Human Activity Recognition with Deep Learning: Overview, Challenges & Possibilities

Pranjal*, *DoCSE, NIT Hamirpur;* Siddhartha Chauhan, *DoCSE, NIT Hamirpur*

Abstract

The growing use of sensor tools and the Internet of Things requires sensors to understand the applications. There are major difficulties in realistic situations, though, that can impact the efficiency of the recognition system. Recently, as the utility of deep learning in many fields has been shown, various deep approaches were researched to tackle the challenges of detection and recognition. We present in this review a sample of specialized deep learning approaches for the identification of sensor based human behavior. Next, we present the multi-modal sensory data and include information for the public databases which can be used in different challenge tasks for study. A new taxonomy is then suggested, to organize deep approaches according to challenges. Deep problems and approaches connected to problems are summarized and evaluated to provide an analysis of the ongoing advancement in science. By the conclusion of this research, we are answering unanswered issues and providing perspectives into the future.

Learning (artificial intelligence); Neural networks; Activity recognition; Multimodal sensors

I. Introduction

Recent advancements in the understanding of human behavior have enabled various applications, for instance intelligent homes [1], health care [2], and increased production [3]. The detection of events is vital for humanity as it tracks the actions of persons using data to track, interpret and assist computer systems in their everyday lives. There are two mainstreams of identification methods of human activity: camera devices and sensor systems. Camera technologies are using sensors to take pictures or images to understand the actions of people [4]. Sensor systems use on-body or environmental sensors to track movement information or to monitor their activity routes. Despite the privacy concerns involved with the deployment of our cameras in our personal room, our routine movements have been tracked by sensor-based devices. Furthermore, sensors gain from generality. During this span of time sensors can be built into handheld devices like tablets, watches, sunglasses and other items including vehicles, walls and furnishings by a common use of smart devices and Stuff Internet. Intruding and non-intrusively, sensors are commonly incorporated around us, recording knowledge about human activity.

A. Challenges

With recognition of human behavior several forms of computer learning have been used. However, many technological challenges still face this area. Many of the problems are associated with other areas of pattern recognition, including machine vision and analysis of natural languages, while others are specific to sensor-based behavior recognition. Below are few examples of issues that the recognition group will tackle.

- The first challenge is the question of extraction of features. The identification of operation is basically a classification concern, and it shares a similar difficulty with other classification problems, including the elimination of features. Feature selection is more difficult to identify sensor-based activity as there are differences in the inter-activities [5]. Related features of various behaviors (for example, walking and running) can be noticed. Therefore, distinctive characteristics that reflect operations are difficult to create uniquely.
- Wide annotated data samples are needed for training and assessment of learning techniques. However, gathering and annotating sensory experience data is costly and time intensive. Annotation scarcity thus poses a major obstacle to understand sensor behavior. Furthermore, it is especially difficult to collect data about any emergent or unpredictable events (e.g., accidental falling).
- Recognition of human behavior comprises three elements: consumer, time and sensor. Second, the habits of behavior depend on people. Different users may have different types of operation. Third, the definitions of operation differ over time. This is unworkable to conclude that consumers can remain static in their market habits for a long time. In addition, as modified, new behaviors may occur. Fourth, numerous sensor systems are installed in human bodies or ecosystems on an opportunistic basis. Driven by events, the structure and configuration of sensors greatly affects results. These three allows the sensory input for action identification to be heterogeneous and desperately need to be mitigated.
- One factor that threatens understanding is the nature of the data connection. Data connection refers to the number of users and the number of operations for which data is associated. The identification of behavior is driven by sophisticated data association and entails several individual challenges. Composite behaviors are the first challenge. Many activities are focused on basic tasks, such as walking and sitting. Nonetheless, synthetic tasks that consist of a series of atomic events are more practical ways to document human everyday routines. For e.g., shooting on the tap, soaping, rubbing the hands,

 doi:10.20944/preprints202102.0349.v1

turning off the tap, ”washing hands” are provided. Data segmentation is one problem powered by composite operation. An activity composite can be described as an activity sequence. Accurate task identification thus depends heavily on specific methods of data segmentation. The third challenge is posed by overlapping events. Simultaneous events occur as the individual engages concurrently in multiple tasks such as listening to a phone call while watching TV. The scope of the data interaction is often related to the multi-occupant behaviors. Recognition becomes difficult when a variety of individuals perform a series of actions, which typically occur in multi-resident situations.

- The reliability of the identification method of human behavior is another consideration which needs to be concerned. Efforts must be taken to make the program accessible to a significant number of people, since knowledge of human behavior can be multiplied in human everyday life. Next, the program would be usable to suit portable devices to provide an instant response. The problem of calculative costs will also be dealt with. Additionally, because the users' lives are constantly tracked by the identification program, there are chances of personal data leakage. Driving the device into private space is yet another matter that should be discussed.
- In comparison to photographs or text, sensory data is elusive and unreadable for action recognition. In addition, due to inherent sensor imperfections, sensory data invariably contains a lot of noise information. Therefore, accurate approaches for recognizing sensory data should be interpretable and able to recognize what aspect of the data makes identification simpler and which aspect can deteriorate that.

B. Context of Deep Learning

Many prior studies have implemented methods of machine learning in consideration of human activity [6]. We rely heavily on techniques of abstraction, including transformation of time-frequencies [7], mathematical approaches [5], and symbolic representation [8]. The derived properties are nevertheless carefully developed and heuristic. There were no standardized or systemic methods to derive distinguishable characteristics for human behaviors effectively.

In recent years, in many areas of computational vision, natural language processing and voice analysis, deep learning has increased prominence in modeling high-level abstractions of nuanced data [9]. Following early research [10]–[12], including investigating the effectiveness of deep education in the understanding of human behavior, the related studies were carried out. In addition to the eventual creation of fundamental awareness of human behavior, latest research is performed to face the unique challenges. Deep learning however, due to its sudden growth, busy progress, and lack of technical support, is facing resentful support by the researcher. It is also important to explain why profound learning in human behavior is possible and effective given the difficulties.

- Deep learning is ”deep”, the most appealing attribute. Deep model layer by layer architectures make it possible to learn scalably from easy to abstract functionality. Advanced computing tools such as GPUs often allow deep models to learn descriptive functions from complex data. The outstanding ability to understand also helps the behaviors identification system to closely evaluate multimodal sensory data and correctly identify them.
- Various neural network architectures represent multi-faceted functions. For example, convolutionary neural networks (CNNs) are able to capture multimodal sensory input locally and the local translation invariance is accurate [13]. Recurring neural networks (RNNs) remove temporal addiction and slowly acquire information over time to transmit sensory input to understand human behavior.
- Deep neural networking can be detachable and scalable into interconnected networks with a global optimization feature that promotes various profound learning strategies like profound communication learning [14], deep active education [15], a framework for deeper attention [16] and other approaches that are not systemic and effective [17], [18]. Works which take these techniques into account serve to numerous deep learning challenges.

C. Contributions

Throughout recent years , hundreds of deep learning approaches have been tested to understand human behavior. Very little is being undertaken to study recent trends in a systematic manner. Wang et al. [19] explored many fundamental approaches for the perception of visual human behavior. Nweke et al. [20] published a report on the classification and categorization of smartphone and wearable sensor-based approaches into generative, biased and mixed processes. Li et.al [21] has introduced numerous deep neural networks for the detection of radar-based behavior. Nonetheless, no research is available yet on topics such as the recent works with a view to the challenges of understanding human behavior and the creation of deep learning models and techniques which are inspired by the particular challenges. The main results of this work can be summarized as follows relative to prior surveys.

- We perform an exhaustive study of fundamental learning approaches to sensor-based perception of human behavior. To order to provide novices and seasoned scientists with an outline of recent developments and an in-depth study of the approaches that have been tested.
- In consideration of the complexities of behavior identification, we suggest a new taxonomy of profound learning approaches. The readers are invited to investigate the course of study which is of concern to them. We review the new

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

technologies and examine how deep networks and advanced learning can be used to solve challenges. In addition, we supply knowledge and extension to identify particular issues on the available public datasets. The goal of the new taxonomy is to establish a problem solving framework in the hope of providing a rough guide in the collection or creation of readers' research topics.

- We address a few topics that are barely discussed and illustrate future developments of science.

II. Sensor and Datasets

Depending on the sensor type used, the output of an activity detection system is important. We group the sensor modes into four techniques in this section: wearable sensors, ambient sensors and object sensors.

A. Wearable Sensors

Since wearable sensors can monitor body motion directly and effectively, These are used more commonly for identification of human behavior. Such devices can be combined easily with laptops, clothing and watches.

An accelerometer is a measurement device used to measure intensity Modification of the target velocity. Measurements per second (m/s^2) and Gforces (g) are per measurement unit. Normally, the sample frequency is in the range of 10 to 100 Hz. Accelerometers can be connected to various areas of a body to detect human movement, such as tail [22], arm [23], ankle [24], wrist [25], and others. A commonly used accelerometer comprises three axes. Therefore, an accelerometer can produce a marginal time sequence.

The gyroscope is a measurement device for direction and angular distance. The angular speed ratio is expressed in degrees per second. Tens to hundreds of Hz is also the sampling rate. Usually an accelerometer is installed into a gyroscope and is connected to the same body sections. Therefore, a gyroscope has three axes and thus three time series.

A magnetometer is a handheld tracker, and is usually connected to an inertial device using an accelerometer and a gyroscope. This tests the difference in a certain direction of the magnetic field. The units are Tesla (T), and is also the sample scale of tens to hundreds of Hz. Likewise, a magnetometer usually has three axes.

The electric activity produced by skeletal muscles is measured and registered using an EMG sensor. In comparison to the three different types of sensors listed above, EMG sensors have to be directly connected to human skin. As such, it's less common in typical contexts than in fine grain gestures like hand [26] or arm [27] and facial expressions. The EMG gives a univariate pulse amplitudinal loop.

ECG is another biometric instrument for the detection of behavior which measures heart-generating electrical activities. The sensor also has to directly touch the human skin. As numerous hearts of people vibrate considerably differently, ECG signals are hard to manage variations in the subject. An ECG system contains a standardized time array.

B. Ambient Sensor

Environmental sensors are typically built into the atmosphere to detect human-climate interactions. A major benefit of room sensors is that they can track multi-occupancy movements, unlike wearable sensors. The environmental sensor devices can also be used to identify indoors, which are difficult to do with wearable sensors.

WiFi is a local wireless network communication system that transmits signals to a receiver via a transmitter. The foundation of the Wi-Fi-based detection of human behavior is that human activities and positions conflict with the transmitter 's signal transmitting direction to the receiver, both through direct transmitting pathways and influencing propagation. The WiFi signal's signal intensity (RSS) is the standard for behavior detection that is best to use and calculate [28]. Nevertheless, even without a complex environmental change, RSS is not robust. Recently, a more advanced channel state (CSI) WiFi signal analysis has been widely studied for the identification of both amplitude and phase operation [29]. CSI may also be used to detect minor gestures such as lip moving [30], keystrokes [31], and heart beats [32], aside from hard behaviors such as walking or jogging.

RFID automatically detects and records tags attached to objects containing electronically-saved information using electromagnetical fields. Two RFID tags are available: active and inactive. tags are available. Active tags rely on a nearby power source (for example, a battery) to constantly relay signals observable by an RFID reader hundreds of meters by them. Passive RFID tags then capture energy from the questioning radio waves of a nearby RFID reader to transmit stored information. Passive RFID tags are thus much cheaper and lighter. The most popular RFID behavior recognition tool is RSS [2], [33]. The working process is that the actions of humans will modify the RFID reader's single power.

Like Wi-Fi or RFID, the radar broadcasters and antennas, which have transmitters and receivers to position on opposite sides, are placed on the same side of the device. The radar-based system is based on the Doppler effect [34], [35]. Recent work primarily uses Doppler spectrograms and machine learning to analyze these spectrograms [35], [36].

C. Object Sensor

Sensors are used to track individual activities themselves through wearable and environmental sensors. In addition to physical activity such as cycling, walking , jogging and other things, though, human behavior is complemented by the constant contact

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

with the natural world (e.g. drinking / eating, dining, playing, etc.) by realistic scenarios. Consequently , it is important for understanding of more nuanced human behaviors to include the knowledge of using objects.

RFID sensors are the most commonly used for defining the use of artifacts in terms of cost effectiveness, precision and ease of deployment. RFID tags need to be applied to target objects, such as cups, magazines, computers and toothpaste [37] as they serve as object sensors rather than environmental sensors. A worn RFID reader is also needed in the detection process. Taking into account both convenience and performance, braceled RFID readers are one of the most common choices [38], [39]. Benefits are often passive RFID tags, since an object needs a special RFID tag and a individual usually remains close to objects while it's used.

There are other modalities for different uses in addition to the aforementioned sensor modalities.

Current handheld apps typically have a built-in speaker pair and a microphone to identify human behaviors. Ultrasound signal propagation is conducted using the speaker, and the microphone receives ultrasound signals. The reason is that human activity will change the ultrasound and hence represent the motion information. It is particularly ideal for the detection of fine-grained human gestures as regulation of moving bodies because there is no need for external sensors and signals [40]. There are also other potential uses. In order to understand chewing behaviors, Lee et al., for example, attempted to use ultrasound signals from a pair of speakers and a microphone [41].

In comparison with the aforementioned environmental sensing modalities, the sensor relies on mechanical systems involving direct physical interaction, which utilize electromagnetic or sound waves to comprehend human behavior. Especially in smart cities or in connected systems, it can be implemented. Implanted in a clever setting, pressure receptors, such as a chair [42], a table [42], a bed [43], and the floor [44], may be placed in different locations. Tiny gestures or specific static postures can be observed due to their direct touch characteristics. This may also be appropriate for other situations, such as preparation tracking [42] and writing attitude corrections [45]. Pressure sensors can be used particularly for energy production when operating as wearable devices, and can therefore be used for self-sustaining applications [46]. The shoes [47] and wrist bands [48] and individual chests [49] are normally mounted.

Of multiple research reasons, there are several freely accessible data sources of identification of human behavior. "Everyday life" refers to people conducting ordinary everyday tasks under orders in the sense of data acquisition. Section 3 describes the problems in more depth.

III. Techniques and Associated Issues

A. Feature Extraction

1) Temporal Features: Recognition of human behavior remains a difficult challenge though progress has been made. Partly because of the wide scope of human interaction and the rich disparity between how a single task should be carried out. It is important to use roles that specifically differentiate operations. Feature extraction is an important step in understanding human interaction as it can collect contextual information in order to differentiate between specific behaviors. The accuracy of action detection methods relies significantly on the characteristics obtained from raw signals. Time features are the most common apps used for the identification of events. Certain technologies for the activity identification, including multimodal and predictive characteristics, are also explored by researchers, which go beyond time-domain technologies.

Human actions are typically a mixture of multiple repetitive fundamental motions that can last between a few seconds and several minutes. Therefore the details of human activity are represented by time series signals, given the comparatively high sensing frequency (tens to hundreds Hz). In this sense, the fundamental streaming movements appear to exhibit smooth variations, while transitions between consecutive basic movements that, in turn, cause major changes. It is important to draw useful temporal features, both within and between successive fundamental gestures, to capture these signal characteristics of human behavior.

Some researchers excel in adapting conventional approaches to derive time characteristics and use deep learning strategies to understand the behavior. Basic sign statistics and waveform characters for deep learning recognition are widely used, including mean and variation of time series signals [50]. This form of function is rugged and scalable. A more sophisticated approach for obtaining time features is by transforming the time series from the time domain into the frequency domain, to use spectral energy shifts. The short time discrete Fourier transform (STDFT) is applied to time-based signals and a time-frequency-spectral picture has been developed in [51] Jiang and Yin. CNN is then used to control the picture and understand basic daily behaviors such as walking and standing. More recently, through a combination of time frequencies and spectral functions, Laput and Harrison [52] has established a finely-grained hand movement sensor-system. It demonstrated an accuracy of 95.2% over 25 atomic hand activities of 12 people. The spectral characteristics can not only be used to detect wearable sensors, but can also be used to detect devices free of operation. Fan et al. [53] suggested the creation in the spatial angles of RFID signal of time-angle spectrum frames that would reflect spectral power differences in time.

Since the amazing ability of automated features learning is one of the most beneficial benefits of deep-learning technology, it is easy to remove temporal features from a neural network to create a deep-grade model. End-to-end learning enables and encourages the integrated learning and recognition processes. Different deep learning methods, including RNN, time CNN and their variants have been applied for the extraction of temporal information. RNN is a deep time retrieval technique in many

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

environments that is commonly used [54], [55]. Classic RNN cells have issues with the absence and acceleration of gradients, which limits the use of the EEG analysis. Used for temporal retrieval of an RNN are the Long-Short Term Memory Units (LSTM) which have solved this problem [56]. When processing sequential data [57], the depths of an efficient LSTM-based RNN must be at least two. Because the sensor signal is a continuous stream, a sliding window typically divides the raw data into discrete sections, each of them being the RNN cell input [58]. Hyperparameters need to be carefully calibrated for achieving acceptable results are the duration and moving phase of the slider pane. In the area of identification of human behavior, ongoing progress is also underway in the early use of the simple LSTM network in various RNN variants. A significant variant of RnN in different contexts, including human activity recognition, is the bidirectional LSTM (Bi-LSTM) structure which has two traditional LSTM layers for extracting temporal dynamics from the forward and backward directions. However, Guan and Plötz [28] proposed a dynamic method with multiple deep LSTM networks and demonstrated superior results on three benchmark datasets for individual networks. In addition to the RNN structure types, some scholars have also researched different RNN cells. Instead of LSTM cells, for instance, Yao et al. [59] used Gated Recursory Units (GRUs) to construct an RNN and used it to detect operation. Nevertheless, experiments have shown that the other kinds of RNN cells can not have a substantially superior classification accuracy value to the traditional LSTM cell [56]. On the other hand, GRUs are best suited for mobile devices that have minimal computing resources due to their computational performance.

For temporary functionality selection, CNN is another attractive deep learning architecture. Contrary to RNN, for streaming data segmentation, a temporal CNN does not require a sliding window. The convolution operations with small kernels are implemented explicitly in the time dimension of sensor signals to obtain local time dependencies. Any plays used one-dimensional (1D) condenses for temporarily extracting time series signal [3], [12], [60]–[62]. Multivariate time series will be generated, requiring the separate application of 1D convolutions, if many sensors or multiple axes existed. Conventional 1D CNNs are usually a constant kernel, such that signal variability can be observed within a constant time span. Taking this distance into account, Lee et al. [63] merged several CNN arrangement with different kernel sizes to reach a time-scale. However, the multi-kernel CNN structure will need greater computing energy, and the time scale to be addressed by a mere CNN is also insufficient. In comparison, a package between two CNN layers is normal, which would lead to information loss if a large time scale was desired. A greatly extended CNN, Xi et al [64] applied to time series for the solution of the problems. The CNN dilatates the dilatated convolution kernel to the sensitive convolution region (i.e., time length) without loss of resolution instead of the traditional convolution kernels. Since the dilated kernel just adds empty elements within the kernel 's components, there is no additional computational cost. In fact, the temporal difference of multiple sensing modalities (for instance , various sensors, axes or channels) is a core issue as the CNN is used in many situations to handle different modalities similarly. Ha and Choi [65] implemented a new CNN system with unique 1D CNNs for multiple modalities in learning and temporal properties dependent on modalities. Many forms of CNN variants are considered with the development of CNNs for efficient incorporation of time characteristics. The gated CNN was used by Shen et al. [66] to track audio signal everyday operation and showed better precision than the naive CNN. In a two-stream CNN system grappling with various time scales, long and others have taken residual lines. Another interesting phenomenon in a human activity culture is the creation of a broad hybrid paradigm to discuss different viewpoints on temporal dynamics. Based on the advantages of RNN and CNN [67], Ordóñez and Roggen proposed that all local and global temporal features would be mitigated. In order to provide effective regional temporal representation, Xu et al. [68] have implemented the advanced initiation CNN framework for the multiple scales of local time extraction. Yuki et al. [69] used a dual-stream ConvLSTM network with a stream covering less time and a longer time to evaluate more complex temporal hierarchical structures. Used an autoencoder to optimize software extractions first and then used the CNN-LSTM cascade to extract local and global software for Wifi operations. Gumaei et al. [70] suggested the hybrid model for managing various aspects of temporal information, which consisted of different types of recurring units (SRUs and GRUs).

2) Multi-modal Features: Recent work on identification for human behavior is usually performed using many instruments, including accelerometers , gyroscopes and magnetometers. Some work has also shown that integrating different sensing methods will yield better results than just one sensor [71]. In the area of fundamental learning-based human activity understanding, then the analysis of intermodality interactions along with knowledge intramodality is an important task. Fusion of sensing modes can be done using two strategies: Fusion function that blends various approaches in order to generate single classification feature vectors and Classifier Ensemble in which classification outputs are paired with the functions of just one modality. Münzner et al [72]. studied the manner in which deep neural networks merge profoundly for perception of multimodal behavior. In conjunction with various network phases, they have grouped the combination modes into four groups. Their research, however, focuses only on CNN architectures. In this case, we extend their concepts of feature fusion approaches to all deep learning architectures and succeed in disclosing more perspectives and unique aspects.

Early Fusion (EF) incorporates the data from all sources, irrespective of the sensing methods, at the outset. As a tactic, it can be attractive in terms of convenience, but there are no thorough parallels. By measuring the Euclidean standards x, y and z, a basic fusion solution in [63] converted the acceleration data in raw x, y and z to a vector magnitude. Gu et al. [73] horizontally mounted time-serial signals in a single 1D vector using a linear auto encoder to achieve reliable representations The intermediate layer output has been used to feed the final sound limit classification system. In comparison, Ha et al. [10] suggested that all signal sequences be vertically stacked to form a 2D matrix and 2D-CNNs specifically used to simultaneously

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

monitor local and spatial dependency over time. In [74] all the modalities have been pre processed into 2D by the authors for the raw sign series of a single modality, but only restructured and stacked around the profile to enter finally the 3D data matrices. Then a 3D-CNN was used to use inter and intra-modality features. The modern CNN is restricted to investigating the similarities of organized modalities within the neighboring region. To solve this problem, Jiang and Yin [51] arranged signal sequences of various modalities in a new structure, which requires each signal sequence to be adjacent to some other sequence, unlike the way separate information sources are normally structured. The DCNN will derive detailed associations of individual sensing axes through this organisation. Another approach is to take advantage of non-adjacent approaches without lack of information and extra costs for computing [64]. In addition to wearable sensors the detection of RFID-based operations often includes the fusion of numerous RFID signals and early fusion of CNNs [2].

Next, **Sensor Fusion (SF)** takes each modality into account separately and then, fuses various modalities. Such an architecture not only derives modality-specific data from separate sensors but also enables adaptive allocation of complexity as the modality-specific branch architectures can vary. In [75], [76] Radu et al. suggested to promote intramodality learning a dynamically linked deep neural network (DNN) architecture. Each sensor model is allocated to separate branches of DNN, and a unifying cross-sensor layer integrates both branches to unleash the information on inter-modality. Each dimension of the sensor was vertically stacked to form 2D matrices, Yao et al. [59] also generated individual CNNs to learn intra-modality relations for each 2D matrix. To order to eliminate the correlations between different sensors, the sensor-specific properties of different sensors are first flattened and placed to a new 2D matrix before integrating into a merge CNN. Choi et al. [77] suggested a more sophisticated fusion method to effectively fuse different modalities by controlling each sensor's contribution point. The authors developed a layer of trust calculations for the automatic determination of the trust score of a sensor modality and the confidence score for the corresponding parameter fusion was normalized and multiplied by pre-processed devices. Instead of fusing the sensor-specific function only later, Ha and Choi [65] suggested developing a vector of various modalities early on as well and extracting similar characteristics between the modalities along with sensor-specific characteristics. Through treating each sensor axis independently, Axis Fusion (AF) handles signal sources in more detail. This removes the conflict from various sensor axes. The late fusion channel-based (CB-LF) was addressed by [72] this way. The sensor channel in CNN can, however, be mistaken with the 'line,' so in this paper then we use the word 'axis.' A typical AF technique for each univariate time series for each sensing channel is to create a special neural network [78], [79]. Data from the final classification network is eventually combined with data representations from any source. 1D-CNNs is commonly used as each sensing channel's interactive learning network. In order to extraction of different timespecific characteristics of each axis to merge the characteristics before feeding a totally attachment plate, Dong and Han [80] suggested using separable convolution operations. The axis-specific method is a prerequisite for the analysis of the complexity of the application to handcrafted apps. For eg, in [17], the time characteristics of acceleration and gyro are represented by the FFT spectrogram image and then merged vertically in a wider picture for inter-modality features in the following DCNN. Moreover, work has integrated the depth aspect of the spectrogram images to create a 3D format [52] that can be conveniently handled as a CNN input channel by 2D CNNs.

In comparison to EF, Sensor Fusion (SF) explores individual modes first and then fuses different modalities. This architecture not only derives modality information from different sensors, but also enables the adaptive spread of complexity, since the architectures of the branches that vary. In [75], [76], Radu et al. suggested to promote intramodal learning the complete integrated deep neural network (DNN) architecture. Each sensor mode is allocated with separate DNN branches and a unifying cross-sensor layer fuses all branches to detect intermodal data. Yao et al. [59] stacked per sensor axis vertically into 2D matrices for each 2D matrix to know the intramodal relationship and constructed individual CNNs. Sensors are then flattened and stacked into a new 2D matrix before being fed to a merge CNN to obtain correlations between the sensors more specifically. Choi et al. ([77] suggested a safer solution to fusion by controlling each sensor's contribution rates, in order to effectively fuse various modalities. In order to automatically calculate the confidence score of a sensing system, the authors built a trustworthy measurement layer and then normed the confidence score and multiplied it with the features prepackaged for the following function fusion. Instead of combining only late-stage sensing characteristics, Ha and Choi [81] proposed building a vector of different modalities in the early stage and taking the similar features from different modalities along with the sensor-specific characteristics.

Axis Fusion (AF) handles signal sources in greater detail by independent treatment of each sensor axis. This prevents interaction between the different sensor axes. The Channel-based late fusion (CB-LF) was the way [72] alluded to. The sensor channel in CNNs, however, can be confused with the "channel," so we use instead the word "axis" in this post. A typical AF technique is to develop for each one of the univariable time series of each sensing channel a different neural network [78], [79]. Output to the final classification network will be eventually concatenated with information representations from all sources. 1D-CNNs are commonly used as an human sensing channel feature learning network. Dong and Han [80] suggested the use of divided turbulence operations to isolate the specific temporal characteristics for each axis and to aggregate all characteristics before a fully connected layer was introduced. The axis basic method is a prerequisite for the study of application of a deep learning to hand-crafted apps. For eg, in [17], acceleration and gyro signal time-specific features are first represented by the FFT spectrogram and then vertically merged into a larger image in order to know inter-modality features for the DCNN below. In addition, other work merged the profile images of the spectrogram in order to create a 3D format [52] which can be conveniently treated as a CNN input channel by 2D CNN's with the depth aspect.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

Compared to the AF method, **Shared filter Fusion (SFF)** filters treat the univariate time serial data of the sensor axis separately. In all time series, the same filter exists. Thus, all feedback participants affect the filters. SFF is smoother and includes less workable parameters compared to the AF way. SFF more commonly proposes arranging the raw sensing sequences into a two-dimensional matrix by piling the model dimension and then using 2D-CNN for the 2D vector with 1D filter [12], [77], [82]. The design is thus equivalent to the application of similar 1D-CNNs to different univariate time series. Although the features of all detection modes are not directly merged, the common 1D filters interact with each other.

Classifier Ensemble, in comparison to the features prior to intervention, several modalities can be combined by combining the effects of identification from each model. A variety of methods for the fusion of recognition findings to create a general lesson have been established. Guo et al. [71], for example , proposed to use MLPs in order to establish a simple classification for any sensing mode and to incorporate all classifiers by allocating ensemble weights to the level of classifying. The writers not only took into account the consistency of the consistency of identification but also emphasized the richness of the base classifiers by causing different steps. The variety in different means of addressing over-fit problems and increasing the overall potential for generalization are thus preserved and essential. Khan et al. [83] not only addressed the fall detection problem but also added an ensemble of reconstruction errors in increasing sensor modality from the auto-encoder.

Scalability of additional sensors is the most desirable advantage of the classifier ensemble process. By just configuring the entire component, a well defined model of a certain sensing modality can easily be integrated with an existing device. In the other hand, the identification paradigm can be voluntarily modified to this hardware adjustment when a sensor is withdrawn from a device. An inherent drawback of the ensemble fusion, however, is that due to the late fusion process inter-modality similarities can be underestimated.

3) Statistical Feature Extraction: In comparison, function engineering approaches may extract useful functions, such as statistical information, rather than the deep learning feature extraction. For the manual design of these apps, however, domain awareness is typically required. Qian et al. [84] have recently managed to build a DDNN to incorporate an end-to - end statistical attribute extraction method for behavior recognition. The idea of the kernel integrating distributions into a deeper architecture was encrypted so that any sequence of statistical moments could be derived as features that reflected each section of the sensors and then used in end-to-end training for the operation classification. In particular, the authors aim at designing a network f which learns statistical functions from many kernels that do not require a manual parameter tuning, i.e. $f(X) = \varphi(X)$, where X is the sensor, and φ is a function mapping function that extracts broad or even infinite-sized features from d-dimensional data space to Hilbert space H. Because the kernel embedding technique to describe an arbitrary distribution needs injective functionality mapping, the neural network will satisfy $f^{-1}(f(X)) = X$ condition for all conceivable X applications. The writers then used an autoencoder to ensure function mapping was injectable. We also added an additional MMD loss feature to allow the auto-encoder know strong data characteristic representations. Extended studies in four datasets have shown that statistical characteristics extraction methods are efficient. Although statistical features have been studied in detail, the reasons for the derived features are still undeveloped. more logical and substantive.

B. Labelling Scarcity

Section 3.1 studies recent methods of deep learning for distinguishable characteristics from sensory data. Some of these are guided processes, we can see. The need for a pool of classified data to establish the differential paradigm is one of the key characteristics of supervised learning approaches. But, for two reasons, having a large volume of accurate labeling data is not always available. The first is a expensive , time-consuming and very boring annotation process. Second, labels are subject to various noise sources, such as sensor noise, segmentation and behavior discrepancies among different individuals, which makes the annotation process error prone. Researchers have also started investigating unintended learning and semi-supervised methods to raising their reliance on huge annotated data.

1) Unsupervised Learning: Unsupervised learning is used primarily for data exploration to find correlations between data. In [21], the authors considered whether unattended learning methods could be incorporated into the recognition of business. For analysis of temporal acceleration data in [85], the algorithm for expectation maximization and the Hidden Markov model regression are applied. However the culture of identification of behavior also lacks more efficient approaches for working with sensory details that are high-dimensional and heterogeneous to identify behavior.

Deep generative structures such as Deep Belief Networks (DBNs) and autoencoders have recently become influential in unattended analysis. Multi layers of hidden modules include DBN's and autoencoders. We are helpful in separating features in large data and identifying trends. Furthermore, in comparison with discriminatory models, deep generative models are stronger against overfitting problems [85]. As a result , researchers continue to use them to retrieve unlabeled data as the processing of unlabeled data is quick and cost-effective. Erhan et al. in [86] report that a generative deep model pretraining guides the training of discrimination toward better solutions of generalization. It has thus become popular to pre-train a deep network on broad unmarked data sets in an unregulated way. The entire identification cycle can be separated into two sections. First, the input data are generated for pre-training functions by extractors that are typically deep generative models. Second, a top layer is introduced and then trained in a supervised classification process with labeled results. The weights of the function extractor can be fine-tuned during the supervised learning. In [87], for instance, pattern recognition DBN based operation is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

introduced. The unsupervised pre-training is followed by the updating of the trained weights with labelled examples available. In [81], a parallel method of pre-training has been carried out, but Restricted Boltzmann Machines (RBMs) are employed to develop an input pattern. In another work [88] Plötz et al . suggested the use, in ubiquitous computing, of autoencoders for the non-controlled learning of the function as an alternate to Principal Component Analysis (PCA). In [60], [89], [90] the authors used the autoencoder variants such as stacked auto encoders [89], stacked denoising autoencoders [73] and CNN autoencoders [90] in a single interconnected neural network for behavior recognition to incorporate supervised characteristic learning and dimensionality reduction. Bai et al. suggested in a recent work [91] a tool called Motion2Vector to transform a time movement data into an embedded motion vector within a multidimensional domain. They use a bidirectional LSTM to encode the input blocks of the temporary wrist sensing data to fit the activity recognition context. Two hidden states are connected to the embedded vectors whose representation of the input movement can be called sufficient. Earlier classifications are qualified for C4.5, K, closest neighbor, and random woodland. Experiments show that when evaluated on public data sets, this approach can achieve precision of more than 87%.

In addition, unsupervised training can not yet conduct activities separately, given the performance of deep generational models in unsupervised training for human activity identification, because unsupervised training can not recognize true labels of behavior without having labeled samples which display the basic reality. The aforesaid approaches can also be known as semi-supervised learning, which leverages both labeled and unlabeled data for neural network research.

2) Semi-supervised Learning: Because of the difficulties in collecting classified results, semisupervised training has become a recent behavior recognition phenomenon. A semisupervised approach needs less data and broad labelled training data. Why unlabeled data can be used to improve the recognition system has been an important topic. Due to its strong deep learning in the collection of data patterns, different semi-supervised training were incorporated for recognition of activities, including co-training, active learning and data enhancement.

In 1998 Blum and Mitchell recommended **co-training** [92]. It was a self-learning expansion. A slow classifier with a minimal number of classified data is initially educated in self-learning approaches. The unidentified samples are labeled with this label. The samples may be labelled and added to the labelled collection for the retraining of the classifier with high conviction. Multiple classifiers, each trained with a single view of training results, are employed in co-training. In comparison, unlabeled samples are chosen by the classifiers to be included with the marking by trust or plurality vote. For the package in instruction, the classifiers are changed, expanded. Blum and Mitchell [92] have proposed that co-training are entirely successful under three circumstances: (a) different views of training data did not correlated strongly; (b) each view provides enough details to have a reliable classifier; and (c) views are mutually redundant. Co-training is consistent with sensor-based understanding of human behavior, since different modalities can be called multiple viewpoints. Chen et al. [93] co-worked on different data methods with several classifiers. The inertia, angular velocity and magnetism are equipped in three groups. When most classifiers decide to forecast an unlisted sample, this sample is labelled and pushed into the label collection for the next exercise. The procedure is replicated before comfortable samples can be labelled or unmarked. The final label with all modalities is then trained in a new classification.

The co-tarining process is like the human learning process. Additional insights from current experience may be gained and new information used for the description and consolidation of experience. Knowledge and experience interact constantly. In the same way, co-training uses existing models for choosing new samples, and the samples continue to prepare the models for the next search. Automatic marking, however, can lead to mistakes. The creation of appropriate labeling will facilitate specifics.

Active learning in semisupervised training is another category. Unlike the self-learning and co-learning process that automatically identifies unlabeled samples, active learning requires annotators, who are usually experts or users, to manually label the data. The objective of active learning, to alleviate the burden of labeling, is to select the most informative unlabeled instances to label and improve the classification systems with those data to minimize human supervision. The most informative examples here indicate the instances where your labels are available that have the largest impact on the model. This requires an annotator, a classifier, a query technique. A limited amount of classified information is studied by the classifier; one of the most appropriate unlabeled items is chosen by a question strategy; the notifier is asked for true labels; the new labels are used for further testing and next test. The dynamic process of learning is a loop too. It ends when the stop criteria are fulfilled. In choosing the most valuable samples, there are two specific query strategies: complexity and variety. The entropy of information will quantify uncertainty. Larger entropy means greater uncertainty and better knowledge. Diversity ensures that the samples submitted will be exhaustive and the information provided is non-repeating and non-redundant. Two question methods were used in [94]. One sample should be selected with the lowest forecast and one should use the concept of co-training, but on the other hand samples that are strongly divergent among classificators should be selected.

For behavior recognition [95], [96], deep active learning methods are used. Hossain et al. [95] hold that the conventional approaches of active learning only pick the most useful samples with only a limited fraction of the data pool available. This removes a significant number of non-selected samples. Although the samples chosen are important for preparation, the samples rejected do have a major importance. They have therefore suggested a new way of combining active learning with deep learning that not only queries the most informative unmarked samples but also utilizes the less necessary samples. In the first instance, the data is clustered with K-means. Although the basic idea is to search for ideal samples like the centers of the clusters, the next samples are also discussed in this article. The studies have shown that by marking 10% of the data the approach suggested

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

can yield good performance.

The two questions of deep-active learning and identification of individuals were investigated further by Hossain and Roy [96]. The first thing is that outliers for significant samples will easily be misunderstood. In addition to informativeness, entropy can also mean outliers if the entropy is determined for the selection, since outliers are not in any of the classes. A mutual loss function to deal with this issue has therefore been suggested in [96]. In order to reduce the entropy of outliers, the loss of cross-entropy or information is jointly minimized. The second problem is how the workload of annotators is reduced, since annotatives must master domain knowledge on exact labels. For this phase, multiple annotateurs are included. We have been chosen from the user's familiarity. The collection of annotatives is based on the complexity and user interactions of the reinforcement learning algorithm. In order to evaluate users 'and annotators' relations, conceptual similarity is used. Experimental tests reveal that the exactness of the function is 8% increased and the convergence rate is higher.

The principle of restoring the model on unlabelled data labels is based on co-training with active learning. Apart from this another approach is to compile new operation data that can be used in different situations such as resource-limited or high risk environments where data collection is difficult.

The **data augmentation** with replicating data indicates that huge fake data are generated from a limited number of real data so that fake data may help the models learn. Generative Adversarial Network (GAN) is one common method. GAN was published in [97] for the first time. GAN is important in the integration of knowledge that accompanies training data delivery. A GAN consists of two elements, one generator and one discriminator. The generator produces and tests synthetic knowledge for validity through the discriminator. The goal of the generator is to produce evidence that is real enough to fool the discriminator, while the discriminator aims at defining the generator's images as false. The routines are poorly dependent on a principle of min-max. The generator and the discriminator develop their generation efficiency and discrimination jointly through preparation. GAN variants were introduced in different areas, such as the generation of languages [98] or the generation of images [99]. SensoryGANs [100] is the first work on data increment with the synthesis of sensory input. Because sensory data was heterogeneous, Wang et al. employed three task-specific GAN's for the three tasks, which may not be enough to reflect a dynamic spectrum of different task. The synthetic data was sent to the prediction classifiers with original data after the generation. This is important to remember that since this work uses profound networks, this relies on marks to ensure it is not unattended. Zhang et al. [101] proposed the use for activity recognition of semi-supervised GAN. Unlike standard GANs, the discriminator allows a $K + 1$ classification in the semisupervised GAN classification, which involves operation and false recognition of results. A prearranged distribution is given by Varian AutoEncoders (VAEs) as inputs, rather than Gaussan noises, to ensure the delivery of generated data in the authentic distribution pattern. The aim of VAEs is for the dissemination of input data to be generated. In addition, VAE++ has been suggested to ensure that the inputs for each training sample are omitted. The cumulative efficacy of activities identification is the cohesive System integrating VAE++ and semi-supervised GAN.

C. Variation in Class

A huge amount of training data is primarily made accessible by digital information technologies to add to the development of deep learning techniques. Most current work on perception of human behavior follows a supervised learning approach, requiring a large number of labeled data for the creation of a deep model. Nevertheless, it is difficult to collect such sensor data on individual events, such as those associated with dropping elderly persons. Moreover, the unconstrained data was inherently unbalanced in class. It is therefore desperately necessary for an appropriate paradigm for action identification to recognize the issue of class inequality.

Introducing the class with the highest number of samples is the most straightforward approach for working with the disparity. This approach is, however, at risk of reducing the overall number of training samples and omitting other important samples with functionality. In comparison,-new class samples of a minority of samples could not only retain all of the initial samples, but also improve the robustness of the models. [3] Grzeszick et al. used two methods of increase to address a class disparity problem: Gaussian noise disruption and interpolation. The gradual solutions may retain the ground structure, but the sensor sampling process simulates a random time jitter. They also generated more tests of the underrepresented groups to make sure that at least a sufficient amount of results are available in each school. Another way to solve this imbalance is to change the model building strategy rather than to balance the training dataset directly. In [102], Guan and Plötz used the F1 symbol as a failure to cope with imbalance rather than traditional cross-entropy. Thanks to the fact that the F1 score takes into account both the recall aspects and the accuracy aspects, groups of various samples are not considered. In addition to the imbalance of the class of original datasets, a semi-supervised framework also has a non-negligible problem as a progressive labeling of unscheduled samples can create unfair new numbers of labels in different classes. In Small Labeled Datasets, Chen et al. [93] involved class imbalance. They used a semi-supervised system, co-training, to improve the protocol of cyclical training. A pattern-preserving technique was suggested prior for the training phase of the joint teaching process in order to align testing samples across classes while also ensuring the distribution of the samples. The K-means clustering of each operation was first taken up in my latent behavioral patterns. Then, each pattern is subject to sampling. The main objective is to ensure that all the activities have a uniform number of patterns.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

D. Diversity

Most sophisticated approaches to interpretation of human behavior presume that the testing data and results are separate and transmitted in an equal way. It is nevertheless rare as sensory data are heterogeneous for behavior identification. There are three categories of the heterogeneity of sensory information. The first is the variability of consumers that derives from various patterns of motion as different individuals execute tasks. Time is the second heterogeneity. Data distributions of activities shift over time and new events can occur in a dynamic streaming system. The heterogeneity of sensors is the third group. Typically active sensors for monitoring of human activity. A limited sensor variance can cause significant sensory data disruption. Sensor instances, styles, locations and architectures in the environment can lead to heterogeneity in the sensors. In fact, where sensing devices are used unlimitedly, a differential distribution between training data and test data can be found in the three categories of heterogeneity, and a sudden decrease in accuracy in the recognition raises questions.

We quickly incorporate transfer learning before we analyze the factors that affect heterogeneity in sensory results. Transference learning is a common technique of machine learning which transfers the classification power of the learning model from a predefined to a dynamic environment. **Transfer learning** is particularly powerful to solve problems of heterogeneity. This avoids reduced model efficiency if the training data and the evaluation data are allocated differently. This question arises in the sense of the activity recognition as templates for the activity recognition are implemented in a particular set-up to practice. The source domain is the domain of transfer learning , which includes vast and annotated data and knowledge, and the goal is to use the source domain information to list the samples in the destination domain. The source domain correlates to the original setup in the behavior recognition region, and the aim domain applies to a new implementation never experienced by the system (e.g. new events, new users, new sensors). Three categorizations of heterogeneity and how state of the art approaches reduce heterogeneity are explored in more detail in the following pages. Most of them uses transfer learning approach.

1) User Diversity: The same operation will be carried out individually by different people due to biological and environmental influences. Many people walk slowly, for example, and some like to walk quicker. Data from different users are transmitted in different ways because individuals have different behavioural habits. Normally, the accuracy will be very high when models are conditioned and checked using data obtained from a certain person. This is, indeed, unworkable. There is extensive literature on customized versions for a particular consumer. System personalization has proved to be true in [103] for a single user with just a limited amount of knowledge from the target user. Clients in the area of behavior detection have lately been exploring customized deep research models for heterogeneity. Woo et al. [104] proposed a method for each person to create a model of RNN. Learning Hidden Unit contributions (LHUC) is introduced when [105] is used, with the parameters being trained by limited amounts of data, to incorporate a specific layer with few parameters for each two hidden layers of CNN. Rokni et al. [106] recommended that their transference learning models be customized. In the preparation process, a few participants (domain source) are initially assigned to CNN data gathered. At the test point, only the top layers of the CNN are fine tuned for the target users (target domain) with a limited amount of data. Aim users require annotation. GAN may also be used to deal with consumer heterogeneity. The authors provided data on the target domain with GANs directly from the source domain to enhance the classifier 's training in [107]. For people-centered sensing applications, Chen et al. [108] identified more person disparity and task specific consistence.

2) Time Diversity: Dynamic and streaming data that monitor movements of people are obtained by human behavior recognition systems. The initial training data representing a sequence of actions are obtained to train the original model in a real-world recognition system, and then the model is optimized for the potential identification of events. In long-term, more than months or even years, it is normal that the processing of sensory data will change with time. Time can lead to three problems in line with the degree of transition and the need to consider new data definitions.

The first issue in activity recognition of heterogeneity in time is **concept drift** [109]. This explains the distribution change between the field of preparation and the test field (or source and goal domain). Concept drift can be sudden or radical [110]. To accommodate drift, deep learning models should include incremental training, in order to constantly learn from newly arrived data new concepts for human activities. For eg, a multi-column bi-directional LSTM ensemble classifier has been proposed in [111]. The model is slowly using new training samples by systematic analysis. Active learning is an incremental form of learning. Active learning can look for the ground reality of streaming data structures with such samples if changes are observed of data streams. It encourages the selection of the most effective samples for the new concepts to be updated. Effective learning thus will promote deep learning models for the duration of streaming sensory data to reduce heterogeneity [15], [112]. Gudur et al. [15] therefore proposed a deep CNN in Bayesia that had dropped in order to achieve the modell's uncertainties and to select the most information points to be queried based on the strategy of unsafe queries.

Conceptual evolution refers to the creation of modern digital media behaviors. The problem of idea creation is that in the initial learning process it is not feasible to gather labelled data for any form of operation. Firstly, despite attempts, only a small range of tasks are possible in the initial training setup of an action recognition program. Secondly , people will do new things, which they never did until the first testing in the behavior recognition program (for example, first-time learning to play guitar). Thirdly, certain things like individuals falling down are difficult to capture. Both tasks will however also be done in the research or implementation process. Thus, the concepts of the new activities must still be studied during the application phase. Studies of behavior detection mechanisms that can recognise new events in the data sharing settings are important. But,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

this is complicated because of the restricted access in the implementation process to annotated data. One strategy is to break down behaviors into intermediate components, including weapons, arms , hands and thighs. This approach assumes that the middle level attributions for more training be identified by specialists, and where new tasks with new characteristics be added, the potential is limited [113].

The problem of **open set** is a hot subject at present. Prior to that, most state-of-the-art programs are for issues with the ”closed-loop,” with the same collection of activities in the instruction system and the evaluation loop. Open-set also derives from the fact that in the initial training process we will never accumulate enough tasks. However, the solutions to open-set problems are only to decide whether the research cases belong to the target activities in accordance with idea evolutions problems rather than precisely know the activities. An intuitive approach to open-set concerns is to construct a negative set to be taken into account. In [36] we suggest a deep model based on GAN. The authors build false samples for the negative range, and the GAN discriminator can be used easily as an open classification system for GAN.

3) Sensor Diversity: Wearable and environmental sensors are part of the systems used for movement detection. Because of the sensitivity of the sensors, a minor changes in the sensors will result in major changes in the data collection or transmission of the sensors. The sensors control cases, forms, locations and configurations in the environment. Different types of sensors that collect absolutely diverse kinds of data with differing formats, frequencies and scales; wearable sensors mounted to the locations of the body only can catch movements in the respective parts of the body. Device-free sensors’ environmental architectures affect signal propagation. All these considerations will lead to declines in the accuracy of identification where the classifiers are not qualified for different equipment. Seamless profound learning models are therefore important to detect behavior in the wild. [114] Shows that the characteristics acquired through profound learning models are transferable for behavior recognition through the sensor types and sensor deployments, particularly those removed from the lower stage, in keeping with previous conclusions of [115].

Also if data are gathered and only **sensor instances** are special, for example , a person substitutes for his or her smartphone for a new smartphone, accuracy in recognition will soon decrease. It is responsible both for hardware and applications. In reality, the sensor chips display variance under the same conditions because of imperfections in the production process [116]. In fact, system output differs across various mobile platforms [117]. For eg, API’s, resolutions and other variables all influence the output of sensors. Several deep learning models have been developed to solve heterogeneity problems caused by different sensor instances. Data augmentation with GANs [18] has been a notable work. The growth in data is a compromise for improved training sets to satisfy the need of a powerful profound learning paradigm for both scale and efficiency. In [18], a heterogeneity generator is developed that synthesizes heterogeneous data from numerous sensor instances at specific disturbances. The goal is to refill the training curriculum with appropriate heterogeneity. In fact, the writers implement a two-parameter heterogeneity system that monitors the variability of the exercise. This approach explores the problem of system instances heterogeneity.

Different **sensor styles and locations** on human bodies induce the variability of the sensory data as the two causes normally occur together. The wearables and IoT devices allow people to use more than one intelligent system to assist their daily lives. Yet consumers often upgrade their intelligent devices or purchase new electronic goods. Since a variety of apps are based around a common interface (e.g., iPhone and Apple Watch), the behavior detection method is chosen to recognize behaviors easily found by existing devices utilizing templates. The machines must be mounted according to styles in terms of positions on the various body locations. The user’s hand should be connected to a mobile for example while a trousers or top pockets should be placed on a laptop. This is obvious that specific devices’ body locations contribute to tremendous variations in the signals received as the signals are triggered by motions of the related areas of the body. Consequently , two problems resulting from these improvements are urgently needed to deal with the variability of the styles and locations of the sensors.Secondly, the bulk of current experiments often represent old data and new data of the same characteristics in a mediocre manner, which is unlikely because sensor styles and locations are not set. For example, the difference in KL between the CNN parameters trained by the old data and the new data is minimized in [118], respectively. To fix the above-noted problem, Akbari and Jafari [14] have established stochastic features that are not only insensitive to classification but also capable of reserving the inherent sensory data structures. The stochastic extraction function model is based on the generative autoencoder.

In fact, Wang et al. [119] questioned how to pick the right transition source positions when there are many possible sources. This problem is realistic since the intelligent devices can be put in many ways, either in the hand or in the purse, which can induce negative transfer in an incorrect range. [120] shows that the correlation between transfer learning contexts is important. [119] implies that increased correlation signals enhanced transitions between two realms. Chen et al. [121] therefore believed that data samples from similar operations, also from separate sensors, were aggregated in the storage space. They give a stratified distance to quantify distances between entities from a class point of view. Wang et al. [119] suggests a semantic distance and a kinetic distance in order to quantify domain differences where the semantic difference includes spatial connections between the two-station data and the cinematic information is associated with motion kinetic energy interactions between two domains. Device-free sensors such as WiFi and RFID are included in the **sensor model**. The transmitting signals are typically highly influenced by the architectures and the surroundings. The explanation is that the signals are naturally mirrored, refracted and diagrammed through media and obstacles like soil, glass and walls during signals. And the recipient’s spatial locations also play a part. Given the sophistication of building classification models for device-free movement detection, relatively few studies focus on how the sensors in the wild can be similarly precise. For example, the [122] adversarial network uses deep-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

feature extraction frameworks to delete knowledge about the environment and extract features that are environmentally neutral. Remember that all the approaches listed above allow data from the target domain to be labelled or unlabeled to upgrade their models. A single-plate configuration that only requires one-off testing and is appropriate to suit all situations is invaluable in real activity recognition systems. In order to catch domain independence features, Zheng, et al [123] defined the new Body Coordinate Velocity Profile (BVP). The apps reflect power flows at varying speeds and are specific to the diverse movements of the involved body sections. The findings of research have shown that BVP is useful for cross-domain learning and adapts to all forms of domain variables including consumers, sensor sizes and sensor designs. One-fit-all is a different approach to reduce the question of variability of perception of behavior.

E. Complex Activities

Given the successful use of a number of deep learning models for the recognition of human behaviors, the bulk of current work focusses on basic tasks like driving, standing and jogging. The basic operation is central and thus the semántics are smaller. By comparison, more complex behaviors can have a series of basic acts and more semanticized events, e.g., job, dinner and coffee planning, which can best represent individuals. Consequently, with the most realistic human-computer situations , it is important to consider dynamic and high-level human behavior. Because composite behavior identification involves not only human body movement but also background knowledge of the environment, it is a more difficult challenge to identify basic action. Furthermore, the design of effective experiments to capture sensor data for composite tasks often involves thorough expertise with the use of multiple sensors and plannings for applications requiring human-computer interactions.

Existing research on the identification of composite behavior can be split into two types. The first incorporates complicated and straightforward work and attempts to create a single paradigm for understanding all forms of work. Experiments [50] are, for example, built to capture both basic and complex everyday home behaviors. Since the writers used brace sensors, information on the environment, body movement and individual positions could be collected. There are twenty two easy and combined behaviors related to the following four strategies: 1) bicycle (e.g. driving, riding outside, cycling outside); 2) verbal (e.g. washing utensil and cooking); 3) moving (e.g. indoor to outdoor and going upstairs); A basic neural feedback network with several layers was developed to identify all events with a high average 90 percent test accuracy. The findings, however, are obtained in a context-dependent environment in which the experimental context is used in teaching and in research, reducing the adaptability of the proposed process.

The second approach is to find complex activity separately from simple ones and to find a mixture of a set of simple activities more in a complex operation. This hierarchical approach is more pragmatic and has a greater focus in science. Nevertheless, it remains under-explored to apply deep learning methods to this area. The teacher has developed a multi-tasking approach to learning, which seeks to consider basic and complex tasks concurrently. One of these works is [124]. In functional terms, the authors broke a composite operation into many basic acts represented by sequential signal fragments. The signal fragments are first inserted into CNNs to obtain low-level action representations which are loaded for identification of straightforward events in a softmax classifier. Around the same time, the derived CNN characteristics of both segments are used to use the connections in the LSTM network to provide a high degree of semantine operation classification. In this way, the mutual profound extractor uses the priori of basic tasks which are components of a composite operation. In comparison to joint research, [125] uses two conditional probabilistic models to infers a series of basic events and their resulting composite operation. The authors used an approximate sequence of acts to infer the composite behavior where time differences between single activities are drawn for the classification of the composite behavior. In the other hand, the predicted composite operation is used to support the next step in the basic sequence of operation. As a result, the predictions of the basic task chain as well as hybrid events during the assumption are jointly modified.

F. Data Decomposition

Since the initial sensor data is constantly streaming signals, a fixed window often acts as an reference for the detection of the operation model, separating raw sensor data sequences into parts. This is critical if the sample limitation of an experiment is to be resolved by supplying sufficient data. Ideally, a separated data section only performs a single operation, such that simulation for all the samples in one window will predict one mark. Nevertheless, where an operation shift takes place in the middle of the window, objects in one window can not necessarily carry the same name. An optimal division strategy is therefore necessary in order to improve the efficiency of action identification.

An easy approach is to empirically seek various set window sizes. But, although the bigger window offers more detail, a switch in the center of the windows decreases. Alternatively, a narrower window does not have adequate detail. In view of the above issue, [126] describes a hierarchical signal segmentation mechanism, which initially used a wide window and shrank the segmentation slowly until a single operation is in a sub-window. The specific criteria is that the classificator is smaller than a threshold between two successive frames. Unlike the hierarchical structure, some researchers have been investigating how to assign a mark specifically for each move, rather than forecasting an entire window [127], [128]. The authors employed fully - connected networks (FCNs) to accomplish this aim based on the semantic segmentation in the computer vision culture. Data of a large window size is inserted in the FCN, and a 1D CNN layer replaces the final fully connected softmax layer in which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

the map length corresponds to time steps and the number of maps correspond to a set of operation groups at each point to determine a name. Consequently, not only do the FCNs use the details of the corresponding time period but also use the data of their closest time phases. The multi-label architecture was developed by Varamin et al. in [129] to concurrently predict the amount of ongoing research and the possibility for any alternate activity in the window. The cumulative A posteriori inference (MAP) was used to determine the most possible events by integrating the multilabel tests by means of the estimated parameters obtained from the training data collection.

G. Parallel Activities

A person may perform more than one activity simultaneously, which is called competitor activities, in a real-world scenario, one after another in sequential mode. For example, when you watch a TV, you can make a phone call. A piece of data will fit multiple ground reality labels from the sensor signals perspective. Therefore, simultaneous behavior identification as a multi-label function may be abstracted. Notice that a single person executes the concurrent operation.

In addition to mutual multimodal fusion, Zhang et al. [130] have established a single fully linked bridge network for each candidate operation. That operation was separately identified by separate softmax layers on the final judgment sheet. One big downside to this type to arrangement is that with the amount of additional operations the costs of storage will rise dramatically. The writers have suggested the use of a single neuron with sigmoid activation in order to overcome the question for each operation, for each binary distinction (done or not) [131]. In addition, Okita and Inoue [132] addressed simultaneous activity identification and introduced an LSTM multi-layer structure in order to identify any operation within each layer of LSTM. There is already a very sluggish rate in testing deep learning approaches for the identification in parallel behaviors and there is scope for progress.

H. Multi-Tenant Activity

Most state-of-the-art work on identification of human behavior focuses on the observation and assistance for single-employees. Nevertheless, living and working environments typically consist of several subjects; therefore it is of special functional interest to design approaches for multi-occupant treatment. Together with occupants, the occupant carries out individual events such as eating one person, while the other watching TV and collective experiences with other individuals coming together to undertake the same activity, such as two subjects playing tennis [133]. There are primarily two kinds of activities involving several occupants. If only wearable sensors are being used for simultaneous behavior detection, it can be separated into several tasks of identification and addressed by traditional solutions; when environment or object sensors are used, the correlation of data from mapping sensed signals to the occupants that ultimately trigger the data generation is the main issue, which gets more important as the numbers are used. In the multi-occupant case , the question of data affiliation is critical because in the absence of this data is irrational and may also endanger the safety of residents in health applications. Human activities and tools are usually used in group activity; thus the meaning and the purpose of knowledge play an important role in creating strategies for recognition. While the knowledge of the multi-employee operation is of great significance, its profound research is still minimal.

In [134] wearable and ambient sensors were used to identify two occupants' community behaviors. The environmental sensors were used to collect information from the background, which is replicated by different practical indoor settings. The sensor data from various individuals were inserted separately in different RBMs and then fusioned in the group operation into a sequential network, a DBN, and an MLP. About 100% high accuracy was obtained. Nonetheless, most targeting strategies were limited by the fact that two people performed the same job together. Tran etc. [135], on the contrary, did not deter the inhabitants from behaving together. This was meant to classify behaviors individually for each resident. That RNN cell responding to one occupants' behavior identification has been generated using a multi-label RNN. Nevertheless, only ambient sensors were used by authors and no clear approach was suggested on the topic of data association.

I. Cost

While in the sensor based human activity recognition group deep learning models have demonstrated dominant precision, they are usually resources-intense. For instance, AlexNet [136], an early DCNN architecture with five CNN layers and three fully connected layers, uses 61 M (249 Mo of memory) parameters and carries out high-precision 1,5 B operations to predict. Graphical processing units (GPUs) are typically used to speed up computing in non-portable systems. GPUs are therefore very costly and power-hungry so that they aren't ideal for smartphone phones in real-time. In addition , recent research has shown that that the size of the neural network by adding additional layers and nodes is a crucial approach to optimizing model performance, which eventually raises algorithm complexity. Consequently , it is necessary and difficult to overcome the high cost of computing to ensure the identification of human behavior in real time and through profound learning models on mobile devices.

Provided that deep neural networks are more effective in the extraction of features than defects, a mixture of manmade and profound features may help minimize calculation costs. In [137] the authors used the functionality of the spectrogram with only

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2021 doi:10.20944/preprints202102.0349.v1

one CNN layer and two entirely linked layers for the identification of human behavior. The hybrid architecture demonstrated comparative efficiency of identification by testing four test datasets with state-of-the-art approaches. The author has tested the proposed approach on three separate mobile devices to verify the viability of real-time use, including two smartphones and an on-node app. The findings found that the processing time of a projection of tens of milliseconds indicated the likelihood of real-time applications. [138] also illustrates how hand-crafted apps are integrated and that a neural network is a potential solution to accomplish real-time detection of cell phone operation. As well as the hand-crafted structure and deep learning functionality of cascading [137] it proposed organizing in combination with the profound learning and hand-crafted elements in a completely integrated classifier. [137] With a limited amount of device use, this system may improve identification accuracy. Another logical scheme of rising computational complexity is refining simple neural network cells and structure. In [139], Vu et al. did not only decline the sophistication of normal LSTM, but also avoid the depletion issue by using a self-gated recurrent neurotransmitter (SGRNN). Driving time and model size were seen in their experiments with higher computing efficiency than LSTM and GRU. The time remains in hundreds of milliseconds, however, and no real-world testing is performed on mobile devices to demonstrate potential deployment in real time. Decline of the filter size as a means of minimizing device size is an effective way to maximize the use of memory and the amount of computing processes for CNN-based systems. For instance, [137] used 1D-CNNs to monitor the model size instead of 2D-CNNs. The quantization of network [140] is a more detailed method for solving both the data and device problems. Instead of cumulative numbers the weight and outputs of the active functions are limited to just two values (e.g. -1, +1). The three key benefits are to capital costs arising from network quantifying: 1) hardware consumption and complexity of the layout are significantly reduced relative to the complete and reliable networks; 2) bit-size operations are considerably more effective than traditional floating or fixed-point arithmetic;3. When a bit-specific operation is used, most accumulating multiplier (require at least hundreds of logic gates) operations, which are particularly well-suited for FPGAs and ASICs [141] can be replaced with popcount-XNOR (requiring only a single logic gate). With respect to detection accuracy below the maximum precision equivalent, the proposed binary model obtained a stronger performance tradeoff and a 9x acceleration on CPUs, and 11x power saving. The network quantification for developing a lightweight and fast-deep learning model was also investigated by Edel and Köppe [142]. Their binarized bidirectional LSTM network has reached just 2 percent lower accuracy detection than its full-precision equivalent, saving 75 percent computing energy.

J. Privacy

The primary purpose of the detection of human activity is to track human actions so that the sensors can continuously detect user interaction. Offer the adversary a possibility to obtain sensitive data like age through time series sensor data because the way an operation is performed varyes between users (due to age , sex , weight, etc.). In specific, its black-box characteristics may inadvertently expose user-discriminatory characteristics for the deep learning technique. The authors investigated the problem of privacy using CNN technology for detection of human behavior in [143]. The empirical studies indicate that while CNN is qualified only for behavior detection with a lack of cross-entropy, the learned CNN characteristics have also shown powerful abilities for consumer discrimination. When using CNN tools, which were derived effectively for operation, a basic logistic regressor would attain a high user rating accuracy of 84.7 percent and the same rating could only obtain 35.2 percent user rating accuracy for raw sensor data. Therefore, a profound learning paradigm used originally for human behavior identification will tackle the privacy leak risk.

To fix this issue, some researchers have investigated the use of an enemy failure feature to minimize the discrediting quality of the data during the training phase. For example, in order to minimize the user identification precision, Iwasawa et al. [143] proposed adding an adverse outcomes failure into the normal operation description missed. The developers of [144] and [145] have also taken the same notion to avoid loss of information. Our analysis findings indicate that the precision for confidential information has been diminished effectively. Nevertheless, the best way to secure one kind of private information such as user identity and gender may be an opponent loss feature. However, the adverse failure runs contrary to the end-to - end preparation cycle that makes it impossible to converge permanently. In view of this gap, [146] took the concept of changing the picture design from the creative culture in order to secure all private information simultaneously. The author has creatively looked at the raw sensor signals from two aspects: the ”shape” aspect which describes how an operation is conducted and is affected by similar user data, such as age , weight, sex, height, etc. They proposed that raw sensor data be converted to ”material,” but the ”stil” is the same as random noises. Therefore, all confidential information may be secured at once by the system. In comparison to the data transformation approach , data disruption is another common method of addressing the privacy dilemma. For example, Lyu et al . suggested incorporating two forms of methods for data destruction to a stronger relationship between privacy and consistency of recognition [147]: random projection and repetitive gompertz. Recently, due to its good theoretical defense of data, differential privacy has gained more interest in science. To order to maintain the -differential secrecy, Phanet al. [148] suggested to interrupt the objective functions of the conventional auto-encoder. A E-differential privacy preservation with softmax layer was developed for the classification or prediction in addition to preservation of the privacy in feature extraction layers. This method provided theoretical privacy guarantees and error limitations, unlike the above approaches.

 doi:10.20944/preprints202102.0349.v1

K. Deep Learning Models

The data collection may contain a number of different forms in a time span of the test organ (e.g. acceleration, angular speed) and the various positions (e.g. wrist, ankle). Nonetheless, only a few modalities from different locations help to classify these activities [149]. Lying, for example, is distinctive because you have horizontal (magnetism) conditions and the uphill and the forward movement of the human knee will perceive the ascendant steps. Unrelated modalities can add noise and affect the efficiency of recognition. In fact, over time , the value of modes varies. For eg, an inconsistency only occurs in a gait in a short amount of time instead of the entire time frame in a Parkinson disease detection system [150]. Intuitively, as the pieces of the body shift vigorously, the techniques show greater value.

The inner structures of deep learning networks remain unrevealed given the progress of profound learning during action recognition. In consideration of the variety of modalities and cycles, the neural networks must be understood to analyze the variables that affect model decisions. When a deep learning algorithm, for example, defines the individual, we seem to know what modality dictates the time interval. The interpretability of profound thinking approaches is also a modern development in the understanding of human behavior.

The basic principle for interpretable methods in deep learning is to assess the value of each element of the input data automatically and to achieve high precision by eliminating the unimportant components and concentrating on the important components. Indeed, the completely connected standard layers are now capable of growing the masses of smaller neurons during training automatically. Li et al. 2016deep suggested using additional layers of pooling to eliminate low weight neurons. This is, however, completely insufficient, as deep models can still encode noise such as unrelevant methods [150]. Several scientists [151], [152] have demonstrated the versatility of neural networks. After the authors discover their connections with the behaviors of the simulation, salient characteristics are sent to the following models [152]. Nutter et al. [153] converted sensory data into images, allowing for more straightforward interpretability of visualization resources for sensory data.

The process of attention mechanism got popular lately in deep learning fields. A attention mechanism is originally a biological and psychological term that explains how we focus on something that is important for improved cognitive outcomes. Driven by this, the scientists extend principles of neural attention to deep analysis to allow neural networks to concentrate on a specific input subset. As the deep focus process concept is measuring input components, components with higher weights are considered more closely related to the identification function and have a larger effect on product decisions [154]. Several experiments used a system of attention mechanism in the analysis of deep model actions [155]–[157]. With respect to the understanding of human behavior, the focus system not only illustrates the most recognizable modalities and times, it also teaches us of the most appropriate approaches and body parts for such acts. Profound treatment can be split between soft attention and hard attention depending on their distinction. In attention layers, **soft attention** uses softmax functions to measure the weight, such that the whole model is a completely differentiable deterministic system through which gradients can be propagated both to other areas of our network and through the soft attention mechanism [158]. Attention layers for feature extraction [159] are used in line-to- line LSTMs. It is often integrated into the neural networks for sliding windows to change the weight of all samples [16] as samples have different contributors to the detection of behavior at various periods. Shen et al. [66] also took the temporal sense into account. In two ways, Zeng et al. [150] has established treatment mechanisms. In order to obtain salient sensory methods and then use time to filter out inactive parts of the data, they first suggest sensor treatment for the inputs. The systems for spatial and temporal treatment are also used [160]. Spatial dependency in particular is derived by the convergence of self-attention processes.

Hard attention decides whether one aspect of the inputs will be considered or not. The weight that is assigned to an input element is 0 or 1 and the question does not vary. The method involves choosing which component. This includes creating a list. For example , the model uses certain aspects of the input to extract information and chooses where to act on the basis of experience in the next step. A neural network is available to produce the collection. Nevertheless, because the right selection policy is arbitrary, difficult attention can be used as a stochastic process. There is intense reinforcing learning. Deep enhancement programming tackles deep learning selection problems and allows models to distribute graduation in the domain of selection policies [161]. Strong concentration with softmax functions and regular gradients of bottom propagation can be equipped of deep reinforcement learning. Zhang et al. [162] are using dueling, deep Q networks as hard attention on the most relevant aspects of the sensory multimodal data. Chen et al. [93], [163] also weakened critical approaches and eliminated undesirable political characteristics. The LSTM focus is integrated into the cycle to make systematic choices as LSTM absorbs knowledge incrementally in an sequence. Chen et al. [164] have also found the inherent relationship of human body behaviors and sub-motions. They hire several agents to work on sub-motions-related modalities. To represent the operations, several officers collaborate. The visualization of the modalities and components of the body reveals that the focus system provides an insight into how sensory data elements influence the interpretation of models.

IV. Future work

Any prospective avenues for study are deserving of further work in order to build maximum capacity for deep learning in consideration of human behavior. The problems outlined in this research will inspire future directions. Many of them are not thoroughly addressed, for example class disparity, hybrid practices, overlapping events, etc given the effort they have made to

tackle these problems. Although existing work remains incapable of detailed and credible approaches to the problems, it lays the groundwork and offers inspiration for the future. Moreover, other avenues for study have hardly been explored before. We outline different key recommendations for study which must be used as a matter of urgency.

- Recognition of human behavior requires sufficiently annotated samples to train the deep learning models. Unsupervised training can contribute to mitigating those needs. To date, deep unsupervised templates for identification of human behavior are primarily used to distinguish characteristics but can not classify behaviors as there are no ground-breaking facts. One possible approach to unsupervised training to infer true labels is therefore to look for additional information that leads to a common and deep unsupervised approach of transfer learning [165]. Another way forward is to use methods based on results, such as ontology [166].
- There is a huge struggle to identify new things that have never been used in models. A robust model would be able to acquire correct learning without any simple truth and learn new skills online. A good way is to learn functionality that can be used for different activities. Although [113] shows that mid-level characteristics can be used to represent behaviors with a variety of requirements, dissolute features [167] are another helpful approach for new activities.
- Potential anticipation of events expands the knowledge of behavior. The behavior prediction method, unlike behavior detection, will forecast users' actions early. The predictive method is useful for human behavior identification so that it can be used for intelligence systems, crime monitoring and driver behaviour. The actions are generally in a certain order in certain that behavior tasks. Therefore, it is helpful to model the temporal dependency between events to predict future forecasts. For certain functions, LSTMs [168] are acceptable. LSTMs can not, however, incorporate these long-term dependencies for long-term operations. In this scenario, brain impulses [169] will help to encourage the estimation of behavior.
- Although hundreds of works have been examined in profound learning and the perception of human behavior by the sensors, state-of-the-art criteria for realistic comparisons are missing. The research conditions and assessment criteria for assessing behavior detection efficiency differ from document to document. The separation of preparation / evaluation / testing affects the outcomes of identification, while profound understanding is largely dependent on development evidence. Silent contrast is also possible for other considerations, like designing and integrating systems. This is also imperative that all studies have a mature standardization. It is worth noting that in many places such a issue is missing. To order to ensure fair contrast, ImageNet Challenge [170], for instance, specifics are clearly described. Jordao etal. [48] have carried out and tested a number of structured activities, but comprehensive and well-known standardization in the area of human behavior identification is still not possible.

V. Conclusion

The purpose of this study is to introduce to novices and advanced researchers who are involved in deep learning approaches for the identification of sensor-based human behavior. A complete survey is presented to summarize the current methods of deep research for sensor-based identification of human behavior. First, we incorporate the multimodal use and universal use of sensory and public data sets for different challenges. We then outline the problems in the recognition of human behavior on the grounds of their causes and examine how broad approaches are used. We address open topics at the end of the research to offer input into the future.

References

[1] Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. Towards reading trackers in the wild: detecting reading activities by eog glasses and deep neural networks. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 704–711, 2017.

[2] Xinyu Li, Yanyi Zhang, Mengzhu Li, Ivan Marsic, JaeWon Yang, and Randall S Burd. Deep neural network for rfid-based activity recognition. In *Proceedings of the Eighth Wireless of the Students, by the Students, and for the Students Workshop*, pages 24–26, 2016.

[3] Rene Grzeszick, Jan Marius Lenk, Fernando Moya Rueda, Gernot A Fink, Sascha Feldhorst, and Michael ten Hompel. Deep neural network based human activity recognition for the order picking process. In *Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*, pages 1–6, 2017.

[4] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2019.

[5] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):1–33, 2014.

[6] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

[7] Tâm Huynh and Bernt Schiele. Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 159–163, 2005.

[8] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.

[9] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and SS Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.

[10] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. Multi-modal convolutional neural networks for activity recognition. In *2015 IEEE International conference on systems, man, and cybernetics*, pages 3017–3022. IEEE, 2015.

[11] Nicholas D Lane and Petko Georgiev. Can deep learning revolutionize mobile sensing? In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pages 117–122, 2015.

[12] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[13] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[14] Ali Akbari and Roozbeh Jafari. Transferring activity recognition models for new wearable sensors with deep generative domain adaptation. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, pages 85–96, 2019.

[15] Gautham Krishna Gudur, Prahalathan Sundaramoorthy, and Venkatesh Umaashankar. Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, pages 7–12, 2019.

[16] Vishvak S Murahari and Thomas Plötz. On attention models for human activity recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 100–103, 2018.

[17] Chihiro Ito, Xin Cao, Masaki Shuzo, and Eisaku Maeda. Application of cnn for human activity recognition with fft spectrogram of acceleration and gyro sensors. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1503–1510, 2018.

[18] Akhil Mathur, Tianlin Zhang, Sourav Bhattacharya, Petar Velickovic, Leonid Joffe, Nicholas D Lane, Fahim Kawsar, and Pietro Lió. Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 200–211. IEEE, 2018.

[19] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[20] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-Garadi, and Uzoma Rita Alo. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105:233–261, 2018.

[21] Fei Li and Schahram Dustdar. Incorporating unsupervised learning in activity recognition. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[22] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3, 2013.

[23] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection. In *European Conference on Wireless Sensor Networks*, pages 17–33. Springer, 2008.

[24] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. *arXiv preprint arXiv:2004.08821*, 2020.

[25] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 10–19, 2008.

[26] Muhammad Zia ur Rehman, Asim Waris, Syed Omer Gilani, Mads Jochumsen, Imran Khan Niazi, Mohsin Jamil, Dario Farina, and Ernest Nlandu Kamavuako. Multiday emg-based classification of hand motions with deep learning techniques. *Sensors*, 18(8):2497, 2018.

[27] Jian Wu, Zhongjun Tian, Lu Sun, Leonardo Estevez, and Roozbeh Jafari. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6. IEEE, 2015.

[28] Yu Gu, Lianghu Quan, and Fuji Ren. Wifi-assisted human activity recognition. In *2014 IEEE Asia Pacific Conference on Wireless and Mobile*, pages 60–65. IEEE, 2014.

[29] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee. A survey on behavior recognition using wifi channel state information. *IEEE Communications Magazine*, 55(10):98–104, 2017.

[30] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. We can hear you with wi-fi! *IEEE Transactions on Mobile Computing*, 15(11):2907–2920, 2016.

[31] Kamran Ali, Alex X Liu, Wei Wang, and Muhammad Shahzad. Keystroke recognition using wifi signals. In *Proceedings of the 21st annual international conference on mobile computing and networking*, pages 90–102, 2015.

[32] Xuyu Wang, Chao Yang, and Shiwen Mao. Phasebeat: Exploiting csi phase data for vital sign monitoring with commodity wifi devices. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 1230–1239. IEEE, 2017.

[33] Lina Yao, Quan Z Sheng, Xue Li, Tao Gu, Mingkui Tan, Xianzhi Wang, Sen Wang, and Wenjie Ruan. Compressive representation for device-free activity recognition with passive rfid signal strength. *IEEE Transactions on Mobile Computing*, 17(2):293–306, 2017.

[34] Xinyu Li, Yuan He, and Xiaojun Jing. A survey of deep learning-based human activity recognition in radar. *Remote Sensing*, 11(9):1068, 2019.

[35] Mehmet Saygın Seyfioğlu, Ahmet Murat Özbayoğlu, and Sevgi Zubeyde Gürbüz. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. *IEEE Transactions on Aerospace and Electronic Systems*, 54(4):1709–1723, 2018.

[36] Yang Yang, Chunping Hou, Yue Lang, Dai Guan, Danyang Huang, and Jinchen Xu. Open-set human activity recognition based on micro-doppler signatures. *Pattern Recognition*, 85:60–69, 2019.

[37] Michael Buettner, Richa Prasad, Matthai Philipose, and David Wetherall. Recognizing daily activities with rfid-based sensors. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 51–60, 2009.

[38] Kenneth P Fishkin, Matthai Philipose, and Adam Rea. Hands-on rfid: Wireless wearables for detecting use of objects. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*, pages 38–41. IEEE, 2005.

[39] Joshua R Smith, Kenneth P Fishkin, Bing Jiang, Alexander Mamishev, Matthai Philipose, Adam D Rea, Sumit Roy, and Kishore Sundara-Rajan. Rfid-based techniques for human-activity detection. *Communications of the ACM*, 48(9):39–44, 2005.

[40] Wenjie Ruan, Quan Z Sheng, Peipei Xu, Lei Yang, Tao Gu, and Longfei Shangguan. Making sense of doppler effect for multi-modal hand motion detection. *IEEE Transactions on Mobile Computing*, 17(9):2087–2100, 2017.

[41] Ki-Seung Lee. Joint audio-ultrasound food recognition for noisy environments. *IEEE journal of biomedical and health informatics*, 24(5):1477–1489, 2019.

[42] Jingyuan Cheng, Mathias Sundholm, Bo Zhou, Marco Hirsch, and Paul Lukowicz. Smart-surface: Large scale textile pressure sensors arrays for activity recognition. *Pervasive and Mobile Computing*, 30:97–112, 2016.

[43] Nicholas Foubert, Anita M McKee, Rafik A Goubran, and Frank Knoefel. Lying and sitting posture recognition and transition detection using a pressure sensor array. In *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, pages 1–6. IEEE, 2012.

[44] Sankar Rangarajan, Assegid Kidane, Gang Qian, Stjepan Rajko, and David Birchfield. The design of a pressure sensing floor for movement-based human computer interaction. In *European Conference on Smart Sensing and Context*, pages 46–61. Springer, 2007.

[45] Dong-Eun Lee, Sang-Min Seo, Hee-Soon Woo, and Sung-Yun Won. Analysis of body imbalance in various writing sitting postures using sitting pressure measurement. *Journal of physical therapy science*, 30(2):343–346, 2018.

[46] Sara Khalifa, Mahbub Hassan, Aruna Seneviratne, and Sajal K Das. Energy-harvesting wearables for activity-aware services. *IEEE internet computing*, 19(5):8–16, 2015.

[47] Edward S Sazonov, George Fulk, James Hill, Yves Schutz, and Raymond Browning. Monitoring of posture allocations and activities by a shoe-based wearable sensor. *IEEE Transactions on Biomedical Engineering*, 58(4):983–990, 2010.

[48] Artur Jordao, Antonio C Nazare Jr, Jessica Sena, and William Robson Schwartz. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*, 2018.

[49] A Moncada-Torres, K Leuenberger, R Gonzenbach, A Luft, and Roger Gassert. Activity classification based on inertial and barometric pressure sensors at different anatomical locations. *Physiological measurement*, 35(7):1245, 2014.

[50] Praneeth Vepakomma, Debraj De, Sajal K Das, and Shekhar Bhansali. A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities. In *2015 IEEE 12th International conference on wearable and implantable body sensor networks (BSN)*, pages 1–6. IEEE, 2015.

[51] Wenchao Jiang and Zhaozheng Yin. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1307–1310, 2015.

[52] Gierad Laput and Chris Harrison. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[53] Xiaoyi Fan, Wei Gong, and Jiangchuan Liu. Tagfree activity identification with rfids. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–23, 2018.

[54] Sicheng Li, Chunpeng Wu, Hai Li, Boxun Li, Yu Wang, and Qinru Qiu. Fpga acceleration of recurrent neural network based language model. In *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*, pages 111–118. IEEE, 2015.

[55] Dalin Zhang, Lina Yao, Kaixuan Chen, Sen Wang, Xiaojun Chang, and Yunhao Liu. Making sense of spatio-temporal preserving representations for eeg-based human intention recognition. *IEEE transactions on cybernetics*, 2019.

[56] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[57] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[58] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. Lstm networks for mobile human activity recognition. In *2016 International Conference on Artificial Intelligence: Technologies and Applications*. Atlantis Press, 2016.

[59] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web*, pages 351–360, 2017.

[60] Stefan Duffner, Samuel Berlemont, Grégoire Lefebvre, and Christophe Garcia. 3d gesture classification with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5432–5436. IEEE, 2014.

[61] Charissa Ann Ronao and Sung-Bae Cho. Deep convolutional neural networks for human activity recognition with smartphone sensors. In *International Conference on Neural Information Processing*, pages 46–53. Springer, 2015.

[62] Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244, 2016.

[63] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. Human activity recognition from accelerometer data using convolutional neural network. In *2017 ieee international conference on big data and smart computing (bigcomp)*, pages 131–134. IEEE, 2017.

[64] Rui Xi, Mengshu Hou, Mingsheng Fu, Hong Qu, and Daibo Liu. Deep dilated convolution on multimodality time series for human activity recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[65] Sojeong Ha and Seungjin Choi. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 381–388. IEEE, 2016.

[66] Yu-Han Shen, Ke-Xin He, and Wei-Qiang Zhang. Sam-gcnn: A gated convolutional neural network with segment-level attention mechanism for home activity monitoring. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 679–684. IEEE, 2018.

[67] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[68] Cheng Xu, Duo Chai, Jie He, Xiaotong Zhang, and Shihong Duan. Innohar: a deep neural network for complex human activity recognition. *Ieee Access*, 7:9893–9902, 2019.

[69] Yuta Yuki, Junto Nozaki, Kei Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. Activity recognition using dual-convlstm extracting local and global features for shl recognition challenge. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1643–1651, 2018.

[70] Abdu Gumaei, Mohammad Mehedi Hassan, Abdulhameed Alelaiwi, and Hussain Alsalman. A hybrid deep learning model for human activity recognition using multimodal body sensing data. *IEEE Access*, 7:99152–99160, 2019.

[71] Haodong Guo, Ling Chen, Liangying Peng, and Gencai Chen. Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1112–1123, 2016.

[72] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 158–165, 2017.

[73] Fuqiang Gu, Kourosh Khoshelham, Shahrokh Valaee, Jianga Shang, and Rui Zhang. Locomotion activity recognition using stacked denoising autoencoders. *IEEE Internet of Things Journal*, 5(3):2085–2093, 2018.

[74] Quang-Do Ha and Minh-Triet Tran. Activity recognition from inertial sensors with convolutional neural networks. In *International Conference on Future Data and Security Engineering*, pages 285–298. Springer, 2017.

[75] Valentin Radu, Nicholas D Lane, Sourav Bhattacharya, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Towards multimodal deep learning for activity recognition on mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 185–188, 2016.

[76] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–27, 2018.

[77] Jun-Ho Choi and Jong-Seok Lee. Confidence-based deep multimodal fusion for activity recognition. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1548–1556, 2018.

[78] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th International Conference on Mobile Computing, Applications and Services*, pages 197–205. IEEE, 2014.

[79] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *International Conference on Web-Age Information Management*, pages 298–310. Springer, 2014.

[80] Mingtao Dong, Jindong Han, Yuan He, and Xiaojun Jing. Har-net: Fusing deep representation and hand-crafted features for human activity recognition. In *International Conference On Signal And Information Processing, Networking And Computers*, pages 32–40. Springer, 2018.

[81] Nils Yannick Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. Pd disease state assessment in naturalistic environments using deep learning. In *Twenty-Ninth AAAI conference on artificial intelligence*, 2015.

[82] Tahmina Zebin, Patricia J Scully, and Krikor B Ozanyan. Human activity recognition with inertial sensors using a deep learning approach. In *2016 IEEE SENSORS*, pages 1–3. IEEE, 2016.

[83] Shehroz S Khan and Babak Taati. Detecting unseen falls from wearable devices using channel-wise ensemble of autoencoders. *Expert Systems with Applications*, 87:280–290, 2017.

[84] Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. A novel distribution-embedded neural network for sensor-based activity recognition. 2019.

[85] Dorra Trabelsi, Samer Mohammed, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE Transactions on automation science and engineering*, 10(3):829–835, 2013.

[86] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 201–208, 2010.

[87] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[88] Thomas Plötz, Nils Y Hammerla, and Patrick L Olivier. Feature learning for activity recognition in ubiquitous computing. In *Twenty-second international joint conference on artificial intelligence*, 2011.

[89] Belkacem Chikhaoui and Frank Gouineau. Towards automatic feature extraction for activity recognition from wearable sensors: a deep learning approach. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 693–702. IEEE, 2017.

[90] Ming Zeng, Tong Yu, Xiao Wang, Le T Nguyen, Ole J Mengshoel, and Ian Lane. Semi-supervised convolutional neural networks for human activity recognition. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 522–529. IEEE, 2017.

[91] Lu Bai, Chris Yeung, Christos Efstratiou, and Moyra Chikomo. Motion2vector: unsupervised learning in human activity recognition using wrist-sensing data. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 537–542, 2019.

[92] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[93] Kaixuan Chen, Lina Yao, Dalin Zhang, Xianzhi Wang, Xiaojun Chang, and Feiping Nie. A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE transactions on neural networks and learning systems*, 31(5):1747–1756, 2019.

[94] Maja Stikic, Kristof Van Laerhoven, and Bernt Schiele. Exploring semi-supervised and active learning for activity recognition. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 81–88. IEEE, 2008.

[95] HM Sajjad Hossain, MD Abdullah Al Haiz Khan, and Nirmalya Roy. Deactive: scaling activity recognition with active deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–23, 2018.

[96] HM Sajjad Hossain and Nirmalya Roy. Active deep learning for activity recognition with context aware annotator selection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1862–1870, 2019.

[97] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[98] Ofir Press, Amir Bar, Ben Bogin, Jonathan Berant, and Lior Wolf. Language generation with recurrent generative adversarial networks without pre-training. *arXiv preprint arXiv:1706.01399*, 2017.

[99] Jun-Yan Zhu and Jim Foley. Learning to synthesize and manipulate natural images. *IEEE computer graphics and applications*, 39(2):14–23, 2019.

[100] Jiwei Wang, Yiqiang Chen, Yang Gu, Yunlong Xiao, and Haonan Pan. Sensorygans: an effective generative adversarial framework for sensor-based human activity recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[101] Xiang Zhang, Lina Yao, and Feng Yuan. Adversarial variational embedding for robust semi-supervised learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 139–147, 2019.

[102] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–28, 2017.

[103] Gary Mitchell Weiss and Jeffrey Lockhart. The impact of personalization on smartphone-based activity recognition. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[104] Sungpil Woo, Jaewook Byun, Seonghoon Kim, Hoang Minh Nguyen, Janggwan Im, and Daeyoung Kim. Rnn-based personalized activity recognition in multi-person environment using rfid. In *2016 IEEE International Conference on Computer and Information Technology (CIT)*, pages 708–715. IEEE, 2016.

[105] Shinya Matsui, Nakamasa Inoue, Yuko Akagi, Goshu Nagino, and Koichi Shinoda. User adaptation of convolutional neural network for human activity recognition. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 753–757. IEEE, 2017.

[106] Seyed Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. Personalized human activity recognition using convolutional neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[107] Elnaz Soleimani and Ehsan Nazerfard. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *arXiv preprint arXiv:1903.12489*, 2019.

[108] Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. Distributionally robust semi-supervised learning for people-centric sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3321–3328, 2019.

[109] Jeffrey C Schlimmer and Richard H Granger. Incremental learning from noisy data. *Machine learning*, 1(3):317–354, 1986.

[110] Zahraa S Abdallah, Mohamed Medhat Gaber, Bala Srinivasan, and Shonali Krishnaswamy. Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)*, 51(4):1–36, 2018.

[111] Dapeng Tao, Yonggang Wen, and Richang Hong. Multicolumn bidirectional long short-term memory for mobile devices-based human activity recognition. *IEEE Internet of Things Journal*, 3(6):1124–1134, 2016.

[112] Ramyar Saeedi, Skyler Norgaard, and Assefaw H Gebremedhin. A closed-loop deep learning architecture for robust activity recognition using wearable sensors. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 473–479. IEEE, 2017.

[113] Harideep Nair, Cathy Tan, Ming Zeng, Ole J Mengshoel, and John Paul Shen. Attrinet: learning mid-level features for human activity recognition with deep belief networks. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 510–517, 2019.

[114] Francisco Javier Ordóñez Morales and Daniel Roggen. Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pages 92–99, 2016.

[115] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[116] Sanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi. Accelprint: Imperfections of accelerometers make smartphones trackable. In *NDSS*. Citeseer, 2014.

[117] Henrik Blunck, Niels Olof Bouvin, Tobias Franke, Kaj Grønbæk, Mikkel B Kjaergaard, Paul Lukowicz, and Markus Wüstenberg. On heterogeneity in mobile sensing applications aiming at representative data collection. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 1087–1098, 2013.

[118] Md Abdullah Al Hafiz Khan, Nirmalya Roy, and Archan Misra. Scaling human activity recognition via deep learning-based domain adaptation. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2018.

[119] Jindong Wang, Vincent W Zheng, Yiqiang Chen, and Meiyu Huang. Deep transfer learning for cross-domain activity recognition. In *proceedings of the 3rd International Conference on Crowd Science and Engineering*, pages 1–8, 2018.

[120] Martin Gjoreski, Stefan Kalabakov, Mitja Luš̆trek, Matjaž Gams, and Hristijan Gjoreski. Cross-dataset deep transfer learning for activity recognition. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 714–718, 2019.

[121] Yiqiang Chen, Jindong Wang, Meiyu Huang, and Han Yu. Cross-position activity recognition with stratified transfer learning. *Pervasive and Mobile Computing*, 57:1–13, 2019.

[122] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, pages 289–304, 2018.

[123] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Zero-effort cross-domain gesture recognition with wi-fi. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pages 313–325, 2019.

[124] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–16, 2018.

[125] Weihao Cheng, Sarah M Erfani, Rui Zhang, and Ramamohanarao Kotagiri. Predicting complex activities from ongoing multivariate time series. In *IJCAI*, pages 3322–3328, 2018.

[126] Ali Akbari, Jian Wu, Reese Grimsley, and Roozbeh Jafari. Hierarchical signal segmentation and classification for accurate activity recognition. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1596–1605, 2018.

[127] Rui Yao, Guosheng Lin, Qinfeng Shi, and Damith C Ranasinghe. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognition*, 78:252–266, 2018.

[128] Yong Zhang, Yu Zhang, Zhao Zhang, Jie Bao, and Yunpeng Song. Human activity recognition based on time series analysis using u-net. *arXiv preprint arXiv:1809.08113*, 2018.

[129] Alireza Abedin Varamin, Ehsan Abbasnejad, Qinfeng Shi, Damith C Ranasinghe, and Hamid Rezatofighi. Deep auto-set: A deep auto-encoder-set network for activity recognition using wearables. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 246–253, 2018.

[130] Yanyi Zhang, Xinyu Li, Jianyu Zhang, Shuhong Chen, Moliang Zhou, Richard A Farneth, Ivan Marsic, and Randall S Burd. Car-a deep learning structure for concurrent activity recognition. In *2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 299–300. IEEE, 2017.

[131] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Shuhong Chen, Ivan Marsic, Richard A Farneth, and Randall S Burd. Concurrent activity recognition with multimodal cnn-lstm structure. *arXiv preprint arXiv:1702.01638*, 2017.

[132] Tsuyoshi Okita and Sozo Inoue. Recognition of multiple overlapping activities using compositional cnn-lstm model. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 165–168, 2017.

[133] Asma Benmansour, Abdelhamid Bouchachia, and Mohammed Feham. Multioccupant activity recognition in pervasive smart home environments. *ACM Computing Surveys (CSUR)*, 48(3):1–36, 2015.

[134] Silvia Rossi, Roberto Capasso, Giovanni Acampora, and Mariacarla Staffa. A multimodal deep learning network for group activity recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[135] Son N Tran, Qing Zhang, Vanessa Smallbon, and Mohan Karunanithi. Multi-resident activity monitoring in smart homes: A case study. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 698–703. IEEE, 2018.

[136] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[137] Daniele Ravi, Charence Wong, Benny Lo, and Guang-Zhong Yang. Deep learning for human activity recognition: A resource efficient implementation on low-power devices. In *2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, pages 71–76. IEEE, 2016.

[138] Ivan Miguel Pires, Nuno Pombo, Nuno M Garcia, and Francisco Flórez-Revuelta. Multi-sensor mobile platform for the recognition of activities of daily living and their environments based on artificial neural networks. In *IJCAI*, pages 5850–5852, 2018.

[139] Toan H Vu, An Dang, Le Dung, and Jia-Ching Wang. Self-gated recurrent neural networks for human activity recognition on wearable devices. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 179–185, 2017.

[140] S Han, H Mao, and WJ Dally. Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint*, 2015.

[141] Zhan Yang, Osolo Ian Raymond, Chengyuan Zhang, Ying Wan, and Jun Long. Dfternet: Towards 2-bit dynamic fusion networks for accurate human activity recognition. *IEEE Access*, 6:56750–56764, 2018.

[142] Marcus Edel and Enrico Köppe. Binarized-blstm-rnn based human activity recognition. In *2016 International conference on indoor positioning and indoor navigation (IPIN)*, pages 1–7. IEEE, 2016.

[143] Yusuke Iwasawa, Kotaro Nakayama, Ikuko Yairi, and Yutaka Matsuo. Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training. In *IJCAI*, pages 1930–1936, 2017.

[144] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Mobile sensor data anonymization. In *Proceedings of the International Conference on Internet of Things Design and Implementation*, pages 49–58, 2019.

[145] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. In *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems*, pages 1–6, 2018.

[146] Dalin Zhang, Lina Yao, Kaixuan Chen, Guodong Long, and Sen Wang. Collective protection: Preventing sensitive inferences via integrative transformation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1498–1503. IEEE, 2019.

[147] Lingjuan Lyu, Xuanli He, Yee Wei Law, and Marimuthu Palaniswami. Privacy-preserving collaborative deep learning with application to human activity recognition. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1219–1228, 2017.

[148] NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *Aaai*, volume 16, pages 1309–1316, 2016.

[149] Yongjin Kwon, Kyuchang Kang, and Changseok Bae. Analysis and evaluation of smartphone-based human activity recognition using a neural network approach. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–5. IEEE, 2015.

[150] Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. Understanding and improving recurrent networks for human activity recognition by continuous attention. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 56–63, 2018.

[151] Eoin Brophy, Jose´ Juan Dominguez Veiga, Zhengwei Wang, Alan F Smeaton, and Tomas E Ward. An interpretable machine vision approach to human activity recognition using photoplethysmograph sensor data. *arXiv preprint arXiv:1812.00668*, 2018.

[152] Li Xue, Si Xiandong, Nie Lanshun, Li Jiazhen, Ding Renjie, Zhan Dechen, and Chu Dianhui. Understanding and improving deep neural network for activity recognition. *arXiv preprint arXiv:1805.07020*, 2018.

[153] Mark Nutter, Catherine H Crawford, and Jorge Ortiz. Design of novel deep learning models for real-time human activity recognition with mobile phones. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[154] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[155] Dalin Zhang, Kaixuan Chen, Debao Jian, and Lina Yao. Motor imagery classification via temporal attention cues of graph embedded eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[156] Dalin Zhang, Lina Yao, Kaixuan Chen, and Sen Wang. Ready for use: Subject-independent movement intention recognition via a convolutional attention model. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1763–1766, 2018.

[157] Dalin Zhang, Lina Yao, Kaixuan Chen, Sen Wang, Pari Delir Haghighi, and Caley Sullivan. A graph-based hierarchical attention model for movement intention detection from eeg signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(11):2247–2253, 2019.

[158] Dalin Zhang, Lina Yao, Kaixuan Chen, and Jessica Monaghan. A convolutional recurrent attention model for subject-independent eeg signal analysis. *IEEE Signal Processing Letters*, 26(5):715–719, 2019.
[159] Yujin Tang, Jianfeng Xu, Kazunori Matsumoto, and Chihiro Ono. Sequence-to-sequence model with attention for time series classification. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 503–510. IEEE, 2016.
[160] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI*, pages 3109–3115, 2019.
[161] Dalin Zhang, Lina Yao, Sen Wang, Kaixuan Chen, Zheng Yang, and Boualem Benatallah. Fuzzy integral optimization with deep q-network for eeg-based intention recognition. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 156–168. Springer, 2018.
[162] Xiang Zhang, Lina Yao, Chaoran Huang, Sen Wang, Mingkui Tan, Guodong Long, and Can Wang. Multi-modality sensor data classification with selective attention. *arXiv preprint arXiv:1804.05493*, 2018.
[163] Kaixuan Chen, Lina Yao, Xianzhi Wang, Dalin Zhang, Tao Gu, Zhiwen Yu, and Zheng Yang. Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.
[164] Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. Multi-agent attentional activity recognition. *arXiv preprint arXiv:1905.08948*, 2019.
[165] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.
[166] Daniele Riboni, Linda Pareschi, Laura Radaelli, and Claudio Bettini. Is ontology-based activity recognition really effective? In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 427–431. IEEE, 2011.
[167] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.
[168] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Action classification in soccer videos with long short-term memory recurrent neural networks. In *International Conference on Artificial Neural Networks*, pages 154–159. Springer, 2010.
[169] Dalin Zhang, Lina Yao, Xiang Zhang, Sen Wang, Weitong Chen, Robert Boots, and Boualem Benatallah. Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[170] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.