International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Systematic Review of various Fusion Techniques for Human Activity Recognition

Sandeep Kaur Gill[1*], Anju Sharma[2]

Submitted: 19/01/2024 **Revised**: 28/02/2024 **Accepted**: 05/03/2024

Abstract: Human Activity Recognition (HAR) has fetched considerable prominence as it plays a critical role in a wide number of applications ranging from healthcare monitoring to human-computer interaction. Gaining accuracy as well as efficiency in the process of representing and recognizing an activity is one of the critical goals in the domain. Apart from developing via technical perspective, utilizing the resources and technicalities in hand to the fullest is another significant criterion to gain accuracy and efficiency in any process. Embedding multiplicity in the sub-tasks via the fusion of multiple sources is one of the options to ensure that the resources enrolled in the task are being utilized effectively and to the fullest. In HAR, fusion could be considered from three perspectives, namely, data fusion, feature fusion and classifier fusion. In this paper, a survey of research work that implemented fusion from any of the three perspectives in the process of recognizing the activity has been generated. Apart from embedding multiplicity via fusion of each criterion on an individual basis, multiplicity could be embedded in the domain via the perspective of number of modes of fusion as well. The review also presents the work that implemented fusion via multiple criteria to optimize the process of recognizing the activity being executed. Section 1 generates an overview of technicalities in hand to represent and recognize the activity and justifies the criteria of embedding multiplicity in the process of activity recognition, Section 2 discusses three modes of fusion to embed multiplicity and gain both accuracy as well as efficiency in the process of HAR and Section 3 gives an overview of open research issues and finally Section 4 justifies the importance of the criteria of fusion.

Keywords: *Human Activity Recognition; Dimensionality Reduction; Data Fusion; Feature Fusion; Classifier Fusion*

1. Introduction

Motion is an inseparable part of the universe and a fundamental aspect of human life. Apart from getting the necessary and urgent activities executed, motion acts as a vital ingredient for his life. Activity is a generic term that could either be atomic in the form of an action or a gesture or it could be a sequence of multiple primitive actions executed in a particular order to either generate any event or execute an interaction. The process of fetching information about motion via various sensors and associating the fetched data with a particular activity name is referred to as Human Activity Recognition (HAR). HAR is one of the hot topics of research in the domain of computer vision as it is associated with the evolution of applications in important domains like medical, security, virtual reality, sports video analysis and human-computer interaction (HCI) (Liu et al. 2018). As it is in the phase of development, it still encounters several challenges such as interclass similarity, intraclass variation, group activities and complex backgrounds that are required to be tackled by the recognition system. Morshed et al. (2023) generated a survey of recent developments in the domain of activity recognition. Various hand-crafted techniques, deep learning techniques and attention-based approaches to represent and recognize the activity being executed are briefly described. Based on the method implemented for the task of feature extraction, the process of recognition could be based on hand-crafted or deep learning or attention-based approach. As described in Fig. 1, the task of activity recognition is a group of several sub-tasks. After fetching data from multiple sources, data is prepared via preprocessing, segmentation and dimensionality reduction and finally fed to the learning process. Both deep learning features, as well as shallow features are extracted from the resulting data. Deep learning features form the basis of training for the classifier. The classifier fetches order of actions via both shallow and deep learning features to recognize the activity being executed. Finally, the result generated by the classifier is evaluated based on various evaluation metrics .

For an accurate and efficient HAR system, the tasks of representing the data, fetching the relevant features from the acquired data and finally recognizing the associated activity are required to be executed efficiently and in an accurate manner. To achieve this goal, the process could be worked upon from several perspectives. Apart from technical upliftment, embedding multiplicity in the sub-tasks is another criterion to gain accuracy and efficiency in any process. To build a robust activity recognition system that could tackle several issues such as interference from sources, multiplicity could be enrolled

[1,2]*Maharaja Ranjit Singh Punjab Technical University, Bathinda*
[1,2]*Department of Computational Sciences*
E-mail: [1]*sandeepforphd@gmail.com,* [2]*anjusharma@mrsptu.ac.in*

in the process by embedding fusion via several modes (Faiz et al. 2023 a). To start with, multiplicity could be embedded in the number of sensors employed to fetch the data. Depending on how the system is implemented, it could either be associated with a contact-based method or a vision-based remote method. Contact-based systems demand physical interaction with the user, but due to low battery life and data security concerns, they have been abandoned. In vision-based systems, human activity is interpreted as a sequence of changes in the view and pose of the executor with time. Unlike contact-based systems, they do not demand direct contact with the user's body and thus do not intrude on his privacy. The ease of implementation and nonintrusive nature of vision-based systems lead to their broad usage(Channi et al. 2023). After extracting features from data by all sensors, all the features are grouped and processed in a combined manner by multiple classifiers and finally, the results derived by multiple classifiers are combined to generate the final result. An overview of three types of fusion, namely, data fusion, feature fusion and finally classifier fusion has been presented.

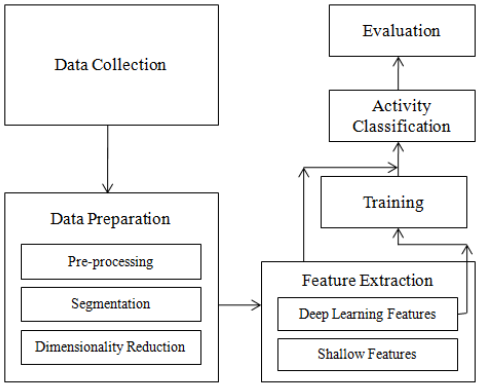

Fig. 1: Human Activity Recognition

To build the base for studying the modes of fusion in the process of HAR, Section 1 generates an overview of the atomic tasks enrolled in the process. It includes the tasks of data collection, data representation, feature extraction to recognize the associated actions and finally combining multiple gestures and actions to conclude the activity being executed. Section 2 recognizes and illustrates three modes to embed fusion, Section 3 lists the open issues and Section 4 concludes the importance of embedding multiplicity in the process of HAR.

1.1 Data Collection

In the realm of activity recognition, the primary undertaking entails the acquisition of data from diverse sources. As described in Fig.2, depending on the mode of acquiring data, HAR may be based on data fetched via sensing devices or smartphone or radar based or vision based devices. Sensing devices such as accelerometer, gyroscope, magnetometer could be embedded in various body-worn entities such as watches, helmets and bands. Sensing devices are cheap and absorb less power but are inconvenient for usage as they are required to be worn by the user. The process of data collection has been rendered notably convenient for the end-user by embedding sensors in the smartphones. Smartphones could be located either in hand or on chest or thigh or a bag and sensors fetch the angular velocity and acceleration of the particular position. Though smartphones are user convenient but they are position dependent as the movement of sensor varies based on their location. In 2018, B. Almaslukh et al. developed a position-independent activity detection system that collected data via smartphone. While the integration of GPS into the smartphone-based system enhanced its functionality, it concurrently engendered elevated costs and augmented power consumption.

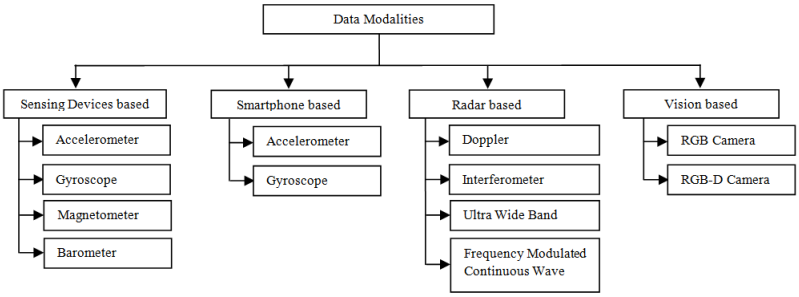

Fig. 2: HAR Data Modalities

Radar based and vision based sources of data are device-free approaches for data collection. Radar-based system is a contactless and insensitive to environmental effects such as daylight. The system is based on radio-waves and data gets collected via reflection. As reflected waves possess change in frequency due to collision with the body, various properties such as shape, size and movement could be extracted from the reflected signal to infer the activity being executed. Vision based sources fetch the data from RGB and RGB-D cameras. As cameras have large coverage, are easy to use, more accessible and cheap, they have been largely implemented. Camera-based systems are the simple and stationary solutions for the task of surveillance.

1.2 Data Representation

After collecting the data, it is prepared for processing by the recognition systems. Depending on the inclusion of motion in the representation, an activity could be represented either in a static or dynamic manner. Static representation does not consider the motion but captures the position and orientation of the portions of body while dynamic representation captures the motion of body parts. As presented in Fig. 3, an activity is composed of a sequence of actions and depending on whether the action enrolled in the execution of activity is considered in a static or a dynamic manner, actions could be described from spatial or temporal perspective respectively.

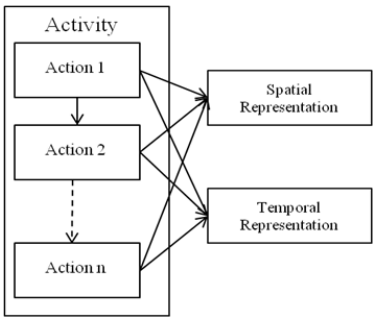

Fig. 3: Criteria of Activity Representation

Spatial representations such as body models, image models and spatial statistics identify and extract the objects of interest from the video and generate the representation of an action. Body models such as kinematic joint model (Ghorbel et al. 2018) fetch the pose of body from the features with the help of spatial structure of the body during the action execution and recognize the action from the group of paths followed by the joints of body. To detect the activity, local features and the related statistics are fetched from the surrounding regions via spatial statistics of those regions. Spatio Temporal Interest Points (STIPs) (Ma et al. 2018) are calculated and distribution of local features across the space is fetched. Spatial statistics could follow volume-based or trajectory-based approaches. Image models such as Motion Energy Images, Motion History Images (Ijjina and Chalavadi 2017), optical flow and silhouettes use a regular grid bounded by the region centered across action executor to describe the action. Temporal representations (Sharaf et al. 2015) such as action templates, action grammars and temporal statistics fetch the temporal attributes and consider the distribution of several features over time to represent the action. They acquire the spatial relations between various components of the executor and fetch the changes between them with time. Action templates such as Wavelet representations (Abid et al. 2021), Fourier Transform (Vemulapalli and Chellappa 2016) acquire the appearance of blocks of features and dynamics on a temporal basis. A set of frames are fetched to compute the templates that could be made the basis for recognizing the actions and the associated activity. Action grammars such as regressive models, Hidden Markov Models (Kuehne et al. 2014) and context-free grammars represent the action as a sequence of moments fetching the appearance features and dynamics.

1.3 Feature Extraction for Activity Recognition

After describing the action, next step in the process of HAR is to fetch the features related to all the actions associated with the execution of any activity. From the spatial and temporal representations of the action, appearance-based recognition approaches fetch the shape and motion feature. Methods based on shape implement foreground segmentation to capture silhouette, local region, contour points and several geometric features while methods based on motion represent the action in the form of Motion History Volume. Depending on the characteristics of features fetched for the recognition of activity, they may be deep features or handcrafted features related to any of the parameters of the action. Non-deep learning approaches (Pareek and Thakkar 2023 ; Xiao and Song 2018; L. Liu et al. 2016; Gao et al. 2015) for feature learning include genetic programming, dictionary learning and evolutionary learning. Genetic programming, as the name specifies, follows evolutionary criteria. It searches the list of possible solutions and

discovers functional relations between several features in data. Pareek and Thakkar (2023) operate upon depth-based data, employ evolutionary learning criteria to fetch the hidden parameters and thus optimize the performance of recognition system. Xiao and Song (2018) represent video via hierarchical model and implement the hierarchical dynamic Bayesian network (HDBN) for the task of activity recognition. L. Liu et al. (2016) fetch the spatiotemporal features and evolve the motion feature descriptor. Support Vector Machine (SVM) calculates the average cross-validation classification error upon the training-set, based on which the genetic programming fitness function is evaluated. This process continues and the best solution obtained is selected as the descriptor. Gao et al. (2015) generate the multi-view bag of words representation and fuse multiple views to recognize the activity being executed. Multiple views are fused, overlapping interest points are removed, latent correlation among the resulting views is discovered and joint dictionary learning criteria is implemented for recognizing the associated activity.

Deep learning methods for HAR (Jain et al. 2021; Xia et al. 2020; Ravanbakhsh et al. 2017; Srivastava et al. 2015) extract the features of activity in a fully automated manner. Based on the learning criteria implemented, deep learning method could either be a generative or a discriminative method. Generative methods implement unsupervised learning criteria to represent unlabelled data. Approaches such as Variational Autoencoders (Jain et al. 2021), Generative Adversarial Networks (Ravanbakhsh et al. 2017) and Autoencoders (Srivastava et al. 2015) are some of the most frequently used approaches in generative models. Discriminative models are the supervised models with a hierarchical structure that categorize the data into several classes using a hierarchical learning strategy. Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are the most frequently used discriminative methods. In 2020, Xia et al. designed an HAR model that combined LSTM and CNN for recognizing the action. LSTM fetched the temporal features while CNN fetched the spatial features of the action and SoftMax function generated the probability distribution of all classes of actions. Generative and discriminative methods could be hybridized to fetch the advantages of both criteria and gain both efficiency as well as accuracy in the task of behavior recognition. Z. Wu et al. (2012) hybridize Post Divergence based discriminative classifiers with Hidden Markov Model (HMM) and propose a hybrid approach for recognizing the activity being executed. With the development of deep learning methodology, supervised and unsupervised approaches have been hybridized. In 2021, Cordell et al. detect the anomaly in the user response on the basis of a hybrid criterion (Cordell et al. 2021) to fetch the anomalous patterns in the test data.

2. Fusion for Ambient Assisted Living

Embedding multiplicity from several perspectives is one of the criteria to gain both accuracy as well as efficiency in any process. In the same manner, depending on the tasks associated with the process of HAR, it could be entitled to multiplicity. First task in the process is the collection of data using hardware components called sensors. As proper representation of data is one of the prerequisites for accurate recognition of activity, hence various data modalities have been proposed for representing the activities. With various affordable and accurate sensors, a variety of criteria for the representation of data were developed (Faiz et al. 2023 b). Major modes for the task of data encoding include RGB, depth, infrared, skeleton, audio, radar, acceleration, event stream, point-cloud and Wi-Fi signal, each of which has distinct properties for usage in different application scenarios.

Decrease in the cost of sensors has generated a trend that uses multiple sensors, which may be mobile or wearable, for the task of fetching the data. Use of multi-modal sensor data increases the accuracy of the process of activity recognition. Instead of relying on single sensor data or on only handcrafted features of the fetched data, multiple classes of features fetched via multiple as well as various kinds of sensors could be combined, and operated upon by multiple classifiers, thus generating a higher level of generalization. It has been proven that embedding multiplicity proves effective as well as efficient in the task of recognizing a complex activity. Gravina et al. (2017) present a review of various techniques that embed multi-sensor fusion and discuss various parameters and properties that decide the choice of fusion at all the levels, namely, data-level, feature-level and classifier-level. Data fusion is an early fusion criterion that combines the data in the input phase before processing and extracting features from it. Integrated data is fed to the classifier to fetch the features and recognize the associated action. Feature fusion embeds fusion in the process at the intermediate level. Multiple features extracted from the input data are combined before being fed to the classifier. Classifier fusion is a late fusion criterion that combines the result produced by each of the classifiers. Fusion from various perspectives that can be embedded in the process of HAR is illustrated in Fig. 4.

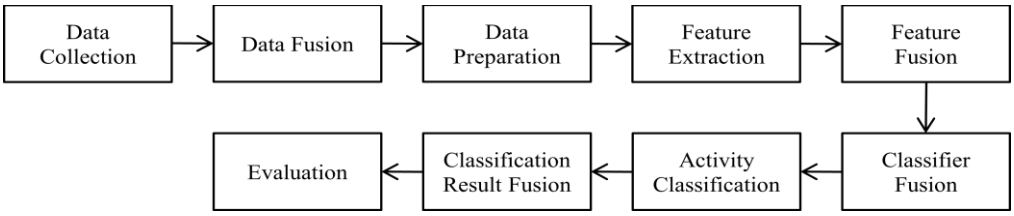

Fig. 4: Levels of Fusion in Human Activity Recognition

2.1 Data Fusion

Depending on the multiplicity in the number of views, the sequence of actions may either belong to a single view or to multiple views. Depending on the nature of sensors employed, multi-view activity recognition may deal with either uni-modal data or with multi-modal data. The first class of data fusion employs a homogeneous set of sensors at multiple points. It assumes that the data fetched from all views is complementary and the extra information fetched helps in recognizing the activity accurately. As data fetched by each viewpoint captures distinctive aspects of the activity, multiplicity in the number of homogenous sensors leads to improvement in the process of recognizing the activity. The second class of data fusion fetches the action sequences from multiple types of sensors. As described in Fig.5, multiple sensors, homogeneous or heterogeneous, are employed to fetch the data. This decreases the uncertainty due to several interfering scenarios such as displacement, and reduces the effect of indirect capture of data, thus embedding robustness in the system. Qiu et al. (2022) describe various kinds of sensors to fetch the data and various criteria related to multi-modal and multi-location data fusion (Chaturvedi et al. 2022)

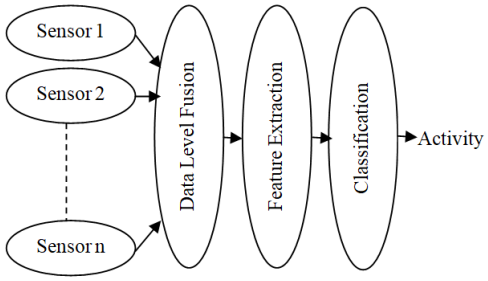

Fig. 5: Data Fusion in Human Activity Recognition

2.1.1 Homogenous Data Fusion

Multiplicity could be embedded in the number of like sensors enrolled in fetching the data to consolidate similar data. Gumaei et al. (2019) implemented a deep hybrid SRU-GRU (SRU - Simple Recurrent Unit, GRU - Gated Recurrent Unit) model on the mobile health dataset collected by attaching multiple sensors on the right wrist, left ankle and chest of ten subjects. As body sensors are inconvenient and may not fetch relevant data, Holte et al. (2011) applied 3D motion description to the data fetched from multiple views via several motion vector fields. Optical flow in 2D multiframes is computed to detect the motion. As 3D data captures more information than 2D data, the flow fetched from each view is reconstructed and extended to 3D mode. D. Wang et al. (2018) proposed a Dividing and Aggregating Network (DA-Net) to execute multi-view HAR by fetching both view-specific representations as well as view-independent representations of the activities related to all views. Both view classifiers and view-specific activity classifiers get employed to fetch the information about the activity being executed. Prior are used to fetch the probabilities of the view, where probabilities are treated as weights while computing the prediction scores by the activity classifier system. Most discriminative information is fetched from multiple view-specific representations. View-specific information fetched by CNNs from all points is shared amongst each other, thus providing a means to refine the information and generate a more accurate representation for the task of recognizing human activity. The fetched view-specific features and classification results from classifiers at multiple viewpoints are fused via average weighted upon view probabilities to generate the final result.

Apart from fetching data via multiple sources, data imputation is another criterion to gain sufficiency in data. L. Wang et al. (2019) proposed a Generative Multi-View Action Recognition (GM-VAR) criterion that integrated complementary information from multiple views to gain improvement in the task of HAR. As missing data is imputed, thus, irrespective of the completeness of information, it gains compatibility in real-world applications. As various data modalities possess distinct properties, multi-view data fusion via naïve concatenation may induce a negative effect and degrade the system's performance. Moreover, as data fetched from multiple

views is commonly incomplete, partial multi-view data degrades the performance of the system in an inevitable manner. To improve the performance, correlation between various action classes and views could be fetched and utilized. GM-VAR framework adopts the View Correlation Discovery Network (VCDN) and generative adversarial training to overcome the incompleteness in the multi-view data. It learns the instance-level pair-wise cross-view as well as intra-view correlation knowledge to leverage the complementary information among the multiple views, thus overcoming the incompleteness in data and improving the performance of the model.

As information from different sensors attached to different body parts possesses different priorities in the process of recognizing the activity being executed, a weighted criterion could be embedded in the process of combining the data fetched from multiple sources. W. Tao et al. (2021) proposed an attention-based fusion criterion that learnt the level of association of each of the sensors in the process of executing each of the activities. Multiple Inertial Measurement Units (IMUs) are attached to different parts of the human body. Body sensors prove advantageous as ambient sensors get across several challenges such as occlusion due to interfering sources, limitations of range and inability to capture the data related to outdoor activities. IMU sensors are attached to multiple body locations to fetch valid data. Features possessing large discriminative power are selected from each of the sensors with CNNs. High-level features fetched by the sensors are combined on the basis of an attention-based fusion mechanism and then processed by the classifier to generate the result vector. The result vector presents the probability distribution of all activities and finally infers the activity. The proposed attention-based fusion criterion outperforms both early fusion as well as late fusion criteria. It gains an understanding of data from each of the sensors and fetches the correlation to generate the final result.

2.1.2 Heterogeneous Data Fusion

Heterogeneous data from multiple sensor modalities could be combined to provide accuracy and robustness and also decrease uncertainty of the system (Pai et al. 2021; Amrita et al. 2023; Rai et al. 2023). As each sensor possesses different properties such as resolution and frequency of sampling, multiple types of sensors could be enrolled in the process of fetching the data. C. Chen et al. (2017) presented a survey of fusion strategies of data from two modalities – visual sensors and inertial sensors. Various probability estimation techniques that fuse heterogeneous sensor data on the basis of probability density function have been proposed. As readings may be erroneous on an individual basis, taking average helps in covering the error incurred during the process of fetching the data. Deep canonical correlation analysis was proposed in order to fuse the data fetched from multiple sources (Chetty and White, 2016).

H. Wu et al. (2002) proposed Dempster-Shafer theory of evidence as an approach for the task of data fusion. Raw data collected by sensors is translated into context information by an Interface Widget that translates the raw data into the form of pre-defined templates. Dempster-Shafer theory is a generalized Bayesian theory that distributes support upon the proposition and upon all the possible and mutually exclusive facts related to the proposition. Each sensor assigned the figures that presented the beliefs upon all the facts. These figures were combined by the weighted- average criterion where weights are related to the rate of correctness in history. As it computed reliability of the sensors before combining them using the combination rule, this increased the robustness as well as reliability of the detection systems. As data fetched from the sensors may be uncertain, a Majority Consensus combination rule was proposed to tackle uncertainty in the domain (Sebbak and Benhammadi, 2017). A 3-layer IoT-based healthcare system fetches various types of data such as medical history, sensor information, spatial information and contextual information via various body and environmental sensors. Thus, the continuous information of the physical state of the person is contextualized and possible activities and risky situations, if they occur, are recognized and notified.

Apart from variation in the nature of data being fused, variation could be embedded in the phase for data fusion as well. S. Münzner et al. (2017) tested the effect of various early fusion and late fusion criteria to fuse the data. Random Forest algorithm is executed with time-domain features and with both time-domain and frequency-domain features. Four fusion criteria, namely, shared filters hybrid fusion (SF-HF), channel-based late fusion (CB-LF), sensor-based late fusion (SB-LF) and early fusion, are executed. CNN fusion models are evaluated on RBK dataset, and it is concluded that late fusion outperforms early fusion and SF-HF fusion generates the highest accuracy.

Both homogeneous as well as heterogeneous criteria for data fusion could be embedded in the process of HAR. A.A. Liu et al. (2019) proposed a supervised multi-domain and multi-task learning (MDMTL) framework. Multi- domain data, the data that is both multi-view as well as multi-modal, is fetched for gaining view-invariance as well as modality-invariance. Domain-invariant information is fetched from the data and inter-relationship between several categories of actions is explored to facilitate the process of recognizing the activity. Several actions that are fetched are correlated

in a latent manner to model the activity. Actions are represented in both RGB and depth modality, thus gaining robustness against background clutter, occlusions and variations in illumination conditions. Apart from working on multi-domain data, MDMTL implements multi-task learning (MTL). MTL fetches the level of inter-relation among multiple tasks associated with the action, thus improving the power of recognizing the action. S. Chung et al. (2019) also implemented both homogeneous as well as heterogeneous data fusion for the task of recognizing the activity. Eight IMUs are worn upon the body for collecting the data. Collected data is used to train LSTM neural network classifier. Additional data related to the circular motion is fetched by a magnetometer and gyroscope. Data from all sensors is acted upon by LSTM network and the final result is computed based on the weighted average of the probabilities of various classes. Accuracy in the prediction of activity determines the weight that is assigned to each of the classifiers. Results conclude that instead of six sensors, placing only two sensors, one upon the right wrist and the other upon the right ankle, are sufficient to fetch the data for recognizing the activity. L. Schrader et al. (2020) designed a caring system for the elderly and diseased population. It fetched data via three types of sensors – SmartCardia wearable, Myo armbands and activPAL monitors. Twelve sensors were attached to different parts of the human body to fetch the data for developing the HAR system. Two Myo armbands and nine activPAL monitors are enrolled thus implementing both homogeneous as well as heterogeneous modes of data fusion.

Table 1 Implementation of Data Fusion in the process of HAR

Author(s)	Year	Nature of Data	Source of Fusion	Classifier	Dataset(s)
Joshi et al.	2023	Multi-sensor data, Multi-modal data	Contribution Significance Analysis (CSA)	HAR_WCNN	One dataset: CASAS
Vidya and Sasikumar	2022	Multi-sensor data	Pearson's correlation	SVM, KNN, Ensemble Classifier, Decision tree	One dataset: UCI ARem dataset
Tao et al.	2021	Multi-sensor data	Attention-based sensor fusion	DNN	Five datasets: Daily, Skoda, PAMAP2, Sensors, Daphnet
Prakash and Yadav	2020	Multi-sensor data, Multi-modal data	Downsampling	Random Forest	One dataset: Opportunity
Chung et al.	2019	Multi-sensor data, Multi-modal data	Two-level stacking and voting ensembles	LSTM	Data collected via eight IMU sensors (Inertial Measurement Units)
Gumaei et al.	2019	Multimodal body sensor data	Reshaping phase	Deep SRU + GRU (Source/ Gated Recurrent Unit) Neural Network	One dataset: MHEALTH
Liu et al.	2019	Multi-sensor data, Multi-modal data	Instance-level fusion	MDMTL	Three datasets: IXMAS, M^2I DailyActivity3D,
Wang et al.	2019	Multi-view Data (RGB + Depth data)	View Correlation Discovery Network (VCDN)	GAN	Three datasets: UWA, MHAD, DHA
Wang et al.	2018	Cross-subject	View-prediction-guided	DA-Net (Dividing	Two datasets:

		data, Cross-view data	Fusion	and Aggregating Network)	NTU, NUMA
Münzner et al.	2017	Multimodal multi-sensor data	_	Random Forest	Two datasets: RBK (Robert Bosch hospital) , PAMAP2
Sebbak and Benhammadi	2012	Multi-view data	Majority-Consensus combination Rule (MCR)	Evidence theory	Simulation studies

B. Vidya and P. Sasikumar (2022) gained an accuracy of 99.63% in the process of recognizing the activities. Four sensors – three wearable sensors and one environmental sensor collected the data. Three wearable sensors - one at the chest and two upon the ankles, and a tri-axial accelerometer from smart-phone collect the RSS data. Statistical features and three entropy-based features – Shannon entropy, log-energy entropy and approximate entropy, fetched via hybrid Discrete Wavelet Transform (DWT) and Empirical Mode Decomposition (EMD) are processed to recognize the activity. EMD algorithm operates upon RSS data to decompose the data into several Intrinsic Mode Functions (IMFs). From the set of features fetched, valid features - features possessing considerable discriminative power, are extracted on the basis of Pearson's correlation approach. Four classifiers, namely, SVM, KNN, EC (Ensemble Classifier) and DT (Decision Tree) are trained on the basis of extracted features for recognizing the associated activity.

Y. Li et al. (2023) measured contribution of sensors by weighing their respective data on the basis of Contribution Significance Analysis (CSA). An HAR system based on wide time-domain CNN (HAR_WCNN) that can fetch multi-environment sensor data is designed. Spatial Distance Matrix (SDM) is built on the basis of spatial variation of activity trajectories to tackle the noise due to multi-person cross-activities. Sensor noise is adaptively constrained and the contribution of each sensor to a particular type of activity is measured on the basis of statistical methodology. Based on accuracy, the designed WCNN system outperforms several HAR methods. Table 1 illustrates several research works that implemented the criteria of data fusion to recognize the action (Saxena et al. 2022).

2.2 Feature Fusion

As the name states, feature fusion combines multiple properties of data. Sufficient number of features are extracted to associate them with actions and recognize the associated activity. For detecting the activity, low-level features such as color, shadow, motion, texture or edges of the entity are captured and high-level features to describe the movement are computed. Features, which may be homogeneous or heterogeneous, are combined and processed by algorithms such as SVM and decision tree to recognize the associated activity. As presented in Fig. 6, depending on the visual characteristics captured, the features could be categorized as appearance features, shape features and motion features. Appearance features are extracted from local regions of the image and several parameters of the image such as its texture, color, intensity, Haar features (Goyani and Patel, 2017) and Local Binary Pattern (LBP) (Hussain and Triggs, 2010) are captured. Redundant and less informative components are excluded and Haar features are computed only for the dominating components. LBP is a simple and robust description of the appearance of the human body and a histogram of LBPs, computed at the pixel level, is used to encode the region in an image. LBP was extended to generate its extensions such as Non-Redundant LBP (NRLBP), centre symmetric LBP (CSLBP), Local Ternary Pattern (LTP) (Hussain and Triggs, 2010) and Local Intensity Distribution (LID) (Nguyen et al. 2011).

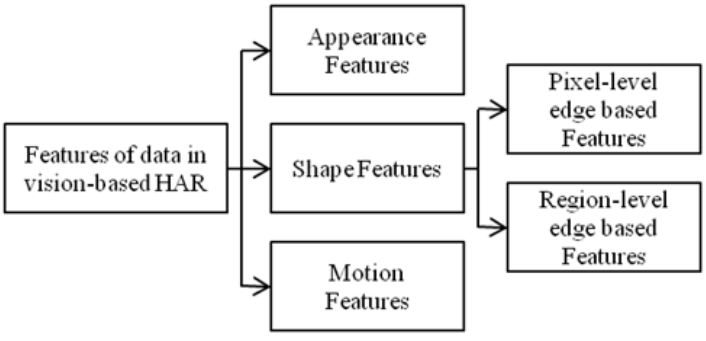

Fig. 6: Features in vision-based HAR

Shape features of an entity are fetched from the sources such as edge-maps. They consider the orientation, location and magnitude of edge pixels. Depending on the atomicity of data, edge-based features could be fetched either at the pixel-level or at the region-level. Data could be interpreted either in the form of rectangular contours or parallel edge segments or small curves named 'edgelets' or binary contours corresponding to various poses and viewpoints. Region-level edge-based features are more adaptive to the deformation of shape of the body at local level. They could be computed by quantizing the region-specific information into multiple discrete values and then accumulated to generate a Histogram of Oriented Gradients (HOG). Motion is one of the most important parameters for describing the entity. After fetching the pose and appearance of object, motion features extract their changes and could be used as a means to differentiate the entities. Motion features capture the temporal difference to generate the temporal features of the image regions. As described in Fig. 7, multiple features could be extracted and fused for the task of classification.

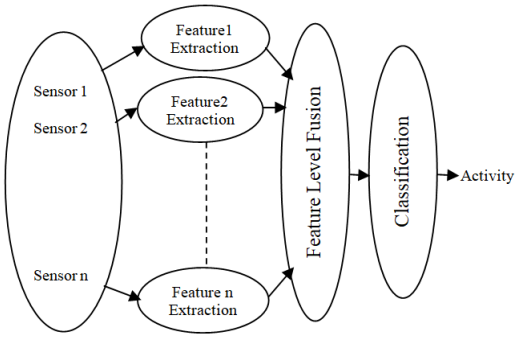

Fig. 7: Feature Fusion in Human Activity Recognition

In 2011, Non-Redundant LBP (NR-LBP) was applied upon the difference images to generate the motion feature. HOGs were computed to fetch the shape features while Histogram of Flow (HOF) (Dalal et al. 2006) was computed to fetch both the boundary as well as motion of various internal regions of the body (Nguyen et al. 2011). Though shape features gained preference due to their ease of use and better differentiating ability, but they are sensitive to interference like clutter while motion features demand the temporal information of the scenario. To describe both kinds of appearance features, the data may be RGB data or skeleton data or depth data. After feature description by various criteria, appropriate feature vectors are selected with various selection methods such as wrapper, filter or embedded feature selection methods. Wrapper-based methods compare the performance of classification algorithms and select the appropriate subset of features. Filter-based methods such as Elitist Binary Wolf Search Algorithm (Li et al. 2017) compare the characteristics of data for an appropriate feature selection. Embedded feature selection methods select the features during classifier training. Some of the feature selection methods include Minimal Redundancy Maximal Relevance (Chernbumroong et al. 2013), RELIEF F (Capela et al. 2015), Discriminant ratio criterion and correlation-based feature selection methods. Selection of appropriate feature vectors is followed by their combination and finally activity recognition (Chen et al. 2017; Chernbumroong et al. 2013; Chetty et al. 2014; Najjar and Gupta 2015; Lara et al. 2012; Peng et al. 2017; Shen et al. 2016; Vidya and Sasikumar 2022; Ward et al. 2006).

Depending on how the features are engineered, they may be categorized as handcrafted features or deep features. Handcrafted features such as time-domain features, frequency-domain features, Hilbert-Huang features (Xu et al. 2016) and ensemble empirical mode decomposition features are the shallow features that are manually fetched and processed by human experts and demand abundant labeled data. Deep features embed automation and deep learning algorithms are employed to fetch them. Various feature sets acquired by the sensors could be combined to integrate multiple properties of the data. A. Abdelgawad and M. Bayoumi (2012) fetch data by multiple sensors and features extracted by each of the sensors are combined using algorithms like Hidden Markov Model (HMM) and Support Vector Machine (SVM). M. Sharif et al. (2017) present a hybrid strategy for classifying the activity being executed. The acquired frames are enhanced by segmenting the moving objects via uniform segmentation and expectation maximization. Contrast stretching technique is implemented to maximally differentiate the foreground from the background. Sliding window concept is used to exclude the static and unnecessary regions and consider only the regions with considerable variation in each successive frame. Moving regions are identified by estimating the velocity and then segmented using expectation-maximization and uniform-distribution-based method. Features are fetched and combined using serial-based fusion technique. Euclidean distance and joint entropy PCA-based methods select the valid features and finally, multi-class SVM

classifies them to recognize the associated activity (Narayan et al. 2023).

S. Islam et al. (2018) recognize the action on the basis of shape of action silhouette. Shape information is derived from the junction points and the patterns followed by the boundary and the action descriptor is constructed on the basis of optical flow. After extracting key frames from several distinct poses, action is described as a geometric pattern (GP) in an 8-directional space. Descriptors are derived on the basis of histograms of GP classes. Lukas-Kanade optical flow (LK OF) points are generated to fetch the temporal variance. Shape information fetched by geometric pattern and flow information fetched by LK OF descriptor are fused in order to exploit both the shape as well as flow information simultaneously. After generating the joint action descriptor on the basis of junction points, optical flow and geometric patterns, the spatial and temporal information gets fused, thus fetching more discriminative power and delivering better performance. M.Uddin and Y.K.Lee (2019) fuse deep spatial features and handcrafted spatiotemporal features for the task of HAR. Deep CNN named Inception-Resnet-v2 fetches the spatial features and a feature descriptor named Weber's law-based Volume Local Gradient Ternary Pattern (WVLGTP) is introduced to fetch the spatiotemporal features. Prior fetches the local features related to all frames to aggregate them and generate the global features related to the video. All the fetched features are pooled to reduce the number of features and gain invariance to translation as well as illumination. WVLGTP is applied to fetch the spatiotemporal features. Both spatial and spatiotemporal features are fed to SVM in the concatenated form to recognize the associated action. C. I. Patel et al. (2020) present a robust descriptor to fetch and fuse the information related to human action and exploit the dissimilarity between various actions. After detecting the moving object and segmenting it from the background, HOG features are fetched and averaged across multiple video frames. Regional features from Fourier HOG fetch the information related to the frequency domain. HOG features, regional features, displacement and velocity of the object are combined, and the resultant feature descriptor is fed to multiple classifiers to prove the effectiveness of the fusion of features in the process of HAR.

Correlation between multiple activities could be fetched and embedded in the HAR system to gain accuracy. Y. Zhang et al. (2021) describe an approach that extracts multiple independent features for recognizing all activities executed either simultaneously or in a sequential manner, on the basis of the correlation between the activities. Activity-specific features are fetched, and activity label-wise correlation map is generated to recognize both exclusive and highly correlated groups of activities. ReLu function is implemented as the activation function and binary cross entropy is used as the loss function for the convolution layers (Mall et al. 2023).

D. Thakur and S. Biswas (2021) declare that as handcrafted features capture the knowledge of experts in the domain, hence extracting the features manually proves important and effective in the process of HAR. Thus, they combine the manually extracted features with the features fetched by deep learning methods. Both features are fed to the softmax layer to detect and recognize the activity. J. Chen et al. (2021) fetch multiple features, select the subset of features and test the selected group via multiple classifiers. Several time-domain, frequency-domain and time-frequency domain features are fetched, and an optimal subset of features is selected on the basis of filter-based methods, wrapper-based methods and embedded-based methods. Irrelevant features are rejected via Relief-F algorithm and correlation between the remaining features is established to minimize the number of features. Selected features are tested with six classifiers, namely, Centre-Nearest neighbors, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naïve-Bayes (NB), Random Forests (RF), Support Vector Machine (SVM) out of which three classifiers – KNN, RF and SVM generate the best result (Mall et al. 2023).

C. Zhang et al. (2022) extract the local features automatically via multiple kernel size CNN and long-term dependencies existing in the data via Gated Recurrent Unit (GRU). After data pre-processing, spatio-temporal features of multiple scales are extracted by three feature extraction blocks differing in the size of the kernel for the process of convolution. CNN extracts the short term and local features from the sensor data. Temporal context between various parts of long-time data is fetched by GRU by integrating the gating units into a cyclic unit thus enabling GRU to capture long-term dependencies. Features extracted by all blocks are fed to the SoftMax layer in a concatenated fashion to finally recognize the activity. Md. M. Islam et al. (2023) develop a multi-level feature fusion and multi-sensor-based data fusion approach in the process of HAR. An IoT (Internet of Things) system is deployed to implement the proposed architecture in the medical domain in an IoHT environment. CAM (Channel Attention Module) and SAM (Spatial Attention Module) are embedded in CNN to extract the channel features and spatial features from the data respectively (Mall et al. 2023).

Table 2 Implementation of Feature Fusion in the process of HAR

Author (s)	Year	Source of Feature extraction	Sources of Fusion	Classifier(s)	Dataset(s)
Islam et al.	2023	Multi-head CNN, CBAM	Spatial features, High level temporal features	Softmax function	One dataset: UP-Fall detection dataset
Indhumathi et al.	2022	Attentive Correlated Temporal Feature	Structural features, Temporal features	MultiSVM	Two datasets: HMDB51, UCF101
Zhang et al.	2022	CNN, GRU	Local features, Long term dependencies	Softmax function	Three datasets: WISDM, UCI-HAR, PAMAP2
Chen et al.	2021	Genetic algorithm based feature selection algorithm	Time-domain features, Frequency-domain features, Time-frequency domain features	Centre-Nearest neighbors, KNN,LDA, Naïve-Bayes (NB), RF, SVM	Recorded by NORAXON
Zhang et al.	2021	Independent Spatio-temporal attention	Correlation map to map Activity Specific Features	PyTorch	Three datasets: Charades, AVA, Volleyball dataset
Muralikrishna et al.	2020	VGG-19 deep NN	Structural features, Temporal features	SVM, SoftMax function	Three datasets: KTH, UTKinect, MSR Action3D
Patel et al.	2020	Human Visual Attention Model	HOG, velocity, displacement	ANN, SVM, Multiple Kernel Learning (MKL) Meta-cognitive Neural Network (McNN), Late fusion	Five datasets: KTH, Weizmann, UCF11, HMDB51, UCF101
Uddin and Lee	2019	Weber's law based Volume Local Gradient Ternary Pattern	Spatial features, Spatiotemporal features	SVM	Five datasets: KTH,UCF101,Hollywood, UCF sports action dataset, UT-Interaction dataset
Islam et al.	2018	Junction Points, Geometric patterns	Shape information features, Motion Information features	Improved TF-IDF	Two datasets: Weizmann, SBU Kinect Interaction dataset
Sharif et al.	2017	Euclidean Distance, Joint Entropy-PCA-based method	LBP with HOG, Harlick features	Multi-class SVM	Four datasets: Weizmann, KTH, UIUC, Muhavi
Zu et al.	2016	Empirical Mode Decomposition (EMD), Hilbert Spectral Analysis	Instantaneous amplitude (IA), Instantaneous frequency (IF), Instantaneous energy density (IE), Marginal spectrum (MS)	Back propagation (BP) neural network	One dataset: PAMAP2 (Physical Activity Monitoring for Aging People dataset2)

As action possesses spatial and temporal features, considerable research has been executed to recognize the actions on the basis of their combination (Muralikrishna et al. 2020; Indhumathi et al. 2022). S. Muralikrishna et al. (2020) propose a methodology for recognizing the actions using two types of features, namely, structural variation fetching the shape and temporal displacement fetching the dynamics of skeletal joints. After estimating the pose via VGG-19 deep neural network, structural features are captured by fetching angles between all pairs of joints. For KTH dataset, angle between the joints is fetched via OpenPose to extract information related to the pose. For other datasets, namely, UTKinect and MSR Action3D, information about the pose is extracted via readings from the sensors. Angle is quantized to b-bits via angle binning and minute variations in the angles are suppressed via quantization. This embeds robustness in the system and tackles structural variation encountered by action execution. After fetching the pose information, variation in the pose is extracted via temporal features, thus capturing the dynamics of all joints. Both structural and temporal features are fed to SVM for classification via the SoftMax function.

C. Indhumathi et al. (2022) fetch spatial features from the keyframes and temporal features via ACTF (Attentive Correlated Temporal Feature). Both types of features, spatial and temporal, are fed to the SVM classifier to recognize the action being executed. N. Jaouedi et al. (2020) built a strong feature vector based on spatial and motion features to gain accuracy in the process of recognizing human actions. Table 2 illustrates several research works that implemented the criteria of fusion of multiple features in the process of Human Activity Recognition.

2.3 Classifier Fusion

To tackle challenges such as uncertainty, high dimensionality and data ambiguity, complex systems could be handled by Multi-Classifier Systems (MCS) that combine multiple classifiers to generate the final result. Hybrid approaches enhance the performance by fetching the strengths of individual classifiers (Joshi et al. 2020; Ponti 2011). The combination of multiple classifiers, each trained upon different data, helps in tackling overfitting and increases the probability of finding the optimal solution. Multiple classification models, which may be homogenous or heterogeneous, could be combined to gain accuracy as well as efficiency in the process of recognizing the activity. MCS tackles the extreme cases of both data scarcity as well as the event of huge amount of data. If the data is scarce, then bootstrapping methods such as bagging and boosting could be exploited, while if the data is overflowing, then data could be partitioned among multiple classifiers and decision by all classifiers could be merged by any combination rule. Several criteria such as simple majority and Dempster-Shafer theory of evidence (Rogova 1994) could be implemented to combine the results of multiple classifiers (Naraya et al. 2023).

Classifiers could be combined in both sequential as well as parallel fashion to generate a multiple-classifier system. In sequential architecture, the classifiers are arranged in sequence as per their ability to estimate the certainty of classification while in parallel architecture, all classifiers are trained by the same training samples and output of all classifiers is combined to generate the final result. Adaboost, one of the topologies vastly applied in the task of data mining, follows sequential topology to arrange the classifiers (Freund and Schapire 1997). As shown in Fig. 8, in parallel topology, same input data is fed to all the classifiers, each classifier implements its respective support function, and all functions are combined to form the final function and generate the result (Narayan et al. 2023).

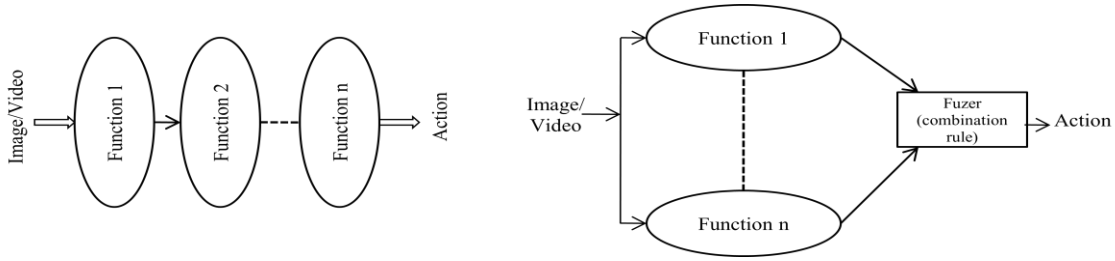

Fig. 8: Support function fusion in serial and parallel fashion

Apart from fusing several support functions implemented by the classifiers by applying fusion at the topology level, fusion could also be implemented upon the class labels generated by multiple classifiers as presented in Fig. 9. Sequential architecture follows an ordered set of rules, where, if the result generated by the primary classifier is not trustworthy due to low confidence, then the data is fed to the next classifier in sequence and the process continues until the result gets generated. In parallel architecture, decisions generated by all the classifiers are combined to generate the final result. Class label fusion implements any of the voting schemes. Major voting schemes include unanimous voting, simple majority and majority voting (Narayan et al. 2023).

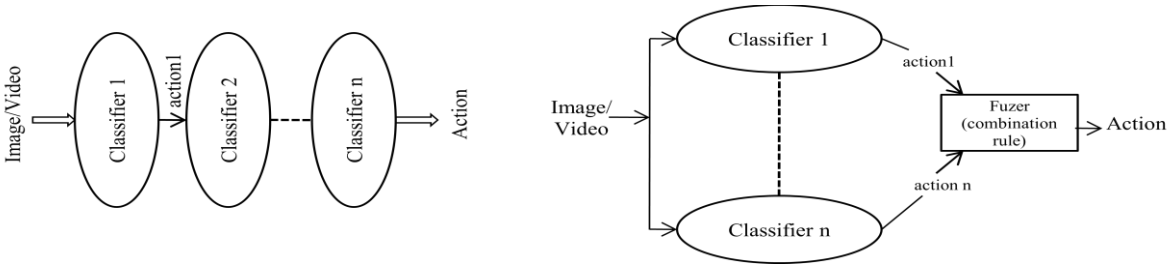

Fig. 9: Class label fusion in serial and parallel fashion

As shown in Fig. 10, apart from all functions generating individual results that are combined to generate the final result, several functions could also be combined. Aggregation is one of the important methods of fusing the support functions as it counteracts overfitting that may be encountered by individual classifiers. Support function fusion implements support functions that provide a score for each of the decisions taken by individual classifiers, and thus derive the estimated likelihood of the class. Various training strategies to fuse the individual results include perceptron-like learning, evolutionary algorithm and ensemble pruning methods. Decisions by heterogeneous classifiers could be combined via stacking or Behavior-Knowledge Space method (Huang and Suen 1995).

For generating an accurate result, that too at the minimal complexity, choosing the appropriate base classifier, appropriate ensemble design technique and strategy for fusion is one of the challenges in the task of recognizing the human activity (Kumar et al. 2015; Wozniak et al. 2014). Implementation of multiple classifier systems was justified by T. G. Dietterich (2000). To deal with insufficient training data, bootstrapping ensemble methods chose the subset of data with replacement on a random basis and plurality voting was implemented to combine the output generated by each classifier. This process of combining the classifiers, each trained upon a separate subset of data, tackled the scenario of overfitting and increased the likelihood of generating an optimal solution. Various fusion strategies include class label fusion, trainable fusion and support function fusion (Nweke et al. 2019).

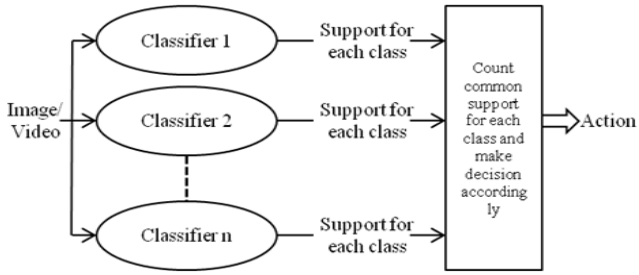

Fig. 10: Support Function Combination in Classifier system

Z. Wu et al. (2012) embed two classifiers in the HAR system. HMM is combined with discriminative classifiers on the basis of PD (Post Divergence) to fetch valid data in an efficient manner. An extension of PD, named PD3 is proposed that proves to be robust in classifying the actions and outperforms the component individual classifiers. Variable-length sequence of joints is mapped to a fixed-dimension feature vector by the feature mappings that are built via post divergence. The mapping derived by PD fetches the discriminative information to generate a vector by utilizing the hidden variables to the fullest. The generated vector is fed to SVM. Valid data is fetched to gain high accuracy in the process of recognizing the associated action by focusing on selective parts of the body.

To embed multiplicity in the number of classifiers in the HAR model, M. A. Bagheri et al. (2014) ensemble five classifiers in the system. Each classifier is fed with a separate set of features as input for heterogeneous action learning and recognition. Three types of classifiers, namely, KNN, NBNN and SVM are embedded in the system. KNN classifier acts upon the positions of joints, NBNN classifier acts upon the information about postures of the body while SVM acts upon three types of information, namely, bag of skeleton words, extreme features and wavelet coefficients of time series. Multiple features such as relative positions of skeleton joints, motion information of postures, bag of skeleton words, wavelet coefficients of time series and extreme features are fed to multiple classifiers, namely, KNN-classifier, NBNN-classifier, SVM classifier. The results generated are combined based on Dempster-Shafer theory of fusion to generate the final result. As both feature representation and classification via an algorithm are improved via

fusion, hence efficiency of the HAR system gets enhanced.

F. Ordóñez and D. Roggen (2016) propose a wearable DeepConvLSTM, a deep neural network consisting of CNN and LSTM layers that recognizes both gestures as well as activities that may be static or dynamic. Convolution layers extract the features from the activity in the form of feature maps while the recurrent layer fetches temporal dynamics. Both homogeneous as well as heterogeneous sensor modalities are worked upon by the framework. It works first upon the data fetched solely from various accelerometers, then upon the data from both accelerometers and gyroscopes and finally upon the data fetched from accelerometers, gyroscopes and magnetic sensors. It does not demand much data pre-processing and thus gains efficiency. Apart from increasing efficiency, multiplicity in the recognition techniques employed helps in tackling many complicated scenarios such as multiplicity and similarity in the actions. Y.Guan and T.Plötz (2017) tackle the scenario by ensembling the LSTM networks and fetching the data using wearables. Ensembles of deep LSTM outperform individual LSTM networks thus generating performance irrespective of shortage of valid and balanced data.

Apart from embedding fusion solely either via fusion of data or features or classifiers, fusion in the process of HAR could also be embedded from multiple perspectives. C. I. Patel et al. (2018) present six fusion models applied at three levels, namely, early fusion, intermediate fusion and late fusion with two models at each of the levels of fusion. Early fusion refers to the process of detecting the moving object and fetching the relevant data from multiple sources. Intermediate fusion refers to the fusion of features enrolled in the task while late fusion refers to the fusion of classifiers in the task of recognizing human action. After resizing the moving object to the dimension 128 * 64, four features are extracted from its periphery, namely, HOG average over ten frames (HOGAVG10), DWT upon ten frames (DWT_TEMP10), displacement of centroid of the object (DISP10) and Local Binary Pattern (LBP10). The average of features over ten overlapping frames is computed to embed robustness in the extracted features. For example, the difference between the centroids (DISP10) is computed upon the offset of ten video frames to fetch the displacement that is used to calculate the velocity of the object. As velocity captures both the speed as well as the direction of motion, the captured data proves efficient in the task of recognizing the action and concluding the activity being executed. Local Binary Pattern (LBP) is fetched to specify the movement of the human body. Features fetched are consolidated and fused via multiple options by concatenation to group the data in an accurate manner and generate finer results. Finally, classification is executed by SVM and ANN and late fusion is embedded via four techniques, namely, Decision Combination of Neural Network (DCNN), Choquet's Fuzzy Integral (CFI), Decision Template (DT) and Dynamic Weighting by Averaged Distances (DWAD). In DCNN, each pair of classifier 'i' and class 'k' has an associated score s_{ik} that is fed to every input node. Weight w_{ijk} acts upon the output generated by the classifier when fed to the output node 'j' and the system generates the result of action recognition based on the maximum response at the output layer node. CFI is another method of combining the results by multiple classifiers. It is the fuzzy average of the classification scores that is computed to generate the final result. DT combines the classification results by all classifiers to generate a fuzzy classifier. Decision is taken on the basis of level of similarity calculated via Euclidean distance between the fuzzy template of the class and the profile of data to be classified. DWAD (Valdovinos and Sanchez 2009) is another method that combines the decisions taken by multiple classifiers.

I. Aydin (2018) built an action recognition framework on the basis of a combination of three classifiers, namely, support vector machine (SVM), Linear Discriminant Analysis (LDA) and neural networks (NN). The proposed hybrid HAR method is based on cuckoo search and fuzzy integral to recognize human action. Sixty-six temporal and frequency-domain features across the axes that possess greater distinguishing ability are fetched from the input signals and fed to three classifiers. The classifiers act upon the features fetched by three accelerometer sensor axes to generate individual results. The results generated by the three classifiers are integrated on the basis of fuzzy integral to generate the final result. Cuckoo search algorithm is implemented to upgrade the confidence level of three classifiers in order to optimize the fuzzy integral parameter (Kumar et al. 2022)

Table 3 Implementation of Classifier Fusion in the process of HAR

Ref.	Year	Mode of Fusion	Source of Fusion	Classifier	Dataset(s)
Khan et al.	2022	Feature Fusion + Classifier Fusion	Spatial Features + Temporal Features	CNN + LSTM	Self generated dataset

Roche et al.	2022	Data Fusion + Classifier Fusion	RGB data + Point Cloud data	CNN + RPN	Two datasets: SYSU-3D, LboroLDNHAR
Ihianle et al.	2020	Feature Fusion + Classifier Fusion	Spatial Features + Temporal Features	CNN + LSTM	Two datasets: WISDM, MHEALTH
Aydin	2018	Classifier Fusion	Frequency domain features + Temporal features	ANN+ LDA+ SVM	Public dataset 6 actions: Sitting, laying, standing, upstairs, downstairs, walking
Patel et al.	2018	Feature Fusion + Classifier Fusion	7 combinations: (HOG+ Centroid) / LBP / HWT / Displacement / (Displacement+ Velocity)/ (Displacement+ Velocity + HWT)/(Displacement+ Velocity+HWT+LBP)	SVM + ANN + MKL	Two datasets: ASLAN, UCF11 benchmark dataset
Guan and Plotz	2017	Classifier Fusion	LSTM Networks	LSTM Networks	Three datasets: OPPORTUNITY, PAMAP2, Skoda
Li et al.	2017	Feature Fusion + Classifier Fusion	Spatial Features + Temporal Features	CNN + LSTM	Three datasets: Trauma Resuscitation, Charades dataset, Olympic Sports dataset
Ordóñez and Roggen	2016	Classifier Fusion	Temporal Features	CNN + LSTM	Two datasets: OPPORTUNITY, Skoda
Bagheri et al.	2014	Classifier Fusion	Positions of skeleton joints, posture information, Bag of Skeleton Words, Wavelet Coefficient of Time series, Extreme Features	KNN + NBNN + SVM	Two datasets: Chalearn, MSR-Action3D
Wu et al.	2012	Classifier fusion	Joint Sequences + Feature Mappings	SVM	3D MoCap

Apart from embedding multiplicity via any of the three fusion criteria, multiplicity could be enrolled in the number of fusion criteria as well. Some researchers embedded feature fusion and classifier fusion in HAR system (Li et al. 2017; Ihianle et al. 2020; Khan et al. 2022). X. Li et al. (2017) generated a multi-modal CNN-LSTM structure by attaching ConvNet and LSTM in a serial order. CNN extracts the spatial features while LSTM deals with temporal features in order to recognize the associated activity. Analysis of dataset related to forty-two trauma resuscitations concluded that approximately 50% of the instances were associated with multiple activities. To fetch and recognize multiple activities, the temporal relation between several spatial features related to multiple activities getting executed simultaneously is fetched as a binary code. A prediction code is generated by the encoder framework. After data pre-processing, ConvNet generates feature vectors and LSTM generates the temporal association between them and based on the spatio-temporal features fetched, sigmoid activation layer recognizes the concurrent activities being executed. I. K. Ihianle et al. (2020) design a multi-channel CNN Bidirectional LSTM (MCBLSTM) that consists of three stacked channels of CNN and BLSTM layers that read the same sensor data, extract the features and concatenate them to finally recognize the activity. Khan et al. (2022) generated a dataset consisting of twelve activities via Kinect Sensor V2 that extracted

twenty-five joints from the human body. Twenty different participants execute twelve activities and data is divided into sequences of length thirty, sixty, ninety, one hundred twenty and one hundred fifty frames. Two layers of 1-D CNN, filter size sixty-four and one hundred twenty-eight respectively, activated by ReLU function, fetch the features that are forwarded to the LSTM layer. Two LSTM layers are followed by a flatten layer and a dense layer with SoftMax activation function to generate the final result.

Roche et al. (2022) embed data fusion and classifier fusion for recognizing the activity. The authors aim at addressing three major challenges. First, recognizing complex interactions that are dependent on the factors such as gap between the agents, direction of motion and location of the actor is one of the challenges. Second, as not all the features that are desired in the fetched data could be captured by each type of data source, multi-modal data is opted as the choice. Third, as supervised machine learning criteria demand an abundant quantity of data for training, available data is required to be utilized fully and in the most efficient manner. To address these challenges, the authors propose a framework that embeds sensor fusion as well as classifier fusion. RGB data and point cloud data are fetched and processed by CNN and Region Proposal Network (RPN) which recognize the region of interest (ROI). The detected region is projected upon the LiDAR data to generate 3D ROI that is passed to the classifier to recognize the activity.

Table 3 illustrates several research works that combined multiple classifiers to generate the result of the process of recognizing the activities. Apart from implementing each of the fusion criteria on an individual basis, multiple criteria of fusion have also been associated with the process of HAR simultaneously. As described in Fig. 11, more than 17% of research in the domain embedded multiplicity in the number of criteria that are associated with fusion.

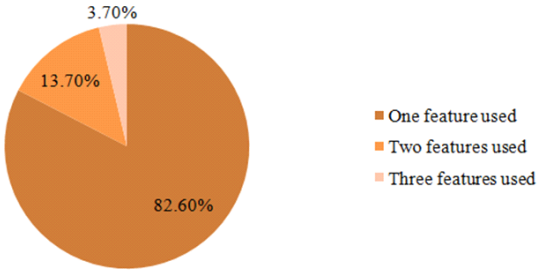

Fig. 11: Multiplicity in the criteria of fusion in HAR

Among the three fusion strategies, data fusion is logically the simplest and technically straightforward to implement, as opposed to classifier fusion which is the most complicated among all the fusion criteria. Owing to simplicity, research in the domain has majorly focused on data fusion. As described in Fig. 12, nearly 53.69% of research has implemented the criteria of data fusion, either solely or in combination with other fusion criteria.

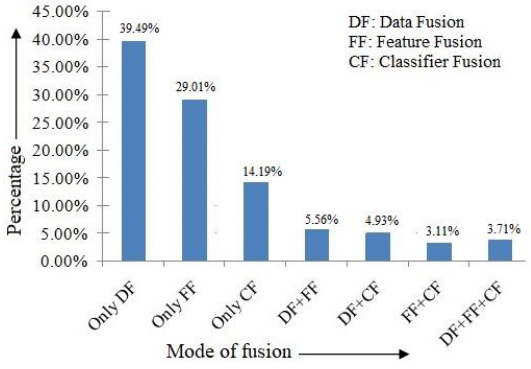

Fig. 12: Multiplicity in the criteria of fusion in HAR

As described in Fig. 13, research in the domain of HAR that embeds fusion from any of the perspectives has been increasing on a regular basis from the last two decades. It is still on the rise to gain accuracy and efficiency in the process of action recognition via several enhancements from the technical front. Apart from implementation for the purpose of HAR, the criteria of fusion have been embedded in many other applications in the domain of computer vision. Both data fusion and feature fusion were embedded in the medical domain for the task of analyzing the data (S.P.Yadav and S.Yadav 2020; Xia et al. 2020). Data fusion lead to decrease in the cost of data transfer

due to decrease in bandwidth required for data transmission and feature fusion enabled to gain effectiveness in the process of treatment by fetching multiple features from medical images. Apart from development within the domain of HAR, research in the domain is proceeding towards the task of predicting the human activities that would be executed (Fortino et al. 2021). I.E.Jaramillo et al. (2023) designed a Human Activity Prediction system based on Bi-LSTM classifier method. It processed the forecasted data and achieved an accuracy of 97.96% in the task of predicting the activity.

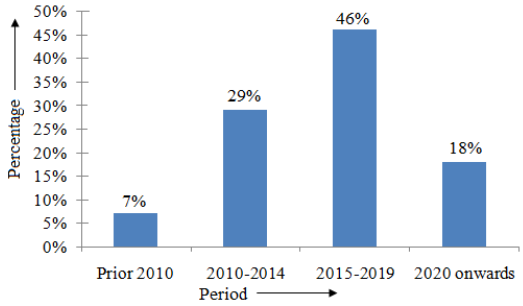

Fig. 13: Research embedding fusion in the domain of HAR

3. State of the Art

Although single-subjected atomic activities are recognized effectively as well as efficiently, there still exist open issues that are required to be tackled by the existing HAR systems. While basic activities such as walking and running are simple to recognize, there are many complications in the real life scenario that are required to be tackled by the action recognition systems. Modeling of composite activities such as culinary preparations or medical interventions upon the patient and their subsequent recognition, pose a more intricate challenge. Apart from nature of activity, complexity also gets embedded in the system due to multiplicity in the number of personnel enrolled in the execution. Both due to complexity in the nature of data as well as due to lack of associated datasets, multi-person and group activities recognition systems are required to be developed. Apart from actions enrolled, contextual information such as location, time, surrounding environment and repetitiveness of any action plays a vital role to interpret the associated activity correctly. To gain accuracy in the action recognition system, the contextual information is required to be taken into consideration. Moreover, many-a-times, multiple activities get executed simultaneously. For example, a group of family members might be watching a movie and having group chat while having dinner. Multiplicity in the number of activities being executed in a simultaneous fashion is also required to be tackled. Action recognition datasets contain normal activities, though the ones containing rare activities are not available. Datasets available do not contain violent acts such as fighting or non-judicial acts like cheating and executing theft or medical abnormalities like heart attack or epileptic fit. Due to the lack of training data, systems to recognize them have not been much worked upon. In the datasets available for developing the systems, inter-class similarity and intra-class variance are additional challenges that are required to be tackled by the developers.

4. Conclusion

This paper focuses on generating a survey of research work aimed at gaining accuracy as well as efficiency in the process of Human Activity Recognition by embedding fusion and thus multiplicity in the process from several perspectives. As the performance of a system is boosted by embedding diversity in the system, hence heterogeneity and thus multiplicity is preferred to be embedded in the system. In the process of HAR, multiplicity could be embedded majorly via three modes, namely, data fusion, feature fusion and classifier fusion. Data fusion embeds multiplicity in the initial phase of the process of HAR. It aims at fetching data from multiple sources, which may be homogeneous or heterogeneous. Data gets captured from multiple views and thus more information is gained to recognize the activity being executed. Feature fusion gains multiplicity at the intermediate phase. It fetches multiple features from raw data to derive meaningful information in order to recognize the associated actions and finally conclude the activity being executed. As there are many classifiers that could analyze the fetched features to recognize the action, hence multiplicity could be embedded into the final phase of the process, namely classification as well. Multiple classifiers could be enrolled, and the result generated by all classifiers could be combined via a particular criterion to generate the final result. As HAR is associated with domains such as medical and security that demand high accuracy as well as efficiency, hence embedding fusion in the process is a necessity to tackle the complexity in the scenario. Accuracy as high as 99.8% via data fusion, 99.4% via feature fusion and 99.13% via classifier fusion has been achieved in the process of HAR. Apart from fusion at an individual level, both accuracy as well as

efficiency could also be gained by embedding fusion via multiple criteria simultaneously in the HAR system.

References

[1] Abdelgawad, A., & Bayoumi, M.: Resource-Aware Data Fusion Algorithms for Wireless Sensor Networks 118. *Springer US*., 17-35 (2012). https://doi.org/10.1007/978-1-4614-1350-9

[2] Abid, M. H., Nahid, A.-A., Islam, Md. R., & Parvez Mahmud, M. A.: Human Activity Recognition Based on Wavelet-Based Features along with Feature Prioritization. *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, 933–939 (2021). https://doi.org/10.1109/ICCCA52192. 2021.9666294

[3] Almaslukh, B., Artoli, A. & Al-Muhtadi J.: A Robust Deep Learning Approach for Position-Independent Smartphone-Based Human Activity Recognition. *Sensors*, *18*(11), 3726 (2018).

[4] Amrita, Joshi, S., Kumar, R., Dwivedi, A., Rai, V., & Chauhan, S. S.: Water wave optimized nonsubsampled shearlet transformation technique for multimodal medical image fusion. *Concurrency and Computation: Practice and Experience*, *35*(7), e7591(2023). https://doi.org/10.1002/cpe.7591

[5] AYDIN, I.: Fuzzy Integral and Cuckoo Search Based Classifier Fusion for Human Action Recognition. *Advances in Electrical and Computer Engineering*, *18*(1), 3–10 (2018). https://doi.org/10.4316/AECE.2018.01001

[6] Bagheri, M. A., Hu, G., Gao, Q., & Escalera, S.: A Framework of Multi-classifier Fusion for Human Action Recognition. *2014 22nd International Conference on Pattern Recognition*, 1260–1265 (2014). https://doi.org/10.1109/ICPR.2014.226

[7] Capela, N. A., Lemaire, E. D., & Baddour, N.: Feature Selection for Wearable Smartphone-Based Human Activity Recognition with Able bodied, Elderly, and Stroke Patients. *PLOS ONE*, *10*(4), e0124414 (2015). https://doi.org/10.1371/journal.pone.0124414

[8] Channi, H. K., Sandhu, R., Faiz, M., & Islam, S. M. (2023, August). Multi-Criteria Decision-Making Approach for Laptop Selection: A Case Study. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5). IEEE.

[9] Chaturvedi, Pooja, A. K. Daniel, and Vipul Narayan. "A Novel Heuristic for Maximizing Lifetime of Target Coverage in Wireless Sensor Networks." Advanced Wireless Communication and Sensor Networks. Chapman and Hall/CRC 227-242.

[10] Chen, C., Jafari, R., & Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications*, *76*(3), 4405–4425 (2017). https://doi.org/10.1007/s11042-015-3177-1

[11] Chen, J., Sun, Y., & Sun, S.: Improving Human Activity Recognition Performance by Data Fusion and Feature Engineering. *Sensors*, *21*(3), 692 (2021). https://doi.org/10.3390/s21030692

[12] Chen, Z., Zhu, Q., Soh, Y. C., & Zhang, L.: Robust Human Activity Recognition Using Smartphone Sensors via CT-PCA and Online SVM. *IEEE Transactions on Industrial Informatics*, *13*(6), 3070–3080 (2017). https://doi.org/10.1109/TII.2017.2712746

[13] Chernbumroong, S., Cang, S., Atkins, A., & Yu, H.: Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications*, *40*(5), 1662–1674 (2013). https://doi.org/10.1016/j.eswa.2012.09.004

[14] Chetty, G., White, M., Singh, M., & Mishra, A.: Multimodal activity recognition based on automatic feature discovery. *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, 632–637 (2014). https://doi.org/10.1109/IndiaCom.2014.6828039

[15] Chetty, G., & White, M.: Body sensor networks for human activity recognition. *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, 660–665 (2016). https://doi.org/10.1109/SPIN.2016.7566779

[16] Chung, S., Lim, J., Noh, K. J., Kim, G., & Jeong, H. : Sensor Data Acquisition and Multimodal Sensor Fusion for Human Activity Recognition Using Deep Learning. *Sensors*, *19*(7), 1716 (2019). https://doi.org/ 10.3390/s19071716

[17] Cordell, K. D., Rao, H., & Lyons, J.: Authentic Assessments: a method to detect anomalies in assessment response patterns via neural network. *Health Services and Outcomes Research Methodology*, *21*, 439-458 (2021). https://10.1007/s10742-021-00245-9

[18] Dalal, N., Triggs, B., & Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance. *Computer Vision – ECCV 2006*, 428–441 (2006). https://doi.org/10.1007/11744047_33

[19] Dietterich, T. G.: Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15 (2000). https://doi.org/10.1007/3-540-45014-9_1

[20] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., & Beghdadi, A.: A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*, *51*(2), 690–712 (2021). https://doi.org/10.1007/s10489-020-01823-z

[21] Faiz, M., & Daniel, A. K. (2023). A hybrid WSN based two-stage model for data collection and forecasting water consumption in metropolitan areas. International Journal of Nanotechnology, 20(5-10), 851-879.

[22] Faiz, M., Sandhu, R., Akbar, M., Shaikh, A. A., Bhasin, C., & Fatima, N. (2023). Machine Learning Techniques in Wireless Sensor Networks: Algorithms, Strategies, and Applications. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 685-694.

[23] Fortino, G., Guzzo, A., Ianni, M., Leotta, F., & Mecella, M.: Predicting activities of daily living via temporal point processes: Approaches and experimental results. *Computers & Electrical Engineering*, *96*, 107567 (2021). https://doi.org/10.1016/j.compeleceng.2021.107567

[24] Freund, Y., & Schapire, R. E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139 (1997). https://doi.org/10.1006/ jcss.1997.1504

[25] Gao, Z., Zhang, H., Xu, G. P., Xue, Y. B., & Hauptmann, A. G.: Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. *Signal Processing*, *112*, 83–97 (2015). https://doi.org/10.1016/j.sigpro.2014.08.034

[26] Ghorbel, E., Boutteau, R., Boonaert, J., Savatier, X., & Lecoeuche, S.: Kinematic Spline Curves: A temporal invariant descriptor for fast action recognition. *Image and Vision Computing*, *77*, 60–71 (2018). https://doi.org/10.1016/j.imavis.2018.06.004

[27] Goyani, M., & Patel, N.: Multi-Level Haar Wavelet based Facial Expression Recognition using Logistic Regression. *Indian Journal of Science and Technology*, *10*(9), 1–9 (2017). https://doi.org/10.17485/ijst/2017/ v10i9/108944

[28] Gravina, R., Alinia, P., Ghasemzadeh, H., & Fortino, G.: Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, *35*, 68–80 (2017). https://doi.org/10.1016/j.inffus.2016.09.005

[29] Guan, Y., & Plötz, T.: 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). https://doi.org/10.1145/3090076

[30] Gumaei, A., Hassan, M. M., Alelaiwi, A., & Alsalman, H.: A Hybrid Deep Learning Model for Human Activity Recognition Using Multimodal Body Sensing Data. *IEEE Access*, *7*, 99152–99160 (2019). https://doi.org/10.1109/ACCESS.2019.2927134

[31] Holte, M. B., Moeslund, T. B., Nikolaidis, N., & Pitas, I.: 3D Human Action Recognition for Multi-view Camera Systems. *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 342–349 (2011). https://doi.org/10.1109/3DIMPVT.2011.50

[32] Huang, Y. S., & Suen, C. Y.: A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(1), 90–94 (1995). https://doi.org/10.1109/34.368145

[33] Hussain, S. ul, & Triggs, B.: Feature Sets and Dimensionality Reduction for Visual Object Detection. *Proceedings of the British Machine Vision Conference 2010*, 112.1-112.10 (2010). https://doi.org/10.5244/C.24.112

[34] Hutchinson, M., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Byun, C., Houle, M., Hubbell, M., Jones, M., Kepner, J., Kirby, A., Michaleas, P., Milechin, L., Mullen, J., Prout, A., Rosa, A., Reuther, A., Yee, C., & Gadepally, V.: Accuracy and Performance Comparison of Video Action Recognition Approaches. *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–8 (2020). https://doi.org/10.1109/HPEC43674.2020.9286249

[35] Ihianle, I. K., Nwajana, A. O., Ebenuwa, S. H., Otuka, R. I., Owa, K., & Orisatoki, M. O.: A Deep Learning Approach for Human Activities Recognition From Multimodal Sensing Devices. *IEEE Access*, *8*, 179028–179038 (2020). https://doi.org/10.1109/ACCESS.2020.3027979

[36] Ijjina, E. P., & Chalavadi, K. M.: Human action recognition in RGB-D videos using motion sequence information and deep learning. *Pattern Recognition*, *72*, 504–516 (2017). https://doi.org/10.1016/j.patcog.2017.07.013

[37] Islam, Md. M., Nooruddin, S., Karray, F., & Muhammad, G.: Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. *Information Fusion*, *94*, 17–31 (2023).

https://doi.org/10.1016/j.inffus.2023.01.015

[38] Islam, S., Qasim, T., Yasir, M., Bhatti, N., Mahmood, H., & Zia, M.: Single- and two-person action recognition based on silhouette shape and optical point descriptors. *Signal, Image and Video Processing*, *12*(5), 853–860 (2018). https://doi.org/10.1007/s11760-017-1228-y

[39] Indhumathi C., Murugan V., & Muthulakshmi G.: Spatio-Temporal Deep Feature Fusion for Human Action Recognition. *International Journal of Computer Vision and Image Processing*, *12*(1), 1–13 (2022). https://doi.org/10.4018/IJCVIP.296584

[40] Jain, Y., Sharma, A. K., Velmurugan, R., & Banerjee, B.: PoseCVAE: Anomalous Human Activity Detection. *2020 25th International Conference on Pattern Recognition (ICPR)*, 2927–2934 (2021). https://doi.org/10.1109/ICPR48806.2021.9412132

[41] Jaouedi, N., Boujnah, N., & Bouhlel, M. S.: A new hybrid deep learning model for human action recognition. *Journal of King Saud University - Computer and Information Sciences*, *32*(4), 447–453 (2020). https://doi.org/10.1016/j.jksuci.2019.09.004

[42] Jaramillo, I. E., Chola, C., Jeong, J.-G., Oh, J.-H., Jung, H., Lee, J.-H., Lee, W. H., & Kim, T.-S.: Human Activity Prediction Based on Forecasted IMU Activity Signals by Sequence-to-Sequence Deep Neural Networks. *Sensors*, *23*(14), 6491 (2023). https://doi.org/10.3390/s23146491

[43] Joshi, S., Kumar, R., & Dwivedi, A.: Hybrid DSSCS and convolutional neural network for peripheral blood cell recognition system. *IET Image Processing*, *14*(17), 4450-4460 (2020). https://doi.org/10.1049/ iet-ipr.2020.0370

[44] Khan, I. U., Afzal, S., & Lee, J. W.: Human Activity Recognition via Hybrid Deep Learning Based Model. *Sensors*, *22*(1), 323 (2022). https://doi.org/10.3390/s22010323

[45] Kuehne, H., Arslan, A., & Serre, T.: The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 780–787 (2014). https://doi.org/10.1109/CVPR.2014.105

[46] Kumar, Vaibhav, et al. "A Machine Learning Approach For Predicting Onset And Progression"“Towards Early Detection Of Chronic Diseases “." Journal of Pharmaceutical Negative Results (2022): 6195-6202.

[47] Kumar, R., Qamar, I., Virdi, J. S., & Krishnan, N. C.: Multi-label Learning for Activity Recognition. *2015 International Conference on Intelligent Environments*, 152–155 (2015). https://doi.org/10.1109/IE.2015.32

[48] Lara, Ó. D., Pérez, A. J., Labrador, M. A., & Posada, J. D.: Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing*, *8*(5), 717–729 (2012). https://doi.org/10.1016/j.pmcj.2011.06.004

[49] Li, J., Fong, S., Wong, R. K., Millham, R., & Wong, K. K. L.: Elitist Binary Wolf Search Algorithm for Heuristic Feature Selection in High-Dimensional Bioinformatics Datasets. *Scientific Reports*, *7*(1), 4354 (2017). https://doi.org/10.1038/s41598-017-04037-5

[50] Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R. A., & Burd, R. S.: Concurrent Activity Recognition with Multimodal CNN-LSTM Structure. arXiv: 1702.01638v1 (2017). https://doi.org/10.48550/arXiv.1702.01638

[51] Li, Y., Yang, G., Su, Z., Li, S., & Wang, Y.: Human activity recognition based on multienvironment sensor data. *Information Fusion*, *91*, 47–63 (2023). https://doi.org/10.1016/j.inffus.2022.10.015

[52] Liu, A.-A., Xu, N., Nie, W.-Z., Su, Y.-T., & Zhang, Y.-D.: Multi-Domain and Multi-Task Learning for Human Action Recognition. *IEEE Transactions on Image Processing*, *28*(2), 853–867 (2019). https://doi.org/10.1109/ TIP.2018.2872879

[53] Liu, L., Shao, L., Li, X., & Lu, K.: Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach. *IEEE Transactions on Cybernetics*, *46*(1), 158–170 (2016). https://doi.org/10.1109/ TCYB.2015.2399172

[54] Liu, T., Chen, Z., Liu, H., Zhang, Z., & Chen, Y.: Multi-modal hand gesture designing in multi-screen touchable teaching system for human-computer interaction. *Proceedings of the 2nd International Conference on Advances in Image Processing*, 198–202 (2018). https://doi.org/10.1145/3239576.3239619

[55] Mall, P. K., et al. "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. Healthcare Analytics, 4, 100216." (2023).

[56] Mall, Pawan Kumar, et al. "Rank Based Two Stage Semi-Supervised Deep Learning Model for X-Ray Images Classification: AN APPROACH TOWARD TAGGING UNLABELED MEDICAL DATASET." Journal of Scientific & Industrial Research (JSIR) 82.08 (2023): 818-830.

[57] Mall, Pawan Kumar, et al. "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities." Healthcare Analytics (2023): 100216.

[58] Ma, S., Zhang, J., Sclaroff, S., Ikizler-Cinbis, N., & Sigal, L.: Space-Time Tree Ensemble for Action Recognition and Localization. *International Journal of Computer Vision*, *126*(2–4), 314–332 (2018). https://doi.org/10.1007/s11263-016-0980-8

[59] Morshed, M. G., Sultana, T., Alam, A., & Lee, Y.-K.: Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities. *Sensors*, *23*(4), 2182 (2023). https://doi.org/10.3390/s23042182

[60] Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., & Dürichen, R. (2017). CNN-based sensor fusion techniques for multimodal human activity recognition. *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 158–165. https://doi.org/10.1145/3123021.3123046

[61] Muralikrishna, S. N., Muniyal, B., Acharya, U. D., & Holla, R.: Enhanced Human Action Recognition Using Fusion of Skeletal Joint Dynamics and Structural Features. *Journal of Robotics*, *2020*, 1–16 (2020). https://doi.org/10.1155/2020/3096858

[62] Najjar, N., & Gupta, S.: Better-than-the-best fusion algorithm with application in human activity recognition. *SPIE Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2015*, 949805-949810 (2015). https://doi.org/10.1117/12.2177123

[63] Narayan, Vipul, et al. "A Comprehensive Review of Various Approach for Medical Image Segmentation and Disease Prediction." Wireless Personal Communications 132.3 (2023): 1819-1848.

[64] Narayan, Vipul, et al. "Severity of Lumpy Disease detection based on Deep Learning Technique." 2023 International Conference on Disruptive Technologies (ICDT). IEEE, 202

[65] Narayan, Vipul, et al. "7 Extracting business methodology: using artificial intelligence-based method." Semantic Intelligent Computing and Applications 16 (2023): 123.

[66] Nguyen, D. T., Li, W., & Ogunbona, P. O.: Local intensity distribution descriptor for object detection. *Electronics Letters*, *47*(5), 321 (2011). https://doi.org/10.1049/el.2010.3256

[67] Nguyen, D. T., Ogunbona, P., & Li, W. Human detection with contour-based local motion binary patterns. *2011 18th IEEE International Conference on Image Processing*, 3609–3612 (2011). https://doi.org/10.1109/ICIP.2011.6116498

[68] Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-garadi, M. A.: Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*, 147–170 (2019). https://doi.org/10.1016/j.inffus.2018.06.002

[69] Ordóñez, F., & Roggen, D.: Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, *16*(1), 115 (2016). https://doi.org/10.3390/s16010115

[70] Pai. M. M. M., Ganiga, R., Pai, R. M. & Sinha, R. K.: Standard Electronic Health Record (EHR) framework for Indian healthcare system. *Health Services and Outcomes Research Methodology*, *21*, 339-362 (2021). https://doi.org/10.1007/s10742-020-00238-0

[71] Pareek, P., & Thakkar, A.: RGB-D based human action recognition using evolutionary self-adaptive extreme learning machine with knowledge-based control parameters. *Journal of Ambient Intelligence and Humanized Computing*, *14*(2), 939–957 (2023). https://doi.org/10.1007/s12652-021-03348-w

[72] Patel, C. I., Garg, S., Zaveri, T., Banerjee, A., & Patel, R.: Human action recognition using fusion of features for unconstrained video sequences. *Computers & Electrical Engineering*, *70*, 284–301 (2018). https://doi.org/10.1016/j.compeleceng.2016.06.004

[73] Patel, C. I., Labana, D., Pandya, S., Modi, K., Ghayvat, H., & Awais, M.: Histogram of Oriented Gradient-Based Fusion of Features for Human Action Recognition in Action Video Sequences. *Sensors*, *20*(24), 7299 (2020). https://doi.org/10.3390/s20247299

[74] Peng, L., Chen, L., Wu, X., Guo, H., & Chen, G.: Hierarchical Complex Activity Representation and Recognition Using Topic Model and Classifier Level Fusion. *IEEE Transactions on Biomedical Engineering*, *64*(6), 1369–1379 (2017). https://doi.org/10.1109/TBME.2016.2604856

[75] Ponti Jr., M. P.: Combining Classifiers: From the Creation of Ensembles to the Decision Fusion. *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, 1–10 (2011). https://doi.org/10.1109/ SIBGRAPI-T.2011.9

[76] Prakash Yadav, S., & Yadav, S.: Fusion of

Medical Images in Wavelet Domain: A Hybrid Implementation. *Computer Modeling in Engineering & Sciences*, *122*(1), 303–321 (2020). https://doi.org/10.32604/cmes. 2020.08459

[77] Prakash Yadav, S., & Yadav, S.: Image Fusion using Hybrid Methods in Multimodality Medical Images. *Medical & Biological Engineering & Computing*, 58, 669-687 (2020). https://doi.org/10.1007/s11517-020-02136-6

[78] Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R., & Fortino, G.: Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion*, *80*, 241–265 (2022). https://doi.org/10.1016/j.inffus.2021.11.006

[79] Rai, V., Gupta, G., Joshi, S., Kumar, R., & Dwivedi, A.: LSTM-based adaptive whale optimization model for classification of fused multimodality medical image. *Signal, Image and Video Processing*, *17*(5), 2241–2250 (2023). https://doi.org/10.1007/s11760-022-02439-1

[80] Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., & Sebe, N.: Abnormal event detection in videos using generative adversarial nets. *2017 IEEE International Conference on Image Processing (ICIP)*, 1577–1581 (2017). https://doi.org/10.1109/ICIP.2017.8296547

[81] Roche, J., De-Silva, V., Hook, J., Moencks, M., & Kondoz, A.: A Multimodal Data Processing System for LiDAR-Based Human Activity Recognition. *IEEE Transactions on Cybernetics*, *52*(10), 10027–10040 (2022). https://doi.org/10.1109/TCYB.2021.3085489

[82] Rogova, G.: Combining the results of several neural network classifiers. *Neural Networks*, *7*(5), 777–781 (1994). https://doi.org/10.1016/0893-6080(94)90099-X

[83] Saxena, Aditya, et al. "Comparative Analysis Of AI Regression And Classification Models For Predicting House Damages İn Nepal: Proposed Architectures And Techniques." Journal of Pharmaceutical Negative Results (2022): 6203-6215.

[84] Schrader, L., Vargas Toro, A., Konietzny, S., Rüping, S., Schäpers, B., Steinböck, M., Krewer, C., Müller, F., Güttler, J., & Bock, T.: Advanced Sensing and Human Activity Recognition in Early Intervention and Rehabilitation of Elderly People. *Journal of Population Ageing*, *13*(2), 139–165 (2020). https://doi.org/10.1007/s12062-020-09260-z

[85] Sebbak, F., & Benhammadi, F.: Majority-consensus fusion approach for elderly IoT-based healthcare applications. *Annals of Telecommunications*, *72*(3–4), 157–171 (2017). https://doi.org/10.1007/s12243-016-0550-7

[86] Sharaf, A., Torki, M., Hussein, M. E., & El-Saban, M.: Real-Time Multi-scale Action Detection from 3D Skeleton Data. *2015 IEEE Winter Conference on Applications of Computer Vision*, 998–1005 (2015). https://doi.org/10.1109/WACV.2015.138

[87] Sharif, M., Khan, M. A., Akram, T., Javed, M. Y., Saba, T., & Rehman, A.: A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. *EURASIP Journal on Image and Video Processing*, *2017*(1), 89 (2017). https://doi.org/10.1186/s13640-017-0236-8

[88] Shen, C., Chen, Y., & Yang, G.: On motion-sensor behavior analysis for human-activity recognition via smartphones. *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 1–6 (2016). https://doi.org/10.1109/ISBA.2016.7477231

[89] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J.: Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Transactions on Image Processing*, *27*(7), 3459–3471 (2018). https://doi.org/10.1109/TIP.2018.2818328

[90] Srivastava, N., Mansimov, E., & Salakhutdinov, R.: Unsupervised Learning of Video Representations using LSTMs. *International Conference on Machine Learning*, 843-852. PMLR (2015).

[91] Tao, W., Chen, H., Moniruzzaman, M., Leu, M. C., Yi, Z., & Qin, R.: Attention-Based Sensor Fusion for Human Activity Recognition Using IMU Signals. *Engineering Applications of Artificial Intelligence* (2021). https://doi.org/10.48550/arXiv.2112.11224

[92] Thakur, D., & Biswas, S.: Feature fusion using deep learning for smartphone based human activity recognition. *International Journal of Information Technology*, *13*(4), 1615–1624 (2021). https://doi.org/10.1007/s41870-021-00719-6

[93] Uddin, M., & Lee, Y.-K.: Feature Fusion of Deep Spatial Features and Handcrafted Spatiotemporal Features for Human Action Recognition. *Sensors*, *19*(7), 1599 (2019). https://doi.org/10.3390/s19071599

[94] Valdovinos, R. M., & Sánchez, J. S.: Combining Multiple Classifiers with Dynamic Weighted

Voting. *International Conference on Hybrid Artificial Intelligence Systems*, 510–516 (2009). https://doi.org/10.1007/ 978-3-642-02319-4_61

[95] Vemulapalli, R., & Chellappa, R.: Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4471–4479 (2016). https://doi.org/ 10.1109/CVPR.2016.484

[96] Vidya, B., & Sasikumar, P.: Wearable multi-sensor data fusion approach for human activity recognition using machine learning algorithms. *Sensors and Actuators A: Physical*, *341*, 113557 (2022). https://doi.org/10.1016/j.sna.2022.113557

[97] Wang, D., Ouyang, W., Li, W., & Xu, D.: Dividing and Aggregating Network for Multi-view Action Recognition. *European Conference on Computer Vision (ECCV)*. 457–473 (2018). https://doi.org/10.1007/ 978-3-030-01240-3_28

[98] Wang, L., Ding, Z., Tao, Z., Liu, Y., & Fu, Y.: Generative Multi-View Human Action Recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6211–6220 (2019). https://doi.org/10.1109/ ICCV.2019.00631

[99] Ward, J. A., Lukowicz, P., Troster, G., & Starner, T. E.: Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1553–1567 (2006). https://doi.org/10.1109/TPAMI.2006.197

[100] Woźniak, M., Graña, M., & Corchado, E.: A survey of multiple classifier systems as hybrid systems. *Information Fusion*, *16*, 3–17 (2014). https://doi.org/10.1016/j.inffus.2013.04.006

[101] Wu, H., Siegel, M., Stiefelhagen, R., & Yang, J.: Sensor Fusion Using Dempster-Shafer Theory. *IEEE Instrumentation and Measurement Technology Conference* (2002).

[102] Wu, Z., Li, X., Zhao, X., & Liu, Y.: Hybrid generative-discriminative recognition of human action in 3D joint space. *Proceedings of the 20th ACM International Conference on Multimedia*, 1081–1084 (2012). https://doi.org/ 10.1145/2393347.2396388

[103] Xia, K., Huang, J., & Wang, H.: LSTM-CNN Architecture for Human Activity Recognition. *IEEE Access*, *8*, 56855–56866 (2020). https://doi.org/10.1109/ACCESS.2020.2982225

[104] Xiao, Q., & Song, R.: Action recognition based on hierarchical dynamic Bayesian network. *Multimedia Tools and Applications*, *77*(6), 6955–6968 (2018). https://doi.org/10.1007/s11042-017-4614-0

[105] Xu, H., Liu, J., Hu, H., & Zhang, Y.: Wearable Sensor-Based Human Activity Recognition Method with Multi-Features Extracted from Hilbert-Huang Transform. *Sensors*, *16*(12), 2048 (2016). https://doi.org/ 10.3390/s16122048

[106] Zhang, C., Cao, K., Lu, L., & Deng, T.: A multi-scale feature extraction fusion model for human activity recognition. *Scientific Reports*, *12*(1), 20620 (2022). https://doi.org/10.1038/s41598-022-24887-y

[107] Zhang, Y., Li, X., & Marsic, I.: Multi-Label Activity Recognition using Activity-specific Features and Activity Correlations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14620–14630 (2021). https://doi.org/10.1109/CVPR46437.2021.01439

[108] Zhu, J., San-Segundo, R., & Pardo, J. M.: Feature extraction for robust physical activity recognition. *Human-Centric Computing and Information Sciences*, *7*(1), 16 (2017). https://doi.org/10.1186/s13673-017-0097-2