Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 7 Issue 1, January 2025***
(https://fcihib.journals.ekb.eg)

Survey on Multimodal Complex Human Activity Recognition

Mustafa Ezzeldin*, *Software Engineering, Faculty of Computers and Artificial Intelligence, Helwan University*
Amr Ghoneim, *Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University,*
Laila Abdelhamid, *Information System, Faculty of Computers and Artificial Intelligence, Helwan University,*
Ayman Atia, *HCI-Lab, Faculty of Computers and Artificial Intelligence, Helwan University*
mostafaezzeldin_psw@fci.helwan.edu.eg, amr.ghoneim@fci.helwan.edu.eg,
laila.abdelhamid@fci.helwan.edu.eg, aezzat@msa.edu.eg

***Abstract*—Multimodal complex human activity recognition involves the recognition and understanding of human activities using multiple modalities, such as visual, auditory, and sensor-based data. With the proliferation of smart devices and the increasing availability of multimodal data, there is a growing need for robust and efficient methods to recognize complex human activities in real-world settings. This paper presents an overview of multimodal complex human activity recognition techniques, highlighting the challenges and recent advancements in the field. This paper discusses the fusion of multimodal data sources, including visual and audio cues, as well as sensor data from wearable devices or environmental sensors. Furthermore, it explores the use of machine learning and deep learning algorithms for activity recognition and the used datasets in this field. Overall, this paper provides a comprehensive overview of techniques, challenges, and future directions in multimodal complex human activity recognition, aiming to stimulate further research in this exciting and rapidly evolving field.**

***Index Terms*—multimodal, har, complex, classification**

I. Introduction

HUMAN activity recognition (HAR) is a field of study that focuses on developing techniques and systems to automatically identify and understand human activities using data collected from various devices. It involves the analysis and interpretation of data to recognize and classify different activities performed by individuals.

The goal of HAR is to enable machines or computer systems to comprehend and respond to human activities in real-time or offline scenarios. By employing a combination of different technologies, such as cameras, accelerometers, gyroscopes, magnetometers, GPS, and microphones, HAR systems capture and process information about body movements, and environmental conditions to infer the activities being performed.

Sensor-based human activity recognition (HAR) and vision-based HAR are two distinct approaches to recognizing and understanding human activities [1]. While both methods aim to achieve the same goal, they differ in terms of the types of sensors used and the data they capture.

Sensor-based HAR relies on a combination of various sensors such as accelerometers, gyroscopes, magnetometers, and microphones. These sensors capture data related to body movements, and environmental conditions. The data collected from these sensors is processed and analyzed using machine learning algorithms to identify and classify human activities.

On the other hand, vision-based HAR focuses primarily on visual information captured by cameras or depth sensors [2]. This approach utilizes computer vision techniques to extract relevant features from the visual data and recognize human activities based on those features. Vision-based HAR often involves techniques such as object detection, pose estimation, motion tracking, and activity recognition algorithms applied to video or image sequences.

One advantage of sensor-based HAR is its ability to capture a wide range of information beyond visual cues. The additional sensors can provide data about motion, orientation, and environmental context, which can be valuable for understanding activities in different settings. Sensor-based HAR can be particularly useful in scenarios where visual information alone may not be sufficient, such as in low-light conditions or when objects are occluded.

On the other hand, vision-based HAR excels in scenarios where visual cues play a crucial role in activity recognition. It can capture detailed information about human poses, gestures, interactions with objects, and the spatial-temporal context of activities. Vision-based approaches can be effective in applications such as action recognition in sports, surveillance, human-computer interaction, and video analysis.

Both sensor-based and vision-based HAR have their strengths and limitations, and the choice between them depends on the specific requirements of the application. In some cases, a combination of both approaches may be employed to leverage the strengths of each modality and achieve more accurate and robust activity recognition.

It is worth noting that recent advancements in deep learning and the availability of large-scale datasets have enabled the development of hybrid approaches that combine sensor data and vision data for activity recognition which are called multimodal HAR [3]. These hybrid methods aim to exploit the complementary nature of sensor and vision information to improve recognition accuracy and address challenges in complex activity scenarios.

Multimodal human activity recognition refers to the process of identifying and understanding human activities by combining and analyzing information from multiple modalities [4], such as visual (e.g., images or videos), audio (e.g., speech or ambient sounds), and sensor data (e.g., accelerometer or gyroscope readings). It aims to capture a more comprehensive and accurate representation of human activities by leveraging the complementary nature of different modalities.

By combining information from multiple sensors or sources, it becomes possible to obtain a more holistic understanding

of human activities, which can lead to improved activity recognition accuracy and robustness.

The process of multimodal human activity recognition typically involves several steps. First, data is collected using different sensors or modalities. For example, in a healthcare setting, a combination of wearable sensors, video cameras, and microphones may be used to capture relevant information about a person's activities. Next, the collected data is preprocessed to remove noise, normalize the data, and extract relevant features that capture discriminative patterns or characteristics of the activities.

After preprocessing, the multimodal data is usually fused or combined to create a unified representation that captures the complementary information from different modalities. There are various fusion techniques, including early fusion, late fusion, and hybrid fusion, which determine when and how the information from different modalities is combined.

Once the fusion is performed, machine learning algorithms are applied to train models that can recognize and classify human activities. These models can be based on traditional machine learning approaches, such as support vector machines (SVMs) or decision trees, or more advanced techniques like deep learning, which have shown promising results in multimodal activity recognition tasks.

Evaluation of multimodal human activity recognition systems is typically done using labeled datasets, where the ground truth activity labels are provided. Performance metrics such as accuracy, precision, recall, and F1 score are used to assess the effectiveness of the recognition system.

Overall, multimodal human activity recognition is an interdisciplinary field that combines techniques from signal processing, computer vision, machine learning, and sensor technology. By leveraging multiple modalities, it offers the potential for more accurate, robust, and context-aware recognition of human activities, enabling a wide range of applications in various domains.

This paper explores the field of multimodal complex Human Activity Recognition (HAR) by discussing their methodologies, strengths, and limitations. Furthermore, this paper discusses the different data modalities to improve activity recognition accuracy and address challenges in complex activity scenarios. The study also delves into the process of multimodal human activity recognition, outlining steps such as data collection from various devices, data preprocessing, feature extraction, and data fusion techniques. This study presents the multimodal and complex HAR datasets and the different algorithms used to detect these activities.

II. Activities types: Simple and Complex

A. Simple Activities

Simple human activities refer to basic actions or behaviors that are relatively easy to perform and recognize [5]. These activities are typically characterized by straightforward movements, minimal complexity, and can be easily understood by observing their visual or sensorial cues. Simple activities are usually repeatable and short in duration such as walking, running and standing. Samples of simple activities are shown in Fig.1

Fig. 1. Examples of simple activities [6]

B. Complex Activities

Complex activities are not as repetitive as simple activities and may involve various hand gestures [7]; for example, eating, drinking coffee, smoking and giving a talk as shown in Fig.2. Complex activities are longer in duration when compared to the simple ones.

Complex HAR often requires the integration of multiple modalities, such as visual, audio, and sensor data, as well as the use of advanced machine learning techniques like deep learning and sequence modeling. The goal is to accurately recognize and understand the intricacies of human activities, enabling applications in areas such as sports [8], surveillance [9], healthcare [10], robotics, and human-computer interaction [8].

Fig. 2. Examples of complex activities [6]

III. Multimodal HAR types

Multimodal HAR involves the fusion of different modalities for activity data capturing. Multimodal HAR make use of the data captured from different devices such as sensors and vision devices.

These different modalities combinations are discussed in the following section. Fig.3 shows samples of data captured with different modalities.

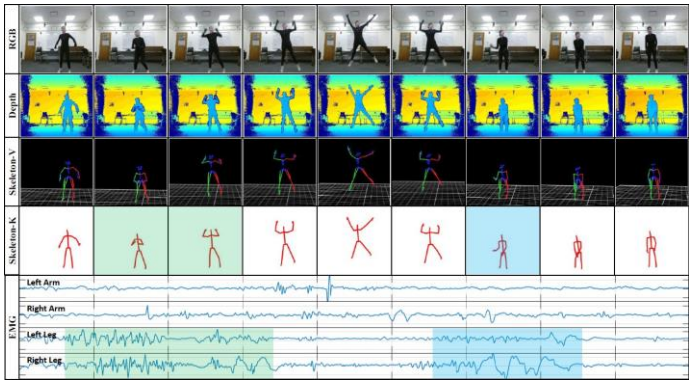

Fig. 3. Samples of data captured with different modalities [11]

A. Fusion of Depth and Inertial Sensors

Several studies have explored the concurrent utilization of depth and inertial sensors in Human Activity Recognition (HAR). Hondori et al. [12] introduced a novel approach by integrating Kinect and inertial sensors for gesture recognition tasks. Kinect technology facilitated the capture of positional and angular displacements, while inertial sensors contributed by analyzing acceleration values. Similarly, Kwolek et al. [13] devised a fuzzy interface methodology to detect falls employing both Kinect and inertial sensors. Notably, their focus was primarily on fall detection, lacking specificity in distinguishing between different activities. In another endeavor, Delachaux et al. [14] integrated Kinect with a set of five three-axis accelerometers. While accelerometers provided acceleration data, Kinect was utilized for positional information. The combined dataset was then subjected to classification using binary neural networks, demonstrating the potential of fusion strategies in HAR research.

B. Fusion of Normal RGB Camera and Inertial Sensors

The fusion of video cameras and inertial sensors has been shown to significantly enhance recognition performance compared to utilizing either modality alone [15]. In practice, internal sensors are commonly employed to capture orientation and acceleration data pertaining to various body segments, while videos are harnessed to glean positional insights [16]. Marcard et al. [17] pioneered the integration of multi-view video cameras with Inertial Measurement Units (IMUs) for motion capture applications. Within their framework, videos were instrumental in achieving precise, drift-free body positioning, whereas IMUs facilitated accurate tracking of 3D limb orientation. Similarly, Trumble et al. [18] devised a sophisticated system incorporating four video cameras alongside inertial sensors to enhance Human Activity Recognition (HAR) via visual hull estimation techniques. Leveraging Long Short-Term Memory (LSTM) networks, they effectively mitigated noise while leveraging temporal information, underscoring the advantages of multimodal fusion in HAR contexts.

C. Fusion of Multiple Wearable Sensors

Several studies have explored combining different wearable sensors for recognizing human activities. Mixing multiple sensors tends to improve recognition accuracy by capturing a wider range of features. Wilson et al. [19] utilized various binary sensors like contact switches, motion detectors, break-beam sensors, and pressure-sensitive mats to create a more accurate activity recognition system. Their system performed better than single-sensor setups, especially with a higher sampling rate [20]. Similarly, Chetty et al. [21] introduced a data analysis approach for activity recognition using both wireless body sensors and inertial sensors from smartphones. They applied algorithms that ranked features based on their importance, along with ensemble learning and classifiers like random forests and lazy learning. Additionally, they incorporated data from smartphone gyroscopes and accelerometers, demonstrating its usefulness in E-Health applications for individuals with special needs.

In the following sections, the steps of multimodal complex HAR system is described. It follows several stages starting from data gathering, pre-processing, feature extraction and classification as shown in the HAR system overview in Fig.6

IV. Data Collection

Multimodal human activity recognition (HAR) systems combine information from multiple sensors or modalities to achieve a more comprehensive understanding of human activities. Here are some devices commonly used in multimodal HAR:

Wearable Sensors [22]: Devices such as smartwatches, fitness trackers, or smart clothing embedded with sensors like accelerometers, gyroscopes, and heart rate monitors can capture motion, orientation, and physiological data. These sensors provide valuable inputs for activity recognition when combined with other modalities.

Smartphones and Tablets [23]: Smartphones and tablets have multiple built-in sensors, including accelerometers, gyroscopes, magnetometers, GPS, microphones, and cameras. They can capture visual, audio, location, and motion data, making them versatile devices for multimodal HAR. Additionally, their computing power allows for real-time data fusion and analysis.

Depth Sensors [24]: Depth sensors like the Microsoft Kinect or Intel RealSense cameras can capture both color and depth information. They provide 3D data that can enhance the recognition of human body movements and interactions with the environment.

Microphones [25]: Audio sensors, such as microphones, can capture ambient sound or specific sounds related to activities. Audio data can be processed for speech recognition, environmental sound analysis, or identifying activity-specific sounds.

Cameras [26]: Visual sensors, such as traditional cameras or RGB-D cameras, capture video or image data. They provide visual cues and information about human activities, body postures, and interactions with objects and the environment.

RFID (Radio-Frequency Identification) [27]: RFID technology uses tags and readers to identify and track objects or

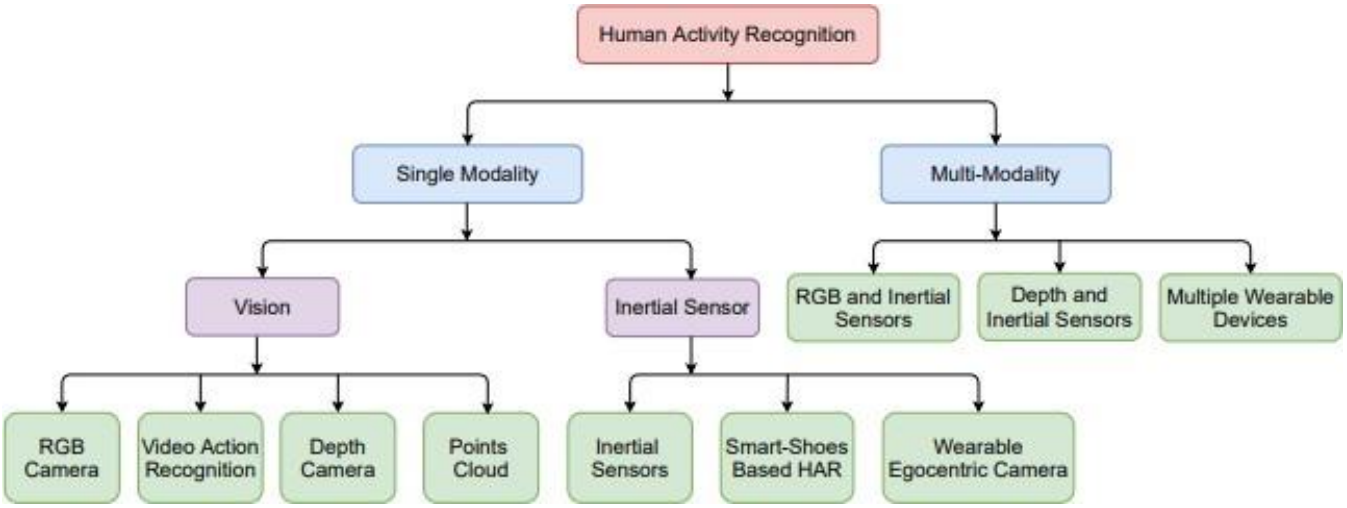

Fig. 4. Types of HAR

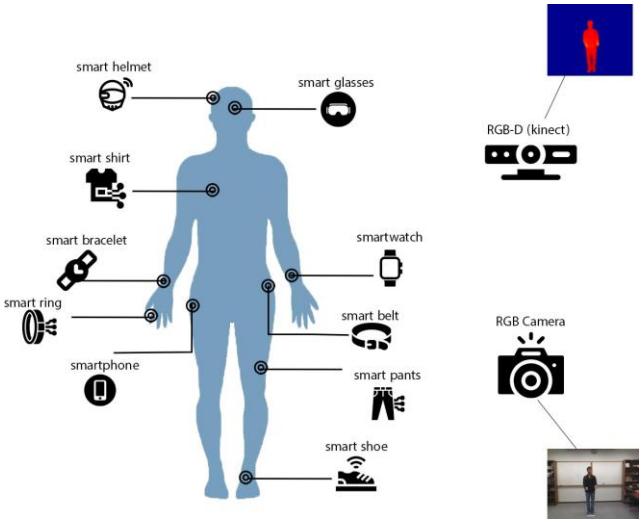

Fig. 5. Different Devices Used for Data Capturing

individuals. In multimodal HAR, RFID tags can be attached to objects or worn by individuals to provide additional contextual information about activities and object interactions.

Inertial Measurement Units (IMUs) [28]: IMUs consist of accelerometers, gyroscopes, and magnetometers integrated into a single device. They are often used in wearable devices or attached to equipment to capture motion and orientation data.

Environmental Sensors [29]: Sensors like temperature sensors, humidity sensors, or gas sensors can provide contextual information about the environment and assist in recognizing activities related to climate control, air quality monitoring, or safety.

The different placements of these different devices are shown in Fig.5.

V. Preprocessing and Windowing

Preprocessing and windowing are important steps in processing for human activity recognition (HAR). These steps help to enhance the quality of the data and prepare it for subsequent analysis and recognition algorithms. Here's an overview of sensor signals and image processing in HAR:

A. Signal processing

Noise Removal: Sensor signals may contain noise or interference that can negatively affect the accuracy of activity recognition. Preprocessing techniques such as filtering (e.g., low-pass, high-pass, or band-pass filtering) can be applied to remove noise and enhance the signal quality. Signal Conditioning: Sensor signals may need to be adjusted or calibrated to account for any biases or offsets. This can involve techniques such as zero-mean normalization or feature scaling to ensure that the signals are in a consistent range or format. Sensor Fusion: In some cases, multiple sensors (e.g., accelerometers, gyroscopes, magnetometers) may be used to capture different aspects of human activity. Sensor fusion techniques, such as data alignment, synchronization, or feature combination, can be employed to integrate the signals from multiple sensors into a unified representation.

B. Image processing

Some machine learning algorithms, like CNNs, face a limitation requiring images in the dataset to be resized to a single dimension. Therefore, images undergoing training or testing on the dataset must be preprocessed and adjusted to match uniform widths and heights prior to input into the learning algorithm.

A typical pre-processing method consists of generating modified versions of the current images and adding them to the dataset. This may involve scaling, rotating, flipping, and other alterations. The goal is to expand the dataset and expose the neural network to diverse image variations. This increases

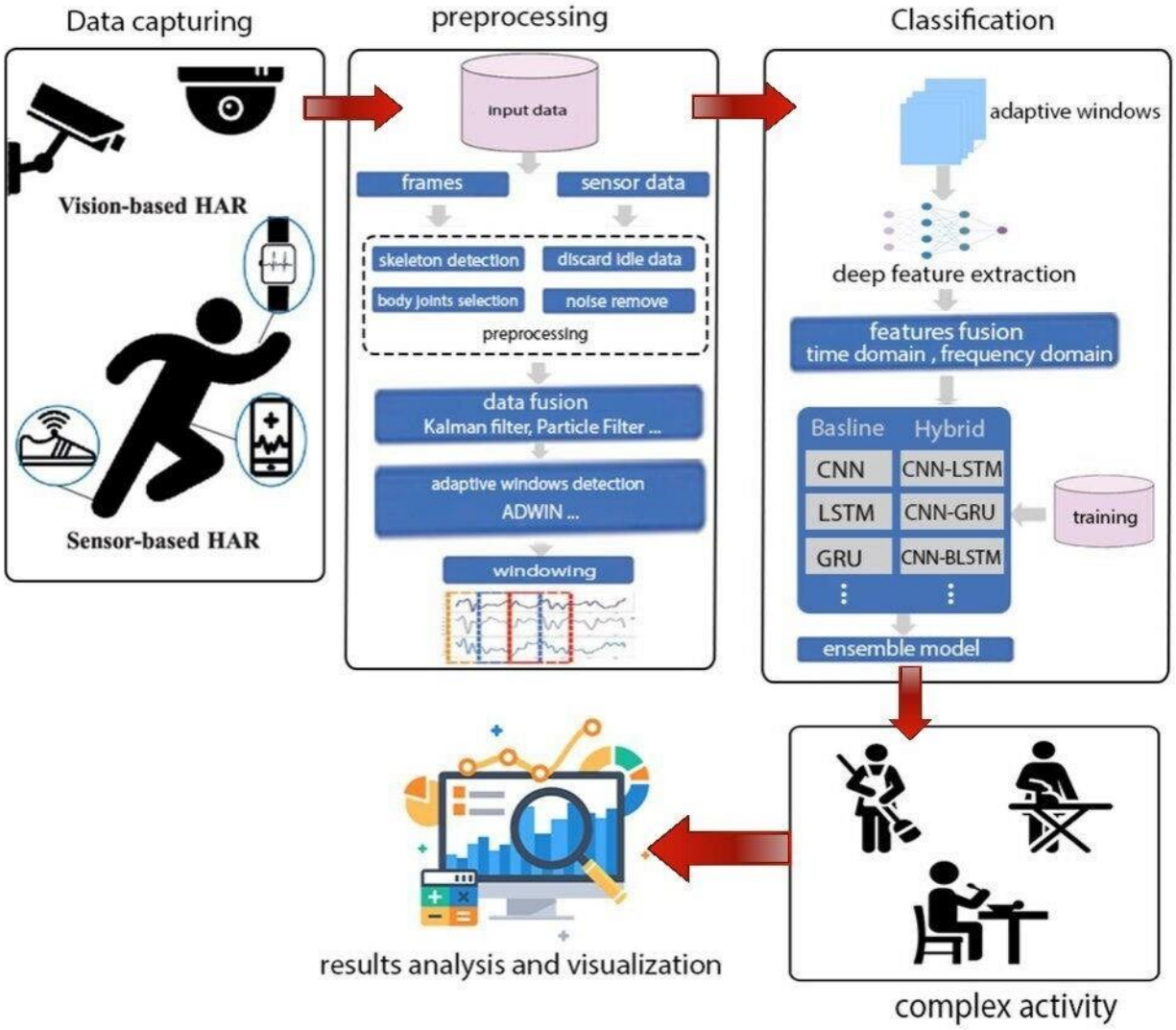

Fig. 6. System Overview of Multimodal Complex HAR

the likelihood of a machine learning model identifying objects in any form.

Image illumination significantly impacts an image's visual characteristics and plays a crucial role in automatic data extraction from images. Consequently, illumination correction is often the initial step in image analysis tasks [30]. This field has seen significant advancements in recent years, with numerous illumination correction techniques available, such as histogram equalization, homomorphic filtering, and gamma correction. These methods can improve image contrast and lighting to a certain extent [31].

Sometimes, removing extra data from images can be beneficial, either to save memory or simplify computation. A common example is changing a color image to grayscale. This is because color is often unnecessary for recognizing and understanding images, as grayscale images can still be sufficient for identifying certain objects. Since color images have more information than black-and-white ones, they can introduce unnecessary complexity and take up more storage space.

Extracting main subjects from a series of images or videos, often referred to as segmentation, involves two main steps: creating a background model and identifying the foreground. The former approach, which involves constructing background information first and then detecting objects by comparing the most recent frame to the extracted background, is particularly effective for tracking quickly moving objects captured by stationary cameras, consumes little computing resources, and is easy to set up [32]. Conversely, the latter approach, which involves extracting the foreground, is used when human activities are recorded by a pan-tilt-zoom camera or a camera mounted on moving objects, such as cars or robots.

Grey-scaled images can be transformed into binary images using thresholding, which helps in isolating areas of interest. A basic thresholding method assigns a black pixel to image pixels with intensity lower than a fixed value, called the threshold, and a white pixel otherwise. Thresholding can be categorized into two types: global thresholding, which uses the same threshold for all image pixels, and local thresholding,

which applies different thresholds to various image regions.

C. Windowing

The collected data are typically continuous time-series data. Windowing involves segmenting the data into smaller, overlapping or non-overlapping windows of fixed duration [33]. This allows for the extraction of features that capture temporal patterns and dynamics within specific time intervals. The choice of window size and overlap is crucial and depends on the characteristics of the activities being recognized. Smaller window sizes capture fine-grained details but may result in increased computational complexity, while larger window sizes provide a broader context but may lose temporal information [34]. Overlapping windows can help capture temporal dependencies between adjacent windows. Window function (e.g., Hamming, Hanning, or Gaussian) can be applied to attenuate edge effects and smooth the signal within each window. By applying preprocessing techniques, the collected data can be cleaned, normalized, and synchronized, reducing noise and inconsistencies. Windowing facilitates the partitioning of the collected data into manageable segments, enabling the extraction of features that capture relevant temporal patterns and dynamics. These preprocessed and windowed data can then be used as input for subsequent feature extraction and machine learning algorithms for human activity recognition.

D. Multimodal Fusion Methods

Fusing the information acquired by different modality sensors is a great challenge due to the dimensionality of data. For a multimodal HAR system, the data acquired can be fused at different stages. The different fusion approaches are shown in Fig.7

1) Early fusion: In the paradigm of early fusion, the amalgamation of features from various modality sensors occurs through dimensionality reduction, culminating in the formation of novel feature vectors. Evangelopoulos et al. pioneered this technique by merging textual and visual signals, meticulously analyzing each modality independently while leveraging saliency scores for both linear and non-linear fusion [35]. Similarly, Neverova et al. compressed all channels into a single dimension within the initial convolution layer, effectively minimizing the number of final parameters to be learned and consequently reducing computational overhead [36].

2) Late fusion: At the Decision Level, the late fusion strategy diverges from early fusion by segregating the data from each modality sensor, independently learning their parameters, and subsequently amalgamating their probabilistic models. This method, as espoused by [37], harnesses the individual strengths of each modality, potentially enhancing recognition outcomes. However, it entails a greater time investment and necessitates a sophisticated learning framework, raising concerns regarding potential loss of inter-modality correlation. A comprehensive evaluation of CNN-based sensor fusion methodologies for multimodal HAR conducted by [38] revealed that late and hybrid fusion techniques consistently outperform early fusion methods, as corroborated by experiments on RBK [38] and PAMAP2 [37] datasets.

3) Slow fusion (Hybrid): The slow fusion paradigm, initially introduced by Karpathy et al. [39], represents a novel approach to multimodal data fusion. This hierarchical technique synergizes elements from both early and late fusion methodologies, sequentially propagating information through successive fusion stages. While offering the advantages of both approaches, slow fusion imposes a substantial computational burden due to its multi-level information processing, necessitating careful consideration of computational resources and efficiency.

VI. Feature Extraction

Obtaining key features from pre-processed data is achieved through feature extraction, followed by feature selection, which picks a subset of features. This method is beneficial when dealing with large datasets, as it reduces resource usage without losing crucial or relevant information. Feature extraction helps minimize redundant data in the dataset, transforming it into the most significant features unique to the activity. Utilizing these features, instead of raw data, decreases the impact of noise and lessens the computational burden of classification algorithms. The typical features are presented in Table I. This section explores traditional and deep learning-based feature extraction methods for HAR.

A. Traditional Feature Extraction

Traditional feature extraction approaches depend on expertise in a particular field and signal processing methods to pull out telling features from raw data. In the context of human activity recognition (HAR), statistical features, time-domain features, frequency-domain features, and time-frequency features are frequently used. These manually crafted features are commonly input into machine learning algorithms for activity classification.

B. Deep Learning Feature Extraction

Deep learning techniques, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have achieved impressive results in automatically extracting discriminative features from raw data. Deep learning-driven feature extraction methods include:

- **Convolutional Neural Networks (CNNs)**: CNNs can automatically learn hierarchical representations of data by employing convolutional and pooling layers. They have been applied to raw data, such as time-series accelerometer data, to extract discriminative features for activity recognition.
- **Recurrent Neural Networks (RNNs)**: RNNs, including variants like long short-term memory (LSTM) networks and gated recurrent units (GRUs), are capable of capturing sequential dependencies in the data. They are well-suited for modeling temporal dynamics in activity sequences.
- **Hybrid Models**: Hybrid architectures that combine CNNs for spatial feature extraction and RNNs for temporal modeling have been proposed for HAR tasks. These

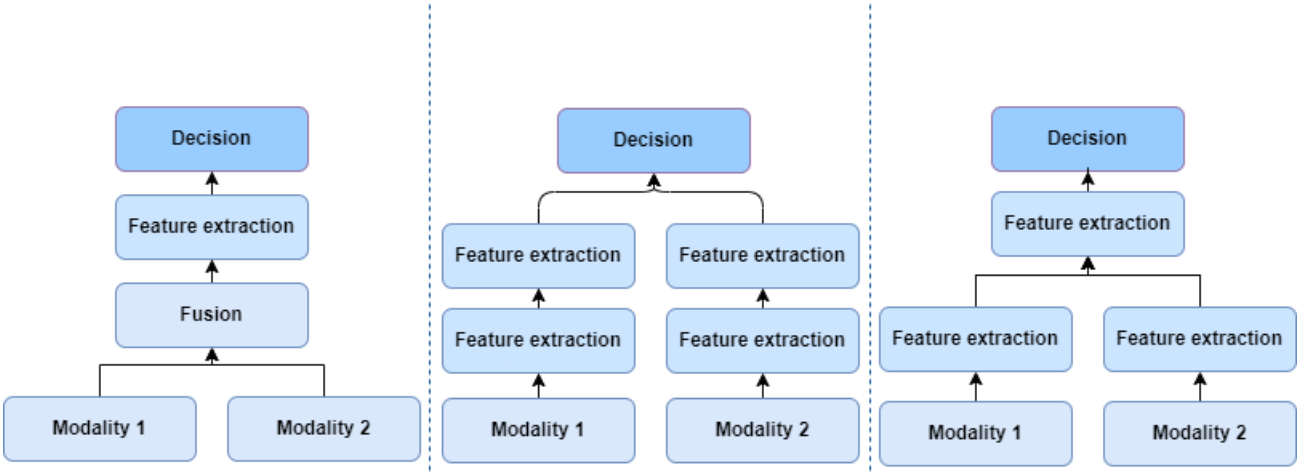

Fig. 7. Different fusion approaches for data from different modalities (a) early fusion, (b) late fusion, and (c) slow fusion

Category	Feature	Abbreviation	Equation	References
Time	Mean	M	$M = \frac{1}{N}\sum_{i=1}^{N} x_i$	[40], [41]
	Variance	V	$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$	[40], [41]
	Mean absolute deviation	MAD	$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$	[42], [43]
	Root mean square	RMS	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$	[42], [43]
	Zero Crossing Rate	ZCR	$ZCR = \frac{1}{N-1}\sum_{i=1}^{N-1}\lvert sign(x_i) - sign(x_{i+1})\rvert$	[40], [42]
	Interquartile Range	IQR	$IQR = Q_3 - Q_1$	[42], [43]
	75 'th percentile	PE	$PE = Q_3$	[41], [42]
	Kurtosis	KS	$KS = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2)^2}$	[43], [44]
	Signal magnitude area	SMA	$SMA = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$	[40], [43]
	Min-max	MM	$MM = \max(x) - \min(x)$	[45], [46]
Frequency	Spectral energy	SE	$SE = \sum_{i=1}^{N}\lvert X_i\rvert^2$	[40], [41]
	Spectral entropy	E	$E = -\sum_{i=1}^{N} P_i \log_2(P_i)$	[41], [43]
	Spectral centroid	SC	$SC = \frac{\sum_{i=1}^{N} f_i\lvert X_i\rvert^2}{\sum_{i=1}^{N}\lvert X_i\rvert^2}$	[40], [47]
	Principal frequency	PF	$PF = f_i$ with the maximum $\lvert X_i\rvert$	[41], [48]
Other	Correlation between axis	CORR	$CORR = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}$	[40], [45]
	Autoregressive coefficients	AR1, AR2	$AR1, AR2$ = coefficients from autoregressive model	[46], [49]
	Tilt Angle	TA	$TA = \arctan\frac{\sqrt{x^2+y^2}}{z}$	[47], [49]

TABLE I
COMMON FEATURES IN HAR

models leverage both spatial and temporal information present in captured data.

Deep learning-based feature extraction methods have demonstrated state-of-the-art performance in HAR tasks, particularly when large amounts of labeled data are available for training.

In complex multimodal human activity recognition, features are extracted from various input data sources such as wearable sensors, video streams, and audio recordings. Key features extracted from different modalities include:

1) Wearable Sensor Data: Features extracted from wearable sensor data include mean, variance, mean absolute deviation, root mean square, zero crossing rate, and interquartile range. These features provide insights into motion and orientation information captured by accelerometers and gyroscopes.

2) Video Data: Features extracted from video data capture spatiotemporal information, including motion and spatial distribution of objects. Common features include spatiotemporal features, histogram of oriented gradients (HOG), and optical flow.

3) Audio Data: Features extracted from audio data include mel-frequency cepstral coefficients (MFCCs), zero crossing rate, and spectral centroid. These features represent the short-term power spectrum and spectral characteristics of audio signals associated with human activities.

In multimodal human activity recognition, features from different modalities are often combined to improve activity recognition accuracy.

VII. DATASETS

In recent years, the field of human activity recognition (HAR) has witnessed significant advancements, driven by the integration of multimodal data and the development of sophisticated machine learning algorithms. The availability of large-scale datasets plays a crucial role in fostering research progress and enabling the development and evaluation of robust HAR systems.

This section focuses on multimodal complex HAR datasets, which encompass multiple sensor modalities and capture a wide range of human activities in diverse real-world scenarios. These datasets are designed to address the challenges of recognizing complex activities that involve intricate temporal and spatial dynamics, as well as variations in the different modalities data. Table III summarizes the available multimodal and complex activities datasets. Table II shows the state of the art results of these datasets.

A. PAMAP2

The Activity Monitoring dataset includes 18 distinct physical activities performed by 9 individuals, consisting of 8 men and 1 woman. It was recorded using 3 inertial measurement units and a heart rate monitor. The dataset features a mix of simple and complex activities, making it commonly used in multimodal complex human activity recognition.

B. Opportunity

The dataset consists of data from motion sensors during common daily activities, gathered through body-mounted sensors such as 7 inertial measurement units, 12 3D accelerometers, and 4 3D localization sensors. It also includes data from object sensors (12 objects with 3D acceleration and 2D rate of turn) and ambient sensors (13 switches and 8 3D accelerometers). Four individuals each carried out 6 sessions, which included 5 sessions of regular activities and 1 supervised ”drill” session.

C. Opportunity++

The Opportunity++ dataset enhances the existing OPPORTUNITY dataset by adding previously unpublished videos and video-based skeleton tracking. This improvement is significant as it fills a gap by incorporating video footage and skeleton tracking, thereby promoting more comprehensive research in multimodal sensor fusion.

D. CSL-SHARE

The CSL-SHARE dataset is an in-house sensor-based collection that includes 22 different daily activities and sports, involving 20 individuals for a total of 691 minutes, with 363 minutes being segmented and labeled. The dataset includes two triaxial accelerometers, two triaxial gyroscopes, four surface electromyography (sEMG) sensors, one biaxial electrogoniometer, and an airborne microphone integrated into a knee bandage. Data was recorded at varying sampling rates and resolutions, with synchronization maintained across multiple recording systems.

E. KU-HAR

The dataset consists of data on 18 distinct activities performed by 90 individuals, with 75 of them being male and 15 being female. The data was gathered using smartphone sensors, specifically Accelerometer and Gyroscope, and includes 1945 raw activity samples from the participants. The activities vary from basic stationary tasks like standing and sitting to more active ones such as walking, running, and playing table tennis.

F. SPHERE

Combines different sensor modality which are (accelerometer, video, environmental sensors). The provided data includes accelerometer readings sampled at 20 Hz in raw format, RGB-D video features related to the center of mass and bounding box of identified persons (instead of raw video to maintain participant anonymity), and raw data from Passive Infra-Red (PIR) sensors for environmental context. The dataset is annotated with twenty activity labels related to posture and ambulation.

G. OPERAnet

It includes RF data such as Channel State Information (CSI) from a WiFi Network Interface Card (NIC), Passive WiFi Radar (PWR) from a Software Defined Radio (SDR) platform, and Ultra-Wideband (UWB) signals. Vision and infrared-based data from Kinect sensors are also incorporated. With approximately 8 hours of annotated measurements from six participants engaging in six daily activities across two rooms.

TABLE II
SUMMARY OF RECENT RESULTS OF MULTIMODAL COMPLEX HAR DATASETS

Dataset	Method	Results
PAMAP2	Bhattacharya et al. [50] Ensemble HAR Sarkar et al. [51] Spatial Attention-CNN Verma et al. [52] Multi-branch CNN GRU Kumar et al. [53] Deep-HAR	Accuracy: 97.45% Accuracy: 98.29% Accuracy: 98.65% Accuracy: 99.64%
Opportunity	Chai et al. [54] InnoHAR Zeng et al. [55] RMFSN Han et al. [56] ResNet+HC Mim et al. [57] GRU-INC Park et al. [58] GTSNet	F1 Score: 94.6% Accuracy: 93.89% Accuracy: 91.55% Accuracy: 90.37% Accuracy: 87.47%
CSL-SHARE	Sakorn et al. [59] Deep Residual NN Hartmann et al. [60] HLF	Accuracy: 91.60%, F1 Score: 92.13% Accuracy: 89.7%
KU-HAR	Akter et al. [61] CNN with Attention Teng et al. [62] Large Receptive Field Attention	Accuracy: 96.86% Accuracy: 91.15%
SPHERE	Ali et al. [63] Deep Learning	CNN-LSTM Accuracy: 93.67%, CNN Accuracy: 93.55%, LSTM Accuracy: 92.98%
OPERAnet	Koupai et al. [64] Fusion Transformer	F1 Score: 95.9%
MMAct	Islam et al. [65] Multi-GAT Liu et al. [66] SAKDN Duhme et al. [67] Fusion-GCN	F1 Score: 91.48% F1 Score: 77.23% F1 Score: 89.60%
UTD-MHAD	Mengyuan et al. [68] CNN with Pose Estimation Islam et al. [69] HAMLET Yang et al. [70] LFMF Liu et al. [66] SAKDN	Accuracy: 94.51% Accuracy: 95.12% Accuracy: 98.20% Accuracy:98.60%
C-MHAD	Wei et al. [71] 3D CNN and 2D CNN	Smart TV Actions F1 Score: 81.8%, Transition Actions F1 Score: 82.3%
WEAR	Bock et al. DeepConvLSTM [72]	F1 Score: 75.78%
CMU-MMAC	Lu et al. [69] RecCapsNet with LSTM Yantao et al. [73] LSTM	Accuracy: 86.6% Recall and precision rates of 85.8% and 86.2%, respectively.
UCSD-MIT	Islam et al. [69] HAMLET Long et al. [74] KEYLESS	F1 Score: 81.52% F1 Score: 74.40%
UT-Kinect	Islam et al. [69] HAMLET Mobaraki et al. [75] LSTM	Accuracy: 97.45% Accuracy: 84.14%
MEX	Martin et al. [76] MLP and CNN	MLP Accuracy: 94.19%, CNN Accuracy: 95.49%
UP-Fall	Mart´ınez et al. [77] MLP and RF Murat et al. [78] ShuffleNet Li et al. [79] Kamtfenet	MLP Accuracy: 94.32%, RF Accuracy: 95.19% Accuracy: 99.7% Accuracy: 99.62%
Berkeley-MHAD	Timothy et al. [80] SVM Ba et al. [81] Logits Liu et al. [66] SAKDN	Accuracy: 97.6% Accuracy: 97.93% Accuracy: 99.33%
MHEALTH	Kutlay et al. [82] Various ML Algorithms	Accuracies: 90.55%-78.09%
REALDISP	Runze et al. [83] SMLDist Aljarrah et al. [84] K-NN, Decision tree, Naive Bayes, SVM, CNN	Accuracy: 94% Accuracies: 98.09%, 94.61%, 83.61%, 93.12%, 99.85%
UCI HAR	Zeng et al. [55] RMFSN Venkatachalam et al. [85] Bi-HAR Han et al. [56] ResNet+HC Mim et al. [57] GRU-INC Wang et al. [86] DMEFAM	Accuracy: 98.13% Accuracy: 97.89% Accuracy: 97.01% Accuracy: 96.27% Accuracy: 96.00%
WISDM	Zeng et al. [55] RMFSN Mim et al. [57] GRU-INC Wang et al. [86] DMEFAM Sun et al. [87] CapsGaNet Park et al. [58] GTSNet	Accuracy: 98.35% Accuracy: 99.13% Accuracy: 97.90% Accuracy: 96.80% Accuracy: 88.87%

H. EV-Action

The EV-Action dataset is a collection of data designed for multi-modal human action analysis, featuring RGB, depth, electromyography (EMG), and two skeleton modalities. It surpasses previous datasets by providing high-quality skeleton data captured using a motion capturing system, along with introducing EMG data, which is a novel addition to the field of motion-related research.

I. MMAct

This dataset has been recorded from 20 distinct subjects with seven different types of modalities: RGB videos, key-points, acceleration, gyroscope, orientation, Wi-Fi and pressure signal. The dataset consists of more than 36k video clips for 37 action classes covering a wide range of daily life activities such as desktop-related and check-in-based ones in four different distinct scenarios.

J. UTD-MHAD

The UTD-MHAD dataset includes 27 separate activities done by 8 participants. Each subject did the action four times, totaling 861 action sequences. The RGB, depth, skeletal, and inertial sensor signals were captured.

K. C-MHAD

It includes two types of actions: smart TV gestures and transition movements. The dataset contains 3-axis acceleration and angular velocity signals acquired by the Shimmer3 wearable inertial sensor at 50Hz via Bluetooth, synchronized with video frames captured at 15 frames per second using a laptop camera with a resolution of 640x480 pixels. Each action stream lasts two minutes, which corresponds to 1801 picture frames and 6001 inertial signal samples. Due to Bluetooth communication latency, 30-40 samples at the start of each action stream are missing and may require no padding before usage. The sensor is worn on the right wrist for smart TV gestures, and on the waist for transition movements.

L. WEAR

WEAR is an outdoor sports dataset designed for human activity recognition (HAR) using both visual and inertial sensors. The dataset includes data from 18 subjects who participated in a variety of fitness activities, as well as untrimmed inertial (acceleration) and camera (egocentric video) data collected at ten distinct outdoor locales.

M. LboroHAR

Contains nine indoor activities carried out by 16 individuals. The dataset collects data from a variety of sensors typically used in indoor applications and autonomous cars, including the depth sensor, RGB color image (RGB-D), LiDAR sensor, and RGB 360 camera. It is the first publically accessible multimodal dataset of its sort and may be used for a variety of HAR applications, including as sports analytics, healthcare aid, and indoor smart mobility.

N. CMU-MMAC

A collection of multimodal measures capturing human activity during cooking and food preparation tasks. Recorded in Carnegie Mellon's Motion Capture Lab, the database includes recordings of twenty-five subjects performing five different recipes: brownies, pizza, sandwich, salad, and scrambled eggs. The modalities recorded encompass video, audio, motion capture, internal measurement units (IMUs), and wearable devices.

O. UT-Kinect

Consists of videos collected with a single stationary Kinect with Kinect for Windows SDK Beta Version. It has ten action types: walk, sit, stand, pick up, carry, toss, push, pull, wave hands, and clap hands. There are ten subjects, and each performs each action twice. The recordings have three synchronized channels: RGB, depth, and skeleton joint positions, with a framerate of 30 frames per second (FPS).

P. MEx

Data from seven distinct physiotherapy exercises, each completed by thirty volunteers, are included in the MEx Multi-modal Exercise dataset. The activities were recorded using four different sensor modalities: two accelerometers, a pressure mat, and a depth camera. Exercise recognition, exercise quality evaluation, and exercise counting are among the activities for which the dataset is appropriate.

Q. UP-Fall detection

The dataset consists of raw and feature sets collected from 17 young, healthy people who completed 11 activities and three falls each. Furthermore, the collection compiles more than 850 GB of data from vision devices, ambient sensors, and wearable sensors.

R. HHAR

This dataset contains measurements from two motion sensors—the accelerometer and the gyroscope—that were taken as users used smartphones and smartwatches to carry out predetermined tasks. The dataset includes recordings of nine individuals' activities, including walking, biking, sitting, standing, climbing stairs, and descending them, made using a total of four smartwatches and eight smartphones. Furthermore, a portion of the dataset comprises accelerometer measurements of devices in six distinct orientations when they are immobile. The devices comprise 31 smartphones, 4 smartwatches, and 1 tablet, covering 13 models and 4 brands, and operating on various versions of Android and iOS.

S. Berkeley MHAD

Data from 12 subjects, ages 23 to 30 (seven male and five female), plus one old subject, are included. Every participant executed five iterations of eleven distinct activities, resulting in roughly 660 action sequences with an approximate recording duration of eighty-two minutes. It includes data from multiple sensors, such as depth sensors, accelerometers, microphones, multi-baseline stereo cameras, optical motion capture systems, and others, that has been synchronized and calibrated.

T. MHEALTH

It consists of recordings of the vital signs and body movements of ten volunteers engaging in twelve different physical activities. The dataset contains information from sensors that measure acceleration, rate of turn, and magnetic field direction. The sensors were positioned on the chest, right wrist, and left ankle. The chest sensor also offers 2-lead ECG readings for potential heart monitoring. The dataset, which was recorded at a sampling rate of 50 Hz, documents a variety of everyday activities in a non-laboratory setting.

U. REALDISP

It is comprised of data gathered with an emphasis on ideal-placement, self-placement, and induced-displacement situations in order to study the impacts of sensor displacement in practical settings. The dataset comprises 17 persons, a variety of physical activities, and sensor modalities (acceleration, rate of rotation, and magnetic field).

V. UCI HAR

30 volunteers' recordings of themselves engaging in six daily tasks while wearing a smartphone strapped on their waist that has an accelerometer and gyroscope integrated are included.

W. UCI HAPT

Recordings from thirty people who carried a smartphone strapped on their waist with inertial sensors and performed twelve distinct fundamental activities and postural transitions.

X. WISDM

Includes time-series accelerometer and gyroscope sensor data from 51 test subjects who completed 18 tasks for three minutes each using smartphones and smartwatches.

VIII. Algorithms Used in Multimodal Complex HAR

Human Activity Recognition (HAR) in multimodal complex scenarios often employs various algorithms for classification. These algorithms can be broadly categorized into machine learning, deep learning, and transformer-based approaches.

A. Traditional Machine Learning

Traditional machine learning algorithms have been widely used in HAR, particularly in early studies where deep learning techniques were not as prevalent. These algorithms often rely on handcrafted features extracted from captured data and perform classification using various classifiers. Some commonly used traditional machine learning algorithms in multimodal HAR include:

- Support Vector Machines (SVM) [108] [109]
- k-Nearest Neighbors (k-NN) [110]
- Decision Trees [111] [112]
- Naive Bayes [113]
- Gradient Boosting Machines [112]

Traditional machine learning algorithms require extensive feature engineering, where domain knowledge is used to extract relevant features from raw data. These handcrafted features may include statistical measures, time-domain features, frequency-domain features, and other engineered representations of the captured data. While traditional machine learning approaches have shown success in HAR tasks, they may struggle to capture complex temporal and spatial dependencies present in multimodal data, and their performance heavily depends on the quality of handcrafted features.

In recent years, with the rise of deep learning techniques, traditional machine learning algorithms have been somewhat overshadowed by deep learning models, which can extract features directly from raw data. However, traditional machine learning algorithms still find applications in multimodal HAR, particularly in scenarios where interpretability and computational efficiency are crucial factors.

B. Deep Learning and Transformers

Deep learning techniques have revolutionized the field of HAR by enabling the automatic learning of features directly from raw data. These techniques have shown remarkable performance in multimodal HAR tasks, surpassing traditional machine learning approaches in many cases. Some commonly used deep learning architectures in multimodal HAR include:

- Convolutional Neural Networks (CNNs) [114] [115]: CNNs are widely used for processing spatial information in captured data, particularly in scenarios involving image-based modalities or sensor data. They consist of convolutional layers followed by pooling layers, enabling them to automatically extract spatial features from input data.
- Recurrent Neural Networks (RNNs) [115] and variants (e.g., Long Short-Term Memory networks, LSTM [114]; Gated Recurrent Units, GRU) [116]: RNNs are well-suited for modeling temporal dependencies in sequential data, making them effective for time-series analysis in HAR. Variants like LSTM and GRU address the vanishing gradient problem in traditional RNNs, enabling them to capture long-range dependencies in temporal sequences.
- Convolutional Recurrent Neural Networks: combine the strengths of CNNs and RNNs by integrating convolutional layers for spatial feature extraction and recurrent layers for capturing temporal dependencies [114] [117] [118]. They are particularly useful for processing multimodal data streams where both spatial and temporal information is essential.
- Transformers: Transformers [119] have emerged as powerful models for sequence transduction tasks, including multimodal human activity recognition (HAR). Unlike traditional recurrent or convolutional architectures, transformers rely on self-attention mechanisms to capture global dependencies across input sequences, making them well-suited for processing multimodal data streams. In the context of multimodal HAR, transformers can effectively

TABLE III
SUMMARY OF ACTIVITY RECOGNITION DATASETS

Dataset	Sensors	Activities
pamap2 [88]	3 (IMUs) and heart rate	18 simple and complex activities
opportunity [89]	Body worn IMUs, ambient sensors, and object sensors	35 activities
opportunity++ [90]	Body worn IMUs, ambient sensors, object sensors, and videos	35 activities
CSL-SHARE [91]	Accelerometers, gyroscopes, EMGs, microphone, and Airborne	22 simple and transition daily living and sports-based activities
ku har [92]	Smartphone accelerometer and gyroscope	18 daily activities
SPHERE [93]	Accelerometers, RGB-D, and environmental sensors	20 daily activities
OPERAnet [94]	WiFi sensing device, ultra-wideband impulse radar, passive WiFi radar, and Kinect motion sensor	6 daily activities
EV-Action [11]	RGB, RGB-D, EMG, and skeleton	10 single actions and 10 object actions
MMAct [95]	RGB, Keypoints, Acceleration, Gyroscope, Orientation, Wi-Fi, and Pressure	37 activities
UTD-MHAD [96]	Kinect camera and wearable IMU	27 activities
C-MHAD [71]	Accelerometer, Gyroscope, and Video	12 actions
WEAR dataset [97]	IMUs and cameras	18 activities
LboroHAR [72]	LiDAR Sensor, RGB, and RGB-D	9 activities
CMU Multi-Modal Activity (CMU-MMAC) [95]	Videos, Audio, Motion capture, IMUs, and wearable watch	5 cooking activities
UCSD-MIT Human Motion [98]	Skeleton, EMGs, and IMUs	11 activities
UT-Kinect [99]	RGB, RGB-D, and skeleton	10 activities
MEx [100]	Accelerometers, pressure, and depth camera	7 activities
HHAR [101]	Accelerometers and gyroscopes from smartphones and smartwatches	6 activities
UP-Fall detection [15]	IMUs, EEG, and infrared	11 activities
Berkeley-MHAD [102]	RGB, RGB-D, accelerometers, and microphones	11 activities
MHEALTH Dataset [103]	ECG, accelerometer, gyroscope, and magnetometer	12 activities
REALDISP Activity Recognition Dataset [104]	Accelerometer, gyroscope, and magnetometer	33 activities
UCI HAR [105]	Accelerometer and gyroscope	6 activities
UCI HAPT [106]	Accelerometer and gyroscope	12 simple and transition activities
WISDM [107]	Accelerometer and gyroscope	18 activities

integrate information from multiple modalities [120] by attending to relevant parts of the input sequences. Transformers have demonstrated state-of-the-art performance in various multimodal HAR benchmarks [121], surpassing traditional deep learning architectures in many cases. Their ability to capture long-range dependencies and effectively integrate information from disparate modalities makes them particularly well-suited for complex real-world scenarios where understanding the relationships between different sensor inputs is crucial for accurate activity recognition.

Deep learning models in multimodal HAR have the advantage of automatically learning relevant features directly from raw data, eliminating the need for handcrafted feature engineering. These models can effectively capture complex temporal and spatial dependencies present in multimodal data, leading to improved classification accuracy and robustness. However, deep learning models often require large amounts of annotated data for training and may suffer from overfitting in scenarios with limited data availability. Regularization techniques and data augmentation strategies are commonly employed to mitigate these challenges and improve model generalization.

IX. EVALUATION METRICS

To evaluate the multimodal HAR system's performance, several metrics can be used. Let N denote the total number of activity instances in the dataset. Additionally, let TP , TN , FP , and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively.

A. Accuracy

The ratio of accurately identified examples to the total number of occurrences is known as accuracy, and it indicates how accurate the HAR system is:

$$\text{Accuracy} = \frac{TP + TN}{N} \tag{1}$$

B. Precision

The percentage of successfully anticipated positive instances among all positively predicted instances is quantified by precision. It is computed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

C. Recall

Recall, which is a statistical measure of the percentage of accurately predicted positive cases among all actual positive

instances, is also referred to as sensitivity or true positive rate.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

D. F1 Score

The F1 score strikes a balance between recall and precision by combining the two into a single metric. The harmonic mean of recall and precision is used to calculate it:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

E. Confusion Matrix

A tabular representation of the classification results that offers a thorough analysis of the classification performance is called a confusion matrix. For each class, it displays the quantity of cases that were successfully and wrongly classified. Evaluation measures including accuracy, precision, and recall can be calculated using the confusion matrix.

It is important to select the appropriate evaluation metrics based on the specific objectives and characteristics of the multimodal HAR system.

X. Applications

Applications of multimodal complex activity recognition span a wide range of domains, benefiting various fields and industries.

Multimodal complex activity recognition can be applied in healthcare settings to monitor patients' activities and provide personalized care. It can aid in fall detection, activity tracking, and assessing overall wellbeing. By combining data from wearable sensors, video surveillance, and audio analysis, healthcare professionals can gain insights into patients' daily activities and identify any anomalies or potential health risks. Multimodal complex activity recognition plays a crucial role in enabling effective and natural interactions between humans and robots. By integrating multiple modalities such as vision, speech, and gesture recognition, robots can better understand human intentions and adapt their behavior accordingly. This technology finds applications in service robots, social robotics, and collaborative robotics, enhancing human-robot communication and cooperation. Moreover, it contributes to the development of smart environments that can understand and respond to human activities. In smart homes, for instance, combining data from sensors, cameras, and audio analysis allows for context-aware automation, energy management, and personalized assistance. These systems can recognize activities like cooking, watching TV, or sleeping, and adjust lighting, temperature, and other environmental factors accordingly.

Multimodal complex activity recognition is valuable in surveillance and security applications. By integrating video analysis, audio recognition, and sensor data, it becomes possible to detect and classify suspicious or abnormal activities in public spaces, airports, or critical infrastructure. This technology enhances situational awareness, facilitates early threat detection, and aids in crime prevention. Additionally, multimodal complex activity recognition has applications in sports training and fitness monitoring. By analyzing data from wearable devices, video analysis, and audio feedback, it becomes possible to track and assess athletes' performance, provide real-time coaching, and offer personalized training programs. This technology can be used in sports analytics, rehabilitation, and general fitness tracking.

XI. Challenges and Future Research

Despite the advancements in multimodal complex HAR, there are several challenges that need to be addressed for further improvements. This section discusses some of these challenges and suggests potential directions for future research.

A. Data Collection and Annotation

Collecting and annotating large-scale multimodal complex HAR datasets is a challenging task. The data collection process involves capturing synchronized data from multiple sensors, such as cameras, inertial sensors, and microphones. The annotation process requires expert knowledge and manual labeling of complex activities, which can be time-consuming and prone to subjective biases. Future research should focus on developing efficient data collection techniques and exploring semi-automatic or automatic annotation methods to facilitate the creation of comprehensive multimodal complex HAR datasets.

B. Sensor Fusion

Effective fusion of information from multiple modalities is crucial for accurate complex activity recognition. Developing robust fusion techniques that can effectively combine information from different sensors is an ongoing research challenge. Additionally, learning effective representations from multimodal data is essential for capturing the inherent complexity and temporal dependencies in complex activities. Future research should explore innovative sensor fusion architectures and representation learning algorithms to improve the performance of multimodal complex HAR systems.

C. Model Interpretability

Multimodal complex HAR systems often utilize deep learning models, which are known for their black-box nature. Interpreting and explaining the decision-making process of these models is a significant challenge. Understanding the rationale behind the predictions can enhance the trust and usability of the system, especially in critical applications such as healthcare and surveillance. Future research should focus on developing interpretable and explainable models for multimodal complex HAR, enabling users to understand how and why certain activity predictions are made.

D. Real-Time Processing and Resource Constraints

Real-time processing of multimodal complex HAR is crucial for applications that require immediate responses, such as human-robot interaction and real-time monitoring. However, multimodal processing can be computationally intensive,

making it challenging to achieve real-time performance, especially in resource-constrained environments. Future research should investigate efficient algorithms, hardware acceleration techniques, and optimization strategies to enable real-time processing of multimodal complex HAR systems on devices with limited computational resources.

E. Domain Adaptation and Generalization

Multimodal complex HAR models trained on one dataset may not generalize well to different domains or unseen activities. Adapting the models to new environments or activities without extensive retraining is a challenge. Future research should focus on developing domain adaptation techniques and transfer learning approaches to improve the generalization capability of multimodal complex HAR models, enabling them to perform well in diverse real-world settings.

F. Privacy and Ethical Considerations

Multimodal complex HAR involves the collection and processing of sensitive personal data. Ensuring privacy, data protection, and ethical considerations are important challenges to address. Future research should focus on developing privacy-preserving techniques, robust anonymization methods, and ethical guidelines for the collection, storage, and usage of multimodal complex HAR data.

Addressing these challenges and exploring the suggested directions for future research will contribute to the advancement of multimodal complex HAR and its applications in various domains, including healthcare, smart environments, and human-computer interaction.

XII. Conclusion

The review article discusses various types of sensors utilized in Human Activity Recognition (HAR), such as visual sensors, wearable inertial sensors, and their combinations. These sensors are cost-effective and readily accessible. However, integrating data from different sensor modalities poses a significant challenge for accurately classifying human activities.

It's important to highlight that research in multimodal HAR is continuously progressing, with scholars exploring innovative algorithms, architectures, and methodologies to enhance the precision, resilience, and real-time capabilities of activity recognition systems.

In summary, multimodal complex HAR demonstrates considerable promise across numerous fields, including healthcare, smart environments, and human-computer interaction. By harnessing the complementary aspects of various modalities, these systems can offer valuable insights and support diverse applications aimed at enhancing human life and interaction with technology.

References

[1] L. M. Danga, K. Mina, H. Wanga, M. J. Pirana, C. H. L. b, and H. Moona, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Elsevier*, 2020.

[2] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[3] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[4] L. Chen, X. Liu, L. Peng, and M. Wu, "Deep learning based multimodal complex human activity recognition using wearable devices," *Applied Intelligence*, p. 4029–4042, jun 2021.

[5] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Computing Surveys (Csur)*, vol. 43, no. 3, pp. 1–43, 2011.

[6] M. O. Gani, *A novel approach to complex human activity recognition*. PhD thesis, Marquette University, 2017.

[7] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook, "Simple and complex activity recognition through smart phones," in *2012 eighth international conference on intelligent environments*, pp. 214–221, IEEE, 2012.

[8] Mueller, Marshall, Khot, S. Nylander, and J. Tholander, ""understanding sports-hci by going jogging at chi," *ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015.

[9] S. Mekruksavanich and A. Jitpattanakul, "Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models," *Electronics*, vol. 10, p. 308, 01 2021.

[10] A. Zahin, T. Le, and R. Hu, *Sensor-Based Human Activity Recognition for Smart Healthcare: A Semi-supervised Machine Learning*, pp. 450–472. 07 2019.

[11] L. Wang, B. Sun, J. Robinson, T. Jing, and Y. Fu, "Ev-action: Electromyography-vision multi-modal action dataset," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 160–167, 2020.

[12] H. Mousavi Hondori, M. Khademi, and C. Lopes, "Monitoring intake gestures using sensor fusion (microsoft kinect and inertial sensors) for smart home tele-rehab setting," 11 2012.

[13] B. Kwolek and M. Kepski, "Fuzzy inference-based fall detection using kinect and body-worn accelerometer," *Applied Soft Computing*, vol. 40, p. 305–318, 03 2016.

[14] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. Satiza´bal Mejia, "Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors," in *Advances in Computational Intelligence* (I. Rojas, G. Joya, and J. Cabestany, eds.), (Berlin, Heidelberg), pp. 216–223, Springer Berlin Heidelberg, 2013.

[15] L. Martinez-Villasen˜or, H. Ponce, J. Brieva, E. Moya-Albor, J. Nun˜ez Martinez, and C. Pen˜afort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, p. 1988, 04 2019.

[16] D. Rav`ı, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. PP, 12 2016.

[17] T. v. Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and imus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1533–1547, 2016.

[18] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3d human pose estimation fusing video and inertial sensors," 09 2017.

[19] D. H. Wilson and C. Atkeson, "Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors," in *Pervasive Computing* (H. W. Gellersen, R. Want, and A. Schmidt, eds.), (Berlin, Heidelberg), pp. 62–79, Springer Berlin Heidelberg, 2005.

[20] L. Gao, A. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," *Medical Engineering & Physics*, vol. 36, no. 6, pp. 779–785, 2014.

[21] M. Yamin and G. Chetty, "Intelligent human activity recognition scheme for ehealth applications," *Malaysian Journal of Computer Science*, vol. 28, 03 2015.

[22] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[23] J. Morales and D. Akopian, "Physical activity recognition by smartphones, a survey," *Biocybernetics and Biomedical Engineering*, vol. 37, no. 3, pp. 388–400, 2017.

[24] J. Park, S. Park, M. Zia Uddin, M. Al-Antari, M. Al-Masni, and T.-S. Kim, "A single depth sensor based human activity recognition via convolutional neural network," in *6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6) 6*, pp. 541–545, Springer, 2018.

[25] A. M. Khan, A. Tufail, A. M. Khattak, and T. H. Laine, "Activity recognition on smartphones via sensor-fusion and kda-based svms," *International Journal of Distributed Sensor Networks*, vol. 10, no. 5, p. 503291, 2014.

[26] M. Z. Uddin, W. Khaksar, and J. Torresen, "A thermal camera-based activity recognition using discriminant skeleton features and rnn," in *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, vol. 1, pp. 777–782, IEEE, 2019.

[27] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall, "Recognizing daily activities with rfid-based sensors," in *Proceedings of the 11th international conference on Ubiquitous computing*, pp. 51–60, 2009.

[28] C. Hou, "A study on imu-based human activity recognition using deep learning and traditional machine learning," in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pp. 225–234, IEEE, 2020.

[29] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE access*, vol. 8, pp. 210816–210836, 2020.

[30] H. Dvir, *Illumination correction for content analysis in uterine cervix images*. 2007.

[31] Y. Wu, Z. Liu, Y. Han, and H. Zhang, "An image illumination correction algorithm based on tone mapping," in *2010 3rd International Congress on Image and Signal Processing*, vol. 2, pp. 645–648, 2010.

[32] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognition*, vol. 76, pp. 635–649, 2018.

[33] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, "A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors," *Sensors*, vol. 19, no. 22, p. 5026, 2019.

[34] A. H. Niazi, D. Yazdansepas, J. L. Gay, F. W. Maier, L. Ramaswamy, K. Rasheed, and M. P. Buman, "Statistical analysis of window sizes and sampling rates in human activity recognition.," in *HEALTHINF*, pp. 319–325, 2017.

[35] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[36] N. Neverova, C. Wolf, G. Lacey, L. Fridman, D. Chandra, B. Barbello, and G. Taylor, "Learning human identity from motion patterns," *IEEE Access*, vol. 4, 11 2015.

[37] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012.

[38] S. Mu¨nzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Du¨richen, "Cnn-based sensor fusion techniques for multimodal human activity recognition," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ISWC '17, (New York, NY, USA), p. 158–165, Association for Computing Machinery, 2017.

[39] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[40] L. Gao, A. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," *Medical Engineering & Physics*, vol. 36, no. 6, pp. 779–785, 2014.

[41] J. Preece, J. Goulermas, L. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 871–879, 2009.

[42] U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *International Workshop on Wearable and Implantable Body Sensor Networks*, (Cambridge, USA), pp. 112–116, 2006.

[43] D. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[44] A. Godfrey, R. Conway, D. Meagher, and G. Olaighin, "Direct measurement of human movement by accelerometry," *Medical Engineering & Physics*, vol. 30, no. 10, pp. 1364–1386, 2008.

[45] A. Bayat, M. Pomplun, and D. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450–457, 2014.

[46] J. Kavanagh and B. Menz, "Accelerometry: a technique for quantifying movement patterns during walking," *Gait & posture*, vol. 28, no. 1, pp. 1–15, 2008.

[47] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.

[48] J. Suto, S. Oniga, and A. Buchman, "Real time human activity monitoring," *Annales Mathematicae et Informaticae*, vol. 44, no. 1, pp. 187–196, 2015.

[49] A. Khan, Y. Lee, S. Lee, and T. Kim, "A triaxial accelerometer-based physical activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1166–1172, 2010.

[50] D. Bhattacharya, D. Sharma, W. Kim, M. F. Ijaz, and P. K. Singh, "Ensem-har: An ensemble deep learning model for smartphone sensor-based human activity recognition for measurement of elderly health monitoring," *Biosensors*, vol. 12, no. 6, 2022.

[51] A. Sarkar, S. S. Hossain, and R. Sarkar, "Human activity recognition from sensor data using spatial attention-aided cnn with genetic algorithm," *Neural Computing and Applications*, vol. 35, no. 7, pp. 5165–5191, 2023.

[52] U. Verma, P. Tyagi, and M. K. Aneja, "Multi-branch cnn gru with attention mechanism for human action recognition," *Engineering Research Express*, 2023.

[53] P. Kumar and S. Suresh, "Deep-HAR: an ensemble deep learning model for recognizing the simple, complex, and heterogeneous human activities," *Multimedia Tools and Applications*, vol. 82, pp. 30435–30462, 2023.

[54] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "Innohar: A deep neural network for complex human activity recognition," *IEEE Access*.

[55] F. Zeng, M. Guo, L. Tan, F. Guo, and X. Liu, "Wearable sensor-based residual multifeature fusion shrinkage networks for human activity recognition," *Sensors*, vol. 24, no. 3, 2024.

[56] C. Han, L. Zhang, Y. Tang, W. Huang, F. Min, and J. He, "Human activity recognition using wearable sensors by heterogeneous convolutional neural networks," *Expert Systems with Applications*, vol. 198, p. 116764, 2022.

[57] T. R. Mim, M. Amatullah, S. Afreen, M. A. Yousuf, S. Uddin, S. A. Alyami, K. F. Hasan, and M. A. Moni, "Gru-inc: An inception-attention based approach using gru for human activity recognition," *Expert Systems with Applications*, vol. 216, p. 119419, 2023.

[58] J. Park, W.-S. Lim, D.-W. Kim, and J. Lee, "Gtsnet: Flexible architecture under budget constraint for real-time human activity recognition from wearable sensor," *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106543, 2023.

[59] S. Mekruksavanich, P. Jantawong, and A. Jitpattanakul, "A deep learning-based model for human activity recognition using biosensors embedded into a smart knee bandage," *Procedia Computer Science*, vol. 214, pp. 621–627, 2022. 9th International Conference on Information Technology and Quantitative Management.

[60] Y. Hartmann, H. Liu, and T. Schultz, "High-level features for human activity recognition and modeling," in *International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 141–163, Springer, 2022.

[61] M. Akter, S. Ansary, M. A.-M. Khan, and D. Kim, "Human activity recognition using attention-mechanism-based deep learning feature combination," *Sensors*, vol. 23, no. 12, p. 5715, 2023.

[62] Q. Teng, Y. Tang, and G. Hu, "Large receptive field attention: An innovation in decomposing large-kernel convolution for sensor-based activity recognition," *IEEE Sensors Journal*, vol. 24, no. 8, pp. 13488–13499, 2024.

[63] A. A. Alani, G. Cosma, and A. Taherkhani, "Classifying imbalanced multi-modal sensor data for human activity recognition in a smart home using deep learning," 10 2020.

[64] A. K. Koupai, M. J. Bocus, R. Santos-Rodriguez, R. J. Piechocki, and R. McConville, "Self-supervised multimodal fusion transformer for passive activity recognition," *IET Wireless Sensor Systems*, vol. 12, no. 5-6, pp. 149–160, 2022.

[65] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1729–1736, 2021.

[66] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 5573–5588, 2021.

[67] M. Duhme, R. Memmesheimer, and D. Paulus, "Fusion-gcn: Multimodal action recognition using graph convolutional networks," in *Pattern Recognition* (C. Bauckhage, J. Gall, and A. Schwing, eds.), (Cham), pp. 265–281, Springer International Publishing, 2021.

[68] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1159–1168, 2018.

[69] M. M. Islam and T. Iqbal, "Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10285–10292, 2020.

[70] Z. Yang, Z. Gui, H. Wu, and W. Li, "A latent feature-based multi-modality fusion method for theme classification on web map service," *IEEE Access*, vol. 8, pp. 25299–25309, 12 2019.

[71] H. Wei, P. Chopada, and N. Kehtarnavaz, "C-MHAD: Continuous Multimodal Human Action Dataset of Simultaneous Video and Inertial Sensing," *Sensors*, vol. 20, no. 10, p. 2905, 2020.

[72] M. Moencks, V. De Silva, J. Roche, and A. Kondoz, "Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset," *arXiv preprint arXiv:1901.02858*, 2019.

[73] Y. Lu and S. Velipasalar, "Human activity classification incorporating egocentric video and inertial measurement unit data," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 429–433, 2018.

[74] X. Long, C. Gan, G. Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multi-modal keyless attention fusion for video classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[75] M. Mobaraki, A. Bannadabhavi, M. J. Yedlin, and B. Gopaluni, "A vision-based deep learning platform for human motor activity recognition," in *2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pp. 1–4, 2023.

[76] Martin, Kyle, Wijekoon, Anjana, Wiratunga, and Nirmalie, "Human activity recognition with deep metric learners.," CEUR Workshop Proceedings, 2020.

[77] L. Mart´ınez-Villasen˜or, H. Ponce, J. Brieva, E. Moya-Albor, J. Nu´n˜ez-Mart´ınez, and C. Pen˜afort-Asturiano, "UP-Fall Detection Dataset: A Multimodal Approach," *Sensors*, vol. 19, no. 9, p. 1988, 2019.

[78] H. Ergu¨der, T. Uzun, and M. Baday, "Advancing fall detection utilizing skeletal joint image representation and deformable layers," *Image Analysis and Stereology*, vol. 43, no. 1, pp. 97–107, 2024.

[79] J. Li, M. Gao, B. Li, D. Zhou, Y. Zhi, and Y. Zhang, "Kamtfenet: a fall detection algorithm based on keypoint attention module and temporal feature extraction," *International Journal of Machine Learning and Cybernetics*, vol. 14, pp. 1831–1844, 2022.

[80] H. Ng, T. Tzen, T. T. V. Yap, H.-L. Tong, C. C. Ho, L. Tan, W. Eng, S. Yap, and J. Soh, "Action classification on the berkeley multimodal human action dataset (mhad)," 02 2017.

[81] L. Ba and R. Caruana, "Do deep nets really need to be deep?," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2654–2662, 01 2014.

[82] M. A. Kutlay and S. Gagula-Palalic, "Application of machine learning in healthcare: Analysis on mhealth dataset," *Southeast Europe Journal of Soft Computing*, vol. 4, no. 2, 2016.

[83] R. Chen, H. Luo, F. Zhao, X. Meng, Z. Xie, and Y. Zhu, "A light-weight deep human activity recognition algorithm using multi-knowledge distillation," 2023.

[84] A. A. Aljarrah and A. H. Ali, *Human Activity Recognition by Deep Convolution Neural Networks and Principal Component Analysis*, pp. 111–133. Cham: Springer International Publishing, 2021.

[85] K. Venkatachalam, Z. Yang, P. Trojovsky`, N. Bacanin, M. Deveci, and W. Ding, "Bimodal har-an efficient approach to human activity analysis and recognition using bimodal hybrid classifiers," *Information Sciences*, vol. 628, pp. 542–557, 2023.

[86] Y. Wang, H. Xu, Y. Liu, M. Wang, Y. Wang, Y. Yang, S. Zhou, J. Zeng, J. Xu, S. Li, *et al.*, "A novel deep multifeature extraction framework based on attention mechanism using wearable sensor data for human activity recognition," *IEEE Sensors Journal*, vol. 23, no. 7, pp. 7188–7198, 2023.

[87] X. Sun, H. Xu, Z. Dong, L. Shi, Q. Liu, J. Li, T. Li, S. Fan, and Y. Wang, "Capsganet: Deep neural network based on capsule and gru for human activity recognition," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5845–5855, 2022.

[88] A. Reiss, "PAMAP2 Physical Activity Monitoring." UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

[89] Roggen, Daniel, Calatroni, Alberto, Nguyen-Dinh, Long-Van, Chavarriaga, Ricardo, Sagha, and Hesam, "OPPORTUNITY Activity Recognition." UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5M027.

[90] M. Ciliberto, V. Fortes Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition," *Frontiers in Computer Science*, vol. 3, 2021.

[91] H. Liu, Y. Hartmann, and T. Schultz, "Csl-share: A multimodal wearable sensor-based human activity dataset," *Frontiers in Computer Science*, vol. 3, 2021.

[92] A.-A. Nahid, N. Sikder, and I. Rafi, "KU-HAR: An open dataset for human activity recognition." Mendeley Data, 2021.

[93] Tonkin, E.L., Holmes, and Song, "A multi-sensor dataset with annotated activities of daily living recorded in a residential setting," *Sci Data*, vol. 10, 2023.

[94] M. J. Bocus, W. Li, S. Vishwakarma, R. Kou, C. Tang, K. Woodbridge, I. Craddock, R. McConville, R. Santos-Rodriguez, K. Chetty, and R. Piechocki, "OPERAnet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors," *Scientific Data*, vol. 9, no. 474, 2022.

[95] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[96] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015.

[97] M. Bock, H. Kuehne, K. Van Laerhoven, and M. Moeller, "Wear: An outdoor sports for wearable and egocentric activity recognition," *CoRR*, vol. abs/2304.05088, 2023.

[98] A. Kubota, T. Iqbal, J. A. Shah, and L. D. Riek, "Activity recognition in manufacturing: The roles of motion capture and semg+inertial wearables in detecting fine vs. gross motion," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6533–6539, 2019.

[99] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27, IEEE, 2012.

[100] Wijekoon, Anjana, Wiratunga, Nirmalie, Cooper, and Kay, "MEx." UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C59K6T.

[101] Blunck, Henrik, Bhattacharya, Sourav, Prentow, Thor, Kjrgaard, Mikkel, Dey, and Anind, "Heterogeneity Activity Recognition." UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

[102] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, 2013.

[103] Banos, Oresti, Garcia, Rafael, Saez, and Alejandro, "MHEALTH Dataset." UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TW22.

[104] Banos, Oresti, Toth, Mate, and O. Amft, "REALDISP Activity Recognition Dataset." UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5GP6D.

[105] Reyes-Ortiz, Jorge, Anguita, Davide, Ghio, Alessandro, Oneto, Luca, Parra, and Xavier, "Human Activity Recognition Using Smartphones." UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C54S4K.

[106] "Smartphone-based recognition of human activities and postural transitions data set." urlhttp://archive.ics.uci.edu/ml/datasets/smartphone-based+recognition+of+human+activities+and+postural+transitions. Accessed: 2022-12-12.

[107] G. Weiss, "WISDM Smartphone and Smartwatch Activity and Biometrics Dataset ." UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5HK59.

[108] P. Choudhary, P. Pathak, and A. Chaubey, "Activity recognition system via unification of cnn and svm in complex domain," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 1–6, 2023.

[109] Y. Nawal, M. Oussalah, B. Fergani, and et al., "New incremental svm algorithms for human activity recognition in smart homes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 13433–13450, 2023.

[110] S. Uddin, I. Haque, H. Lu, and et al., "Comparative performance analysis of k-nearest neighbour (knn) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, p. 6256, 2022.

[111] A. Wang, H. Chen, C. Zheng, L. Zhao, J. Liu, and L. Wang, "Evaluation of random forest for complex human activity recognition using wearable sensors," in *2020 International Conference on Networking and Network Applications (NaNA)*, pp. 310–315, 2020.

[112] K. Gusain, A. Gupta, and B. Popli, "Transition-aware human activity recognition using extreme gradient boosted decision trees," in *Advanced Computing and Communication Technologies* (R. K. Choudhary, J. K. Mandal, and D. Bhattacharyya, eds.), (Singapore), pp. 41–49, Springer Singapore, 2018.

[113] J. Shen and H. Fang, "Human activity recognition using gaussian na¨ıve bayes algorithm in smart home," *Journal of Physics: Conference Series*, vol. 1631, p. 012059, sep 2020.

[114] A. Sarabu and A. Santra, "Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks," *Emerging Science Journal*, vol. 5, pp. 25–33, 02 2021.

[115] S. Mekruksavanich, P. Jantawong, N. Hnoohom, and A. Jitpattanakul, "Heterogeneous recognition of human activity with cnn and rnn-based networks using smartphone and smartwatch sensors," in *2022 3rd International Conference on Big Data Analytics and Practices (IBDAP)*, pp. 21–26, 2022.

[116] X. Huang, Y. Yuan, C. Chang, Y. Gao, C. Zheng, and L. Yan, "Human activity recognition method based on edge computing-assisted and gru deep learning network," *Applied Sciences*, vol. 13, p. 9059, 2023.

[117] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 286–293, 2021.

[118] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[119] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[120] J. Li, L. Yao, B. Li, X. Wang, and C. Sammut, "Multi-agent transformer networks for multimodal human activity recognition," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, (New York, NY, USA), p. 1135–1145, Association for Computing Machinery, 2022.

[121] S. Sowmiya and D. Menaka, "Transformer model for human activity recognition using iot wearables," in *High Performance Computing, Smart Devices and Networks* (R. Malhotra, L. Sumalatha, S. M. W. Yassin, R. Patgiri, and N. B. Muppalaneni, eds.), (Singapore), pp. 287–300, Springer Nature Singapore, 2024.