more common in HAR. Accurate and differentiating features need to be developed and removed from activity videos to address these issues.

3.2 Recognition under Real-World Settings

3.2.1 Unsaturated Backgrounds. Performance recognition's complexity comes when most situations use dynamic recording devices, while applications such as video recognition and fall system use still cameras. A typical example of a powerful recording is a sports event broadcast. People uses portable devices with embedded cameras and smart glasses also to record the streaming video. These streaming videos having more complex background with changing location which make them strong and complex backgrounds to recognise the activity. Second, the actual videos are full of visuals, lighting variations, and viewing changes, making it challenging to see tasks in complex and different situations.

3.2.2 More than two subject Interactions and Crowded Activities. Previous research focused on human activities of low-level types such as running, jumping, and raising hands. One common feature of these functions is based on single type communication without personal or human communication. They involve only an individual in an activity. However, people often do activities with one or more people and objects in the real world. The football game is an excellent example of a partnership where many players (i.e., personal contact) in a team defend the ball (e.g., personal contact) and compete with players from another team. It is a challenge to find and follow many topics in a row or to see all the team's activities as "playing football" instead of "it works."

3.2.3 Distanced and Degraded-Quality Videos. Far distanced and low-impact videos with high visibility are available in most video viewing contexts. Larger and denser areas such as airport

Figure 4: Long-Distance and Low-Quality videos: (a) Abnormal behaviours in surveillance. (b) HAR in long-distance broadcasts.
municipalities and passenger passengers are independent times where things happen more often. Besides, high-quality surveillance cameras cannot deliver high-quality videos like current data sets when the target person is clear and transparent. While we do not expect to follow everyone in these situations, some unusual or criminal-related behaviour should be identified by the HAR system (Figure 4 (a)). Another common long-distance case is the distribution of the ball (Figure 4 (b)). Due to the long-range of cameras, the issue is small, making it difficult to analyze the torso [10], and the low quality of those long-distance videos increases the difficulty.

3.2.4 How to overcome these challenges. To overcome these challenges, we need to include alternatives, namely, (i) backlash [25, 26]; in this case, the system distinguishes a domain that does not function over time (background) in relation to people, moving or changing objects in the framework (front of the earth); (ii) tracking a person, in this case the tracking of a person's position over time, which also helps to follow an unusual behaviour [27, 28, 29]; and (iii) the discovery of a person's object and the acquisition of an act [30, 31, 32], in this system can identify the activities and actions of the person in the image.

4. Human Activity Representation and Categorization

4.1 Depth-Based Representation

Previous HAR researches takes a look at traditional RGB cameras to captured feasible quality videos. Whereas Deep cameras are having complex performance and high cost that’s why they are limited for researches [34]. Some less expensive cameras such as Microsoft Kinetics [35] are present which helps researchers a lot with similar quality of data and depth resolutions and also, they are more accessible and more comfortable. To detect joints in real time scenarios there are kinetic SDK (accept algorithms in [35]). There’s a major impact on computer viewing community due to availability of depth maps and skeletal information. These factors are prominent solution for solving HAR problem i.e., instead of conventional RGB-based methods use of in-depth maps or also can be used as enhancers in RGB designed approaches. This section has reviewed recent developments in job representation separately using in-depth maps or skeletons.

4.1.1 Depth Maps Representation. In-depth maps contain deeper links compared to standard color images and are very informative. The methods presented in this section view depth maps as spacecraft and extract features directly from them. These features are used independently or integrated with the RGB channel to create multiple features. Li et al. [36] used an action graph model that represents tasks that use many key positions that serve as locations in the action graph. All functions share the same standing sets, and each stand looks like a 3D dot bag from a depth map. However, incorporating all 3D points is expensive for a computer; therefore, proposing a more direct and effective method of sampling 3D representation points, achieving an accuracy of more than 90% recognition by sampling approximately 1% point report.

4.1.2 Skeleton-Based Representations. Skeletal structure of Bones and integrated joint’s positions are the main focused features produced in deep maps. The Kinect device is ubiquitous in this case due to easy access to bones and joints. The Kinect v1 SDK system has 20 members, while the latest version (Kinect v2) has 25 members, adding five members close to the hands and neck (see Figure 4). We have reviewed recent papers on skeleton-based presentations and summarized three aspects of efforts to improve the effectiveness of skeleton-based presentations.

First, the skeletal model has a natural defect that is always a problem with the skeletal system when working at the edges (see Figure 5) [36]. Factors from the bones and spine can end up being 1 wrong. Current methods often solve it by combining other solid materials to close or reduce the problems of closure by separating all the bones in different parts of the body and behaving independently because not all body parts are included.
4.2 Human Activity Categorization

The categorization of human activity recognition tasks is the most challenging task as they involve classification among different datasets that the model gets to train and test the system. Many researchers have divided HAR programs into two main categories, i.e., (i) Unimodal approaches and (ii) multidisciplinary approaches. After that further division, the two categories were divided into some more subdivisions.

Unimodal approaches detect activities from single-dimensional data, such as images, and are subdivided into (i) space-time, (ii) stochastic, (iii) statutory-based, (iv) structural-based methods.

As we discuss space approaches, transient factors include job recognition [37, 38] or trajectories [39, 40]. Stochastic methods use a mathematical model to identify human activities (e.g., Markov's hidden models) [41, 42] and, as the name suggests in methods based on the Rule, use rules to identify human activities [43, 44]. The most effective approach is modeling, using high-level thinking to mimic body movements [45, 46].

Multimodal approaches encompass all the collected data from different sources [47] and are then divided into three categories: (i) relevant, (ii) behavioural, and (iii) human communication methods.

Affected methods use communication methods to determine a person's feelings and state of touch [48, 49]. Behavioural approaches focus on facial expressions, hand gestures, and placement of recognition to determine a person's work [50, 51].

5. Activity Classification Methods and Approaches

5.1 Unimodal Methods

Unimodal methods use single modality data, such as images. Many methods present task recognition as a set of visual objects extracted from videos or images and use multiple classification models to detect this function [54, 55]. Some methods use captions of movement trajectories as it will be challenging to ensure that the movement is continuous or not [56, 57], while others use the light flow feature, which helps them see activities from full-motion video.

5.1.1 Space-Time Methods. The excessive amount of work known is based on the space system [79]. It is understandable that the timing approach is based on human movements' movement over time and is used in human tracking systems to monitor abnormal behaviour. This tracking space method is based on positions, footprints and travel descriptions (Wang et al., 2013). (Efros et al. 2003) presented a study on long-distance recognizing where objects have small resolutions. Numerous such research are presenting different applications of space-time methods.

5.1.2 Stochastic Methods. The stochastic approach is the most commonly used approach from the beginning of the research area in computer vision and Machine Learning. Many researchers presented the use of Hidden Markov Model, many researchers also consider Hidden Conditional Random Fields for stochastic models (Bishop 2006, Quattoni et al. 2007). Human behaviour detection forms sequences of actions; this is presented in research by (Robertson and Reid et al. (2006), providing information vectors of positions, velocity, and local descriptors. These features are then used by the HMM model to predict human activities.

5.1.3 Rule-Based Methods. This method used pre-set sets of instructions to perform the recognition. Each sets of instruction is integrated used for development of a proper efficient human activity recognising model. Many types of research are conducted using a rule-based approach in HAR, and the rule-based approach is less regulated than others, as there are specific rules set out in the system.

5.1.4 Shape-Based Methods. Shape-based methods are all about human pose detection and body parts' appearance to keep track of human activities. Human emotions are represented as 2D space as rectangular patches, and the volume of motions are represented as 3D space as volumetric
shapes. Human bodies are presented as limbs jointly connected. By fetching all this information from videos, algorithms can detect the activities more clearly and accurately.

5.2 Multimodal Methods

In recent years, much attention has been given to the identification of multiple jobs. Many approaches are based on two different type of integration: initial integration and time integration. One of the easiest ways to get efficient performance is to integrate features in a wide vector and get the result of feature extraction (Sun et al., 2009). This combination method can improve visual performance, but the vector of the new feature is much larger.

Multimodal indicators are often compiled over time; therefore, the temporary organization of a large event and various approaches is important to understand the data. Other than visual synchronization, audio-visual analysis is used for many other purposes like keeping an eye on human trajectories (Lichtenauer et al., 2011, Perez et al., 2004) and active identity (Wu et al., 2013). Multimodal approaches are considered as three approaches: (i) affective modes, (ii) ethical approaches, and (iii) modes based on social media interaction. These methods describe the action of an atom or a combination that may be related to the emotional state of the person he is talking to emotionally and physically.

5.2.1 Affective Methods. This method is related to a person's emotional state, which can be recognized by hand gestures, speech, facial expressions, psychological changes, and activities. This is a combined form of computer vision, artificial intelligence, pattern recognition, and emotional science. This can be used to advance HAR by recognizing more proper and formal activities like "ongoing meeting" instead of predicting some people are sitting and talking, "a man trying to catch a bus" instead of running. Researches [20] are all about this approach of combining the emotional state of a person with its activity prediction.

5.2.2 Behavioural Methods. This method is very similar to the affective method as in this, we will be talking about a person's behaviour, and in the previous method, we were talking about whether a person is affected or not overtime, which can be viewed by a change in a person's psychological state. Factors that can affect a person's mood or behaviour can be segregated into different features like mood, emotions, interactions with other humans, actions, and activity. Hence, it will be difficult to predict the complex actions and human behaviour. Research in [70] presented an advancement for behaviour detection by integrating speech and visual detection.

5.2.3 Methods Based on Social Networking. A very basic task in every human life is social interaction, this is also a best way of detecting human activity as how frequent a person post their daily updates anyone can get the track of their activities and also there can be misuse of this in kidnapping or any other abnormal activity. Most common platforms for social interactions like Facebook, Youtube, and twitter measure social involvement of any person. A recent survey [71] provide full up-to-date information to automate human behaviour analysis.

5.2.4 Multimodal feature fusion. In this multimodal feature fusion, we try to recognize people's activities in a crowded area where everyone is performing different activities with different sounds. In this case, the HAR system will use visual information to recognize underlying activities easily recognized. However, the HAR system will perform recognition more accurately by taking account of audio-visual analysis. In crowded areas, many humans perform similar actions with their body parts but emit different audio. Through this audio-visual analysis, the HAR system can easily recognize the activities. The audio-visual information will help the system understand who the target person is for test sequences and distinguish different activities performed by different persons. The dimensionality of data from different modalities collected is the biggest challenge of multimodal recognition. As the dimension of video, data is much more complicated than that of audio data, so in this case, dimensionality reduction is useful. To perform these two types of techniques, we can use that proposed in [72, 73].
Early Fusion, is the process that occur on extraction level to extract features from different frames, at the time of extracting features and combining different modality by dimensionality reduction of data and forming a single vector that is viewed as an individual. The main advantage is that when we have so many different modalities that are positively correlated, the result will be more accurate since, in this Fusion, we need only a single learning phase. However, this approach may fail at the time of one modality dominating over others.

Another category is late Fusion; this Fusion is done at the decision level. It learns parameters of each modality separately by several probabilistic models. They will be combined in a supervised framework providing a final score [75, 76]. The strengths of each modality affect the results of recognition in a better way. However, this approach is time-consuming in comparison to another approach.

Also, a third approach is presenting by [74], known as slow Fusion. This approach combines the above-discussed approaches and acts as a hierarchical approach, this will pass all the layers from a fusion layer to get combined data and also this process is slightly a slower process. Furthermore, it is having advantages of both the fusion levels, but there is also a computational restriction because of the different level of raw data that is to be processed.

5.3 Classification Approaches
As the next stage, we will discuss some classification algorithms to recognize human activities from features extracted from videos and images. The classification approaches are considered in mainly three categories: frame pattern matching, discriminative models and generative models. Pattern matching-based methods are relatively more straightforward but sometimes expensive because of its computational requirements. Generative algorithms learn using joint probabilities. Now comes with discriminative methods that give results directly as generative methods include hidden Markov model, dynamic bayesian method, and discriminative approach include models such as Artificial neural network (ANN), relevance vector machine (RVM) and support vector machines (SVM).

5.3.1 Template-Based Approaches.

5.3.1.1 Template Matching. The template matching approach uses two components to match on shapetime, i.e., the motion-energy image that captures any motion over time, motion-history image that captures the last occurrence of any motion. Furthermore, when found in the frame, it generates a template of an activity. They can also be regarded as a shape-time approach. This approach is presented in [68, 69].

5.3.1.2 Dynamic Time Warping. This method is just like its name, a dynamic programming algorithm initially developed for speech recognition which is used as matching two sequences with variances. Its task to represent the words as templates and if found a new word than giving a word score for it. When it comes to applying DTW in HAR since human activities captured videos and videos that are collected online too is considered as a sequence of keyframes. Then the activity recognition task is then transformed to a frame pattern matching problem. Then research [67] proposed the solution of learning gestures using pre-trained models. The result was gained after gestures features after matching patterns among frames can be obtained from means and variances of correlations between features of image frames and new videos.

5.3.2 Generative Models

5.3.2.1 Hidden Markov Model Approach. One of the tasks of hidden markov model is human activity recognition. Hidden Markov Model is first proposed for speech recognition. In research [63], Yamato et al. was the first one to use HMM for activity recognition. They used the number of pixels as feature extracted from each frame. Then they use HMM to train on features combined vector of sequence from each frame including some probability gained by some matrices, confusion matrix and transition matrix. But after all this, the problem gets transformed into a
simple evolution problem. Then some advancements are presented to combine human body parts and track them for feature extraction. Then in combination of all the approaches, researchers presented an approach which is a coupled HMM approach used for human computer interaction in research [64-66].

5.3.2.2 *Dynamic Bayesian Networks.* A dynamic bayesian network, unlike HMMs, has an essential feature as it contains more than one random variable, which can contain only one random variable. In this way, HMMs are the simplified form of DBN with a restricted number of the random variable and a fixed graph structure. Research [62] proposed a model for two hands gestures; in this, we can see three hidden variables. These three hidden variables represent two hand motions and their spatial relation. Furthermore, we can extract at least five features, i.e., two hands motion and relative position with face and spatial relation observed between two hands.

5.3.3 *Discriminative Models*

5.3.3.1 *Support Vector Machines.* This method is the most commonly used model applicable in real world for discriminative models, and now getting started in HAR. Preciously, research [60] presented an SVM model to distinguish between two classes. In this, the author proposed a model to create a margin between two classes. Furthermore, research [61] presented a combination of space-time features and combined their SVM model to implement the HAR system through video processing. A dataset known as KTH sets a benchmark in HAR research areas; they used this data for their model training.

5.3.3.2 *Conditional Random Fields.* This method compactly represents the conditional probability of an event to be held, i.e., the sequence of event $Y$ from the combinations of sequence $X$, and it is a type of undirected graphical model. [59] compared HMMs and CRFs, and they discovered that CRF’s performance is better than HMMs, although HMM’s model features are independent of the assumptions of HMMs. Whereas the HMMs model is independent of the given labels; thus, these complex features invalidate its assumptions, and the method that is no longer a proper generative model is making of HMM.

5.3.3.3 *Deep Learning Architectures.* In the deep learning approach, we are having four approaches, i.e., Recurrent neural network (RNNs), Deep neural networks (DNNs), convolutional neural network (CNNs), and some emergent neural networks. Among these categories, ConVents is the one which is widely used for prediction with deep learning architecture. However, in HAR's case, the dataset that we get is very small compared to the ideal data that Convent needed to be trained on. One research of [58] represents the solution to combining hand-crafted features with improving the performance.

5.3.4 *Temporal Shift Module*

5.3.4.1 *Offline Video Recognition.* After we get videos, we first grouped $T$ frames as a sample to process on 2D CNN models. The model processed each frame and test them on a threshold and gathered all the results for final prediction. Our TSM model performs on same computational cost as 2D CNN performs. TSM will work on each frame’s residual block and this help in better temporal fusion of information at low computation. TSM is also beneficial in conversion of any 2D CNN model into 3D CNN model and this will not require any extra computation. This make our TSM model compatible on any hardware specifications, just with a pre-requisite of 2D CNN compatibility.

5.3.4.2 *Online Video Recognition.* Online video recognition is most important in real-life deployment of HAR. This is sometime difficult task to get a solution with high accuracies and low time
complexities. Online video recognition can be used in Augmented Reality or Self driving cars for crowd detection and their activities and also in its complex applications. Our TSM model is capable of achieving acceptable accuracies for online video recognition.

In Offline mode, TSM uses features of succeeding frames to use in the current frame that’s why it is also known as bi-directional shifts (figure 2). In online mode TSM performs smart functions as it uses preceding frame feature for current frames, in this way our model performs shift in uni-direction. The online video recognition gives us low latency as we use only a minor part of frames to get features whereas some methods [93] uses large number of frames to predict one result which leads to high time-complexity. Low memory consumption will be there as we are saving only that minor part of frames in cache memory.

5.3.5 Transfer Learning. In HAR transfer learning is emerging because of many important applications such as disorder of movement, obesity, and patient monitoring from distance in some deadly diseases [98, 99]. This method starts with clustering of videos as much as possible to segregate similar actions and to predict them in clusters. Then, the analyzing of CNN architecture layers will be done by using CCA similarity [100]. From the report generated we will be able to know what features to transfer and tune the layers again for more accuracy.

5.3.6 BERT in 3D CNN architecture

5.3.6.1 Temporal Modelling. In this method, researchers first perform classification with the help of 3D CNN excluding the BERT based temporal model and they extracted features from the output of classification. Then to preserves the information they use BERT classification approach by adding those extracted features to the parameters of BERT encoding and also additional classification features appended in the results [101].

5.3.6.2 Feature Reduction. BERT is producing too many parameters to be worked on. As if there are approx. 500 feature dimensions then the BERT will have approx. 2-3 million of parameters and if we increase the features dimension as this will depend on the inputs, BERT parameters will increase immensely. There is the need to reduce features for this, researchers come up with two approaches i.e., with help of modified block which will be used for feature reduction (FRMB) and also an approach of feature reduction with additional block (FRAB) is also proposed along 3D CNN model backbone model. In FRMB, the backbone model gets replaced with a block which will be reducing dimensions of features. While in FRAB, they use an additional block for features dimension reduction (for e.g., ResNetXt101 as backbone of 3D CNN architecture figure 5).
Figure 5: Feature Reduction Approaches

5.3.6.3 SlowFast Architecture. SlowFast approach [102] includes both streams fast and slow. As before this approach there were different modalities for each stream, but now there will be only one architecture that will work for both streams namely fast and slow. Now, due to its different streams lead to different temporal resolutions will not allow BERT temporal approach to implement. In this, two different approaches are proposed Early Fusion and Late Fusion. In early fusion, only one BERT model is utilized, and the concatenation of streams will be done before the BERT layer. And to make this merger efficient there will be reduction of temporal resolution for fast stream. In late fusion, the outputs of two BERT layers from each stream will be concatenated as we need two different BERT models in this fusion (figure 6).

Figure 6: Implementation of both Fusion techniques
6. Human Tracking approaches and Futuristic approaches

6.1 Human Tracking
Another topic of concern for research is human tracking for video surveillance. In video surveillance this process will locate human’s trajectories, and then the system analyzes the person's trajectories for better performance. This tracking of human trajectories will help in abnormal behaviours. Human tracking [33] can be performed using two approaches, i.e., (i) kernel-based approach and (ii) filtering-based approach.

6.1.1 Filter Based Tracking. In the filter-based method, there are two most used techniques, i.e., Particle filter (PF) and Kalman filter (KF), and also filter based approach is itself a widely used technique to track the humans. Gaussian noise perturbed a state estimation method i.e., KF and it is based on linear dynamical systems. Furthermore, for non-gaussian and multimodality cases, we have particle filter, it is a conditional density propagation method.

6.1.2 Kernel Based Tracking. In the kernel-based tracking, the system uses color histograms of weighted objects to track the objects from one frame to another over time, and this process is iteratively performed on mean shift procedure is also called mean-shift tracking. The mean shift procedure provides the fast speed of convergence and low requirements of computation.

6.2 Further Discussion
In this section, we will be discussing three futuristic solutions, applications, and advancements.
First, the biggest challenge in HAR is computational power as we have a computational constrained system, and to overcome this, we can go for embedded sensors, microchips in hardware. Furthermore, we can go for more advanced classification approaches to recognize activity in patches from a computer vision perspective. We can also degrade the quality of the input image and focus on balancing algorithm efficiency, information, and accuracy rate.
Second, many tasks in HAR is done on single datasets, but we can go for evolution to combine many large datasets and make complex and single data. This evolution will make our results more efficient, and also it will require advanced algorithms to train the system. Another evolution is to compare the algorithms and unite them to make a more successful model, and it will also work for real world inputs.
Third, we can make our results more attractive by including the behaviour of the surroundings like instead of saying a person is running, we should rather say a person is trying to catch the bus; instead of saying many people sitting and talking, we should rather say a meeting is going on. We should include Natural Language Processing (NLP), which will also provide us with behavioural context with actions and activities.

7. Conclusion
Human activity recognition is a very prominent topic and widely used in video surveillance, healthcare, and human-computer interaction. As technology advances, crime increases, so we need to advance our HAR application for better law regulation. Furthermore, challenges will increase when we come to real scenarios.
In this review, human activities are divided into three categories of different forms of performing actions, first action primitives, then actions/activities, and thirdly interactions. We have discussed the representation of human activity recognition, classification approaches, and the category of different approaches of HAR. We have also summarized the classification approaches, and we have taken from the beginning to the latest technology, i.e., deep learning approaches. We have gone through an advanced application of HAR, i.e., human tracking.
As there are many surveys on human activity recognition, none of them adequately describe human activity recognition as they discuss the computational issues in human activity recognition. At last, we can conclude that, though humans keep on understanding different methods for HAR, there are more severe challenges in real scenarios such as human poses detection, handling background errors, occlusion, and many more.
References

[1] Gupta, A., and Davis, L. S. (2007). “Objects in action: an approach for combining action understanding and object perception,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Minneapolis, MN), 1–8.

[2] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, “A survey of video datasets for human action and activity recognition,” Computer Vision and Image Understanding, vol. 117, pp. 633–659, 2013

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, pp. 1395–1402, Beijing, China, 2005

[4] I. Laptev and T. Lindeberg, “Space-time interest points,” in Proceedings Ninth IEEE International Conference on Computer Vision, vol. 1, pp. 432–439, Nice, France, 2003.

[5] L. Rabiner and B. H. Juang, “An introduction to hidden Markov models,” IEEE ASSP Magazine, vol. 3, pp. 4–16, Prentice Hall, Upper Saddle River, New Jersey, 1986

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, pp. 1097–1105, Lake Tahoe, Nevada, 2012

[7] A. Krizhevsky, D. M. (1999). The visual analysis of human movement: a survey. Comput. Vis. Image Understand. 73, 82–98. doi:10.1006/cviu.1998.0716

[8] Aggarwal, J. K., and Cai, Q. (1999). Human motion analysis: a review. Comput. Vis. Image Understand. 73, 428–440. doi:10.1006/cviu.1998.0744

[9] Wang, L., Hu, W., and Tan, T. (2003). Recent developments in human motion analysis. Pattern Recognit. 36, 585–601. doi:10.1016/S0031-3203(02)00100-0

[10] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Understand. 104, 90–126. doi:10.1016/j.cviu.2006.08.002

[11] Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: a survey. Proc. IEEE Trans. Circuits Syst. Video Technol. 18, 1473–1488. doi:10.1109/TCSVT.2008.2005594

[12] Poppe, R. (2010). A survey on vision-based human action recognition. Image Vis. Comput. 28, 976–990. doi:10.1016/j.imavis.2009.11.014

[13] Chen, L., Wei, H., and Ferryman, J. (2013b). A survey on human motion analysis using depth imagery. Pattern Recognit. Lett. 34, 1995–2006. doi:10.1016/j.patrec.2013.02.006

[14] Ye, M., Zhang, Q., Wang, L., Zhu, J., Yangg, R., and Gall, J. (2013). “A survey on human motion analysis from depth data,” in Time-of-Flight and Depth Imaging, Lecture Notes in Computer Science, Vol. 8200. eds M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb (Berlin Heidelberg: Springer), 149–187.

[15] Guo, G., and Lai, A. (2014). A survey on still image based human action recognition. Pattern Recognit. Lett. 43, 1–43. doi:10.11514/122649.1922653

[16] Chen, L., Wei, H., and Ferryman, J. (2013b). A survey of human motion analysis using depth imagery. Pattern Recognit. Lett. 34, 1995–2006. doi:10.1016/j.patrec.2013.02.006

[17] Ya, M., Zhang, Q., Wang, L., Zhu, J., Yangg, R., and Gall, J. (2013). “A survey on human motion analysis from depth data,” in Time-of-Flight and Depth Imaging, Lecture Notes in Computer Science, Vol. 8200. eds M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb (Berlin Heidelberg: Springer), 149–187.

[18] Aggarwal, J. K., and Xia, L. (2014). Human activity recognition from 3D data: a review. Pattern Recognit. Lett. 48, 70–80. doi:10.1016/j.patrec.2014.04.011

[19] Pantic, M., and Rothkrantz, L. (2003). “Towards an affect-sensitive multimodal human-computer interaction,” in Proc. IEEE, Special Issue on Multimodal Human-Computer Interaction, Invited Paper, Vol. 91 (IEEE), 1370–1390.

[20] Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2006). “Human computing and machine understanding of human behaviour: a survey,” in Proc. International Conference on Multimodal Interfaces (New York, NY), 239–248
[22] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31, 39–58. doi:10.1109/TPAMI.2008.52

[23] Bousmalis, K., Mehu, M., and Pantic, M. (2013a). Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: a survey of related cues, databases, and tools. Image Vis. Comput. 31, 203–221. doi:10.1016/j.imavis.2012.07.003

[24] Rodríguez, N. D., Cuéllar, M. P., Lilius, J., and Calvo-Flores, M. D. (2014). A survey on ontologies for human behaviour recognition. ACM Comput. Surv. 46, 1–33. doi:10.1145/2523819

[25] Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE 90, 1151–1163. doi:10.1109/JPROC.2002.801448

[26] Mumtaz, A., Zhang, W., and Chan, A. B. (2014). “Joint motion segmentation and background estimation in dynamic scenes,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 368–375.

[27] Liu, J., Yan, J., Tong, M., and Liu, Y. (2010). “A Bayesian framework for 3D human motion tracking from monocular image,” in IEEE International Conference on Acoustics, Speech and Signal Processing (Dallas, TX: IEEE), 1398–1401.

[28] Wang, H., Kläser, A., Schmid, C., and Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. 103, 60–79. doi:10.1007/s11263-012-0594-8

[29] Yan, X., Kakadiaris, I. A., and Shah, S. K. (2014). Modeling local behaviour for predicting social interactions towards human tracking. Pattern Recognit. 47, 1626–1641. doi:10.1016/j.patcog.2013.10.019

[30] Pirsiavash, H., and Ramanan, D. (2012). “Detecting activities of daily living in first-person camera views,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2847–2854.

[31] Gan, C., Wang, N., Yang, Y., Yeung, D. Y., and Hauptmann, A. G. (2015). “DevNet: a deep event network for multimedia event detection and evidence recounting,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2568–2577.

[32] Jainy, M., Gemerty, J. C., and Snoek, C. G. M. (2015). “What do 15,000 object categories tell us about classifying and localizing actions?,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 46–55.

[33] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, “Pfinder: real-time tracking of the human body,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 780–785, 1997.

[34] O. Oreifej and Z. Liu, “Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences,” in 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723, Portland, OR, USA, 2013.

[35] J. Shotton, T. Sharp, A. Kipman et al., “Real-time human pose recognition in parts from single depth images,” Communications of the ACM, vol. 56, pp. 116–124, 2013.

[36] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern, pp. 9–14, San Francisco, CA, USA, 2010

[37] Shabani, A. H., Clausi, D., and Zelek, J. S. (2011). “Improved spatio-temporal salient feature detection for action recognition,” in Proc. British Machine Vision Conference (Dundee), 1–12.

[38] Li, R., and Zickler, T. (2012). “Discriminative virtual views for cross-view action recognition,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2855–2862

[39] Li, B., Camps, O. I., and Sznaier, M. (2012). “Cross-view activity recognition using hankelets,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition
[40] Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, I. A. (2013). “Action recognition by matching clustered trajectories of motion vectors,” in Proc. International Conference on Computer Vision Theory and Applications (Barcelona), 112–117.

[41] Lan, T., Wang, Y., and Mori, G. (2011). “Discriminative figure-centric models for joint action localization and recognition,” in Proc. IEEE International Conference on Computer Vision (Barcelona), 2003–2010.

[42] Iosifidis, A., Tefas, A., and Pitas, I. (2012a). Activity-based person identification using fuzzy representation and discriminant learning. IEEE Trans. Inform. Forensics Secur. 7, 530–542. doi:10.1109/TIFS.2011.2175921

[43] Morariu, V. I., and Davis, L. S. (2011). “Multi-agent event recognition in structured scenarios,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3289–3296.

[44] Chen, C. Y., and Grauman, K. (2012). “Efficient activity detection with maxsubgraph search,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1274–1281

[45] Sigal, L., Isard, M., Hauessecker, H., and Black, M. J. (2012b). Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation. Int. J. Comput. Vis. 98, 15–48. doi:10.1007/s11263-011-0493-4

[46] Tran, K. N., Kakadiaris, I. A., and Shah, S. K. (2012). Part-based motion descriptor image for human action recognition. Pattern Recognit. 45, 2562–2572. doi:10.1016/j.patcog.2011.12.028

[47] Wu, Q., Wang, Z., Deng, F., Chi, Z., and Feng, D. D. (2013). Realistic human action recognition with multimodal feature selection and fusion. IEEE Trans. Syst. Man Cybern. Syst. 43, 875–885. doi:10.1109/TSMCA.2012.2226575

[48] Liu, N., Dellandréa, E., Tellez, B., and Chen, L. (2012b). “Associating textual features with visual ones to improve affective image classification,” in Proc. International Conference on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, Vol. 6974 (Memphis, TN), 195–204.

[49] Martinez, H. P., Yannakakis, G. N., and Hallam, J. (2014). Don’t classify ratings of affect; rank them! IEEE Trans. Affective Comput. 5, 314–326. doi:10.1109/TAFFC.2014.2352268

[50] Song, Y., Morency, L. P., and Davis, R. (2012a). “Multimodal human behaviour analysis: learning correlation and interaction across modalities,” in Proc. ACM International Conference on Multimodal Interaction (Santa Monica, CA), 27–30

[51] Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2014b). “Classifying behavioural attributes using conditional random fields,” in Proc. 8th Hellenic Conference on Artificial Intelligence, Lecture Notes in Computer Science, Vol. 8445 (Ioannina), 95–104

[52] Patron-Perez, A., Marszalek, M., Reid, I., and Zisserman, A. (2012). Structured learning of human interactions in TV shows. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2441–2453. doi:10.1109/TPAMI.2012.24

[53] Marín-Jiménez, M. J., Noz Salinas, R. M., Yeguas-Bolivar, E., and de la Blanca, N. P. (2014). Human interaction categorization by using audio-visual cues. Mach. Vis. Appl. 25, 71–84. doi:10.1007/s00138-013-0521-1

[54] Kong, Y., Jia, Y., and Fu, Y. (2014a). Interactive phrases: semantic descriptions for human interaction recognition. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1775–1788. doi:10.1109/TPAMI.2014.2303090

[55] Wang, S., Ma, Z., Yang, Y., Li, X., Pang, C., and Hauptmann, A. G. (2014). Semisupervised multiple feature analysis for action recognition. IEEE Trans. Multimedia 16, 289–298. doi:10.1109/TMM.2013.2293060

[56] Matikainen, P., Hebert, M., and Sukthankar, R. (2009). “Trajectons: action recognition through the motion analysis of tracked features,” in Workshop on Video-Oriented Object and Event Classification, in Conjunction with ICCV (Kyoto: IEEE), 514–521

[57] Raptis, M., Kokkinos, I., and Soatto, S. (2012). “Discovering discriminative action parts from
mid-level video representations,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1242–1249

[58] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos, “Vlad3: encoding dynamics of deep features for action recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1951–1960, Las Vegas, NV, USA, 2016.

[59] D. L. Vail, M. M. Veloso, and J. D. Lafferty, “Conditional random fields for activity recognition,” in Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, p. 235, Honolulu, Hawaii, 2007

[60] V. Vapnik, S. E. Golowich, and A. Smola, “On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes,” in Advances in Neural Information Processing Systems, vol. 9, pp. 841–848, Vancouver, British Columbia, Canada, 1996.

[61] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in Pattern Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, pp. 32–36, Cambridge, UK, 2004

[62] H. I. Suk, B. K. Sin, and S. W. Lee, “Hand gesture recognition based on dynamic Bayesian network framework,” Pattern Recognition, vol. 43, pp. 3059–3072, 2010.

[63] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden markov model,” in Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 379–385, Champaign, IL, USA, 1992.

[64] P. Natarajan and R. Nevatia, “Online, real-time tracking and recognition of human actions,” in 2008 IEEE Workshop on Motion and Video Computing, pp. 1–8, Copper Mountain, CO, USA, 2008.

[65] S. Hongeng and R. Nevatia, “Large-scale event detection using semi-hidden markov models,” in Proceedings Ninth IEEE International Conference on Computer Vision, pp. 1455–1462, Nice, France, 2003.

[66] N. M. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian computer vision system for modeling human interactions,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 831–843, 2000

[67] T. Darrell and A. Pentland, “Space-time gestures,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 335–340, New York, NY, USA, 1993

[68] A. Bobick and J. Davis, “An appearance-based representation of action,” in Proceedings of 13th International Conference on Pattern Recognition, pp. 307–312, Vienna, Austria, 1996.

[69] A. F. Bobick and J.W. Davis, “The recognition of human movement using temporal templates,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 257–267, 2001

[70] Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2007). Audiovisual synchronization and fusion using canonical correlation analysis.IEEE Trans. Multimedia 9, 1396–1403. doi:10.1109/TMM.2007.906583

[71] Candamo, J., Shreve, M., Goldgof, D. B., Sapper, D. B., and Kasturi, R. (2010). Understanding transit scenes: a survey on human behaviour-recognition algorithms. IEEE Trans. Intell. Transp. Syst. 11, 206–224. doi:10.1109/TITS.2009.2030963

[72] Atrey, P. K., Hossain, M. A., El-Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. 16, 345–379. doi:10.1007/s00530-010-0182-0

[73] Shivappa, S., Trivedi, M. M., and Rao, B. D. (2010). Audiovisual information fusion in human-computer interfaces and intelligent environments: a survey. Proc. IEEE 98, 1692–1715. doi:10.1109/JPROC.2010.2057231

[74] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 1725–1732

[75] Westerveld, T., de Vries, A. P., van Ballegooij, A., de Jong, F., and Hiemstra, D. (2003). A probabilistic multimedia retrieval model and its evaluation. EURASIP J. Appl. Signal Process.
Jiang, B., Martínez, B., Valstar, M. F., and Pantic, M. (2014). “Decision level fusion of domain specific regions for facial action recognition,” in Proc. International Conference on Pattern Recognition (Stockholm), 1776–1781.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014

Mahrishi, M., Hiran, K. K., Meena, G., & Sharma, P. (2020). Machine Learning and Deep Learning in Real-Time Applications. IGI Global

Efros et al., 2003; Schuldt et al., 2004; Jhuang et al., 2007; Fathi and Mori, 2008; Niebles et al., 2008

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1933–1941, 2016

Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In European Conference on Computer Vision, pages 849–866. Springer, 2016

Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 923–932, 2016

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In CVPR, volume 2, page 3, 2017

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In International conference on machine learning, pages 843–852, 2015

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision, pages 20–36. Springer, 2016

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. arXiv preprint arXiv:1701.07067, 10, 2017

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: 7092 Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 305–321, 2018

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4694–4702, 2015

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In Joint Pattern Recognition Symposium, pages 214–223. Springer, 2007

Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. arXiv preprint arXiv:1711.08496, 2017

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. arXiv preprint arXiv:1804.09066, 2018

Feng, S.; and Duarte, M. F. 2019. Few-Shot LearningBased Human Activity Recognition. arXiv preprint arXiv:1903.10416

Shin, H.-C.; Roth, H. R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; and Summers,
R. M. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5): 1285–1298

[96] Salem, M.; Taheri, S.; and Yuan, J.-S. 2018. ECG Arrhythmia Classification Using Transfer Learning from 2-Dimensional Deep CNN Features. In 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), 1–4. IEEE

[97] Salem, M.; Taheri, S.; and Yuan, J.-S. 2018. ECG Arrhythmia Classification Using Transfer Learning from 2-Dimensional Deep CNN Features. In 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), 1–4. IEEE

[98] Mahrishi M., Morwal S. (2020) “Index Point Detection and Semantic Indexing of Videos—A Comparative Review”. In: Pant M., Kumar Sharma T., Arya R., Sahana B., Zolfagharinia H. (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, vol 1154. Springer, Singapore. https://doi.org/10.1007/978-981-15-4032-5_94

[99] Heldman, D. A.; et al. 2017. Telehealth Management of Parkinson’s Disease Using Wearable Sensors: an Exploratory Study. Digital biomarkers 1(1): 43–51

[100] Morcos, A.; Raghu, M.; and Bengio, S. 2018. Insights on representational similarity in neural networks with canonical correlation. In Advances in Neural Information Processing Systems, 5732–5741

[101] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (10 2018)

[102] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2019-October, pp. 6201–6210. Institute of Electrical and Electronics Engineers Inc. (10 2019)