ACM-ASC Pre-Internship 2023

Background Study

Literature Review Summary

Coding- Familiarisation

Title : Stacked Ensemble Model for Human Activity Recognition

Group No: 04

Team Lead : Girish S

Faculty Mentor: Dr. Namitha K

Team Members

Girish S	**AM.EN.U4AIE22044**
R S Harish Kumar	**AM.EN.U4AIE22042**
Anuvindh MP	**AM.EN.U4AIE22010**
Harishankar Binu Nair	**AM.EN.U4AIE22023**

Tasks to do

1. **Background study:**
    - **Started familiarizing with the existing implementations on HAR by opencv.** https://learnopencv.com/introduction-to-video-classification-and-human-activity-recognition/
    - **Searched for the datasets generally used for human activity recognition problems in Kaggle** https://www.kaggle.com/datasets/sharjeelmazhar/human-activity-recognition-video-dataset/data
    - **Started watching video tutorials on HAR.** https://www.geeksforgeeks.org/human-activity-recognition-with-opencv/
    - **Write a brief paragraph outlining my approach towards achieving the project deliverables and implementing the solution.**
      The Human Activity Recognition (HAR) is a pattern recognition task that learns to identify human physical activities recorded by different sensor modalities. The application areas include human behavior analysis, ambient assisted living, surveillance-based security,

gesture recognition, and context- aware computing. The importance of Human Activity Recognition (HAR) lies in its ability to monitor and track human activities in real-time, which can provide valuable insights and support for several applications. For instance, in healthcare, HAR system can be employed to monitor the activities of elderly or disabled individuals and provide assistance when needed.
In this project, we introduce an efficient system for Human Activity Recognition by leveraging publicly available datasets containing class-labeled activities like walking, running, sitting, standing, and more. The data undergoes preprocessing, and features are extracted using a proposed model for human detection. Subsequently, these identified features are fed into a deep learning model, trained to accurately recognize various activities, including but not limited to walking, running, sitting, and standing.

Girish S	1. An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier
	2. Human Activity Recognition for Office Surveillance
Anuvindh MP	1. Automated Daily Human Activity Recognition for Video Surveillance Using Neural Network
	2. An Efficient Human Activity Recognition Using Hybrid Features and Transformer Model
R S Harish Kumar	A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network
	Human Activity Recognition System from Different Poses with CNN
Harishankar Binu Nair	1. A Survey on Human Activity Recognition and Classification
	2. Human Activity Recognition in Videos

Literature Review :

1. **An Approach of Understanding Human Activity Recognition and Detection for Video Surveillance using HOG Descriptor and SVM Classifier**

   IEEE/2017 | doi: 10.1109/CTCEEC.2017.8455046

   - **Problem addressed:** This work mainly focuses on multiple human detection and activity recognition using HOG (Histogram Oriented Gradient Descriptor) descriptor and SVM (Support Vector Machine) classifier. This paper have taken in a various techniques from previous works and have improved with the HOG descriptor and SVM classifier, such as background subtraction for detecting moving humans, 2D Median filtering for noise control, RGB2GREY and binary image formation for computation reduction.

   - **Challenges:** Detection and recognition of moving is not an easy task as continuous deformation of objects takes place during movement. Any moving object has several attributes in temporal and spatial spaces. In spatial space objects vary in size whereas in temporal space it varies in moving speed. Previously proposed ideas have two main approaches of detecting and tracking humans: frame difference method and background modeling method. Frame difference method is most suitable for no change in background and when there is a relatively static situation. Background modeling method is based on Gaussian mixture model (GMM), Graph cut method. GMM and Graph cut methods are more complex and a large amount of calculation is involved.
     In this method using Background Subtraction methods, This approach will work well when all the foreground pixels are moving and all the background pixels are static in nature. Background subtraction also fails when there is data occlusion, i.e., when 2 people cross each other or a part of a person is hidden by objects. Taking computational expense into account, various preprocessing techniques are applied on the data before providing to HOG descriptor and SVM classifier such as RGB2GREY, Median Filtering to reduce noise, Binary Image creation. Resulting in the data just being a 2 dimensional matrix of 1s and 0s, before being a 3 dimensional matrix with values varying from 0 to 255.

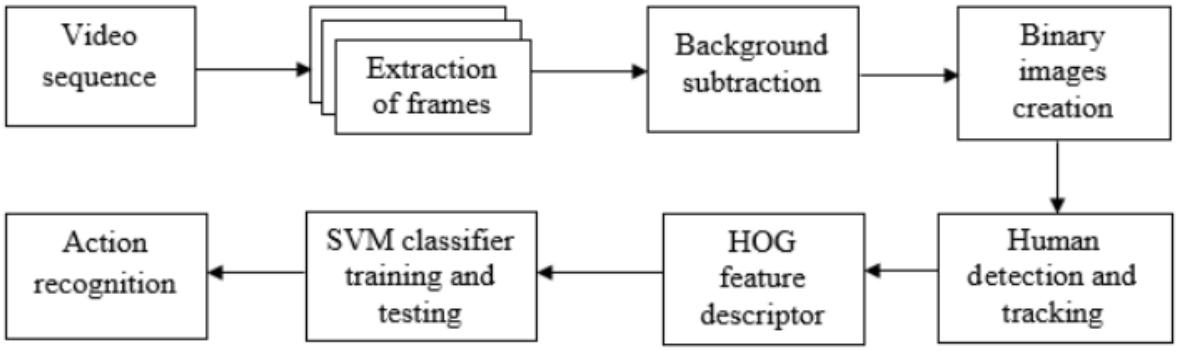


Fig. 1. *Block diagram of approach used in Human Activity Recognition*

- **Summary:**
  In this paper, the input video is read from UT-interaction dataset consisting of continuous activity execution with 5 human-human interactions like shake-hands, hug, push, kick and punch, and key frames are extracted, which are given to background subtraction providing moving humans on the foreground. Then preprocessed by various methods like RGB2GREY and 2D Median Filtering for noise reduction and Binary Image creation for human detection. “Once after noise removal is done the gray scale images will be converted to binary images of 0s and 1s, where binary 1 is used for representing the human region which is filled with white color and apart from moving human region binary 0 is used which represents absence of humans.”
  The detected binary images are given to HOG feature extractor for detecting and tracking humans.
  The extracted features are given to the SVM classifier for training and saving the trained model.
- **Dataset used:**
  UT-interaction dataset consists of continuous activity execution which contains five human-human interactions like shake-hands, hug, push, kick and punch and the lengths of the videos are around 30 seconds.
  Their Own video dataset is recorded in 2 different background settings with static background and two sets of own dataset is considered performing seven different activities like stand, walk, hug, punch, kick, handshake and fallback with duration of 30 seconds and both datasets are considered with static background.
- **Results and Future works:**
  The proposed work gives a solution for human detection and activity recognition. Although many works have been carried out, this proposed work provides excellent results for various kinds of video datasets considered. Human detection using background subtraction for static video gives effective results and with

HOG feature extraction and SVM classifier recognition of human activities provides good recognition results with less minimum number of false detections. Use of UT-interaction and own datasets achieves a higher rate of recognition. The results obtained demonstrate that the method is efficient. Therefore, the proposed technique can be regarded as a better choice for human detection and activity recognition for video surveillance application.
Future work is aimed towards minimizing the false detection like shadow and other reflections of human, it could also incorporate HAR for moving background scenario and recognizing various activity like talking, eating, etc. which can be done by pose estimation of each individual human in the video

2. Human Activity Recognition for Office Surveillance

IEEE/2023 | doi:10.1109/INCET57972.2023.10170132

- **Problem addressed:**
  This paper presents a method for human activity recognition in office surveillance videos using machine learning models including convLSTM,GRCNN and LRCN with three main steps: pre-processing, feature extraction and activity classification.The main targeted activities are walking, sleeping on desk, handshaking, typing, opening or closing door. Experimental results demonstrate the effectiveness of the proposed LRCN approach in accurately recognizing human activities in office surveillance videos with acceptable training and testing accuracy.
- **Challenges:**
  A system capable of inferring the behavior of humans and recognizing or even predicting their activities can have a wide range of applications, from surveillance to more complex functions like an automatic commentary on sporting events like cricket, football, etc. even better when these activities are recognized instantly and automatically. One such environment where human activity recognition plays a vital role is the office environment. People across the world work in different office related jobs and perform a wide range of office activities, but there are some common activities that are similar in most workplaces. In the paper, there has been an attempt to classify some of the common office activities being performed by office employees on a day-to-day basis. Recognizing these common office activities can aid in employee monitoring and detecting unusual behavior. Unusual activities can vary from employees slacking off, an employee falling to the ground due to fatigue, heart-related disease, etc., or even detecting intruders behaving suspiciously, or rapid movement in the form of running by employees indicating some kind of trouble and thereby taking immediate, appropriate actions to tackle the problems

Challenges from video for office surveillance using machine learning (ML) models is the lack of large-scale annotated datasets that cover a diverse range of human activities in office environments
Another challenge is dealing with occlusions, where parts of the body are not visible, leading to incomplete information for activity recognition. The generalizability of ML models is also a major concern as models trained on one dataset may not perform well on different data sets, leading to poor real-world performance. Additionally, the deployment of ML models in real-world scenarios involves challenges related to power consumption, processing speed, and hardware requirements
The entire frame of the video is processed to reduce the error due to occlusion and lighting effects. To achieve generalizability of the model it has trained over large and diverse office video data. This paper deals with the ML models that are comparatively efficient in terms of speed, power and other resources by reducing training time by processing an entire video in a single pass.

- **Summary:** Human activity surveillance video systems are gaining popularity in the field of computer vision due to user demands for security as well as their growing importance in many applications such as elder care, home nursing, and unusual event alarming. Automatic activity recognition is the key to video surveillance.
  This paper presents a method for human activity recognition in office surveillance videos using machine learning models including convLSTM, GRCNN and LRCN with three main steps: pre-processing, feature extraction and activity classification.
  The main targeted activities are walking, sleeping on the desk, handshaking, typing, opening or closing doors.
  Experimental results demonstrate the effectiveness of the proposed LRCN approach in accurately recognizing human activities in office surveillance videos with acceptable training and testing accuracy
- **Dataset used:**
  collecting a suitable office video dataset for human activity recognition research is an important task that requires a considerable amount of effort and resources. While some publicly available datasets exist, they may not cover the full range of activities that occur in typical office environments. One approach to overcome this issue is to collect additional data from online sources. Online platforms such as YouTube and Vimeo contain a vast amount of videos that potentially include office-related activities. However, collecting and annotating videos from online sources requires careful consideration of copyright and privacy issues. Moreover, the quality and consistency of the videos in the dataset must be ensured to avoid bias and ensure the robustness of the recognition model. Therefore, while online video sources can provide a valuable resource for expanding office video datasets, careful curation and ethical considerations are necessary to ensure the reliability and usefulness of the resulting dataset.
  Thus they have sourced videos from online as well as recorded their own data

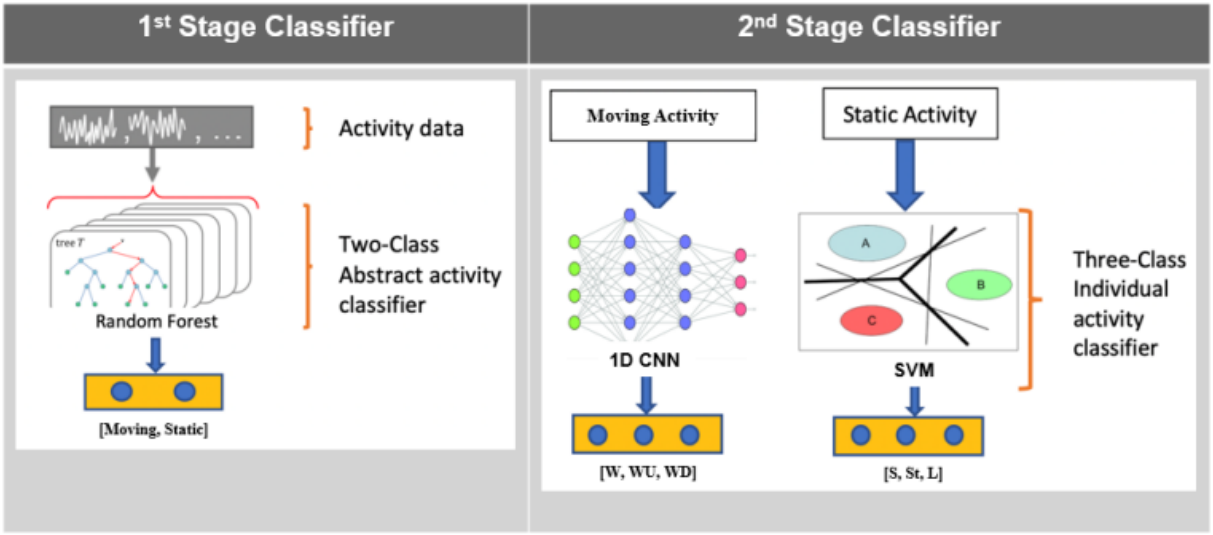
- **Result and Future work:**
  It was discovered that the feature extraction, action representation, and classification steps of processes can be used to construct a human action recognition system. Among the various use cases of human activity recognition, office activity recognition is considered essential for efficient and safe day-to-day working in the office.
  This project looks at some of the common activities that exist in an office and tries to classify them. They have made use of some of the most popular algorithms that are used in the machine learning and artificial intelligence fields, such as the ConvLSTM, which is a LSTM with convolution layers; GRCNNs, which make use of GRU units for feature extraction; a fully connected dense layer for caption generation; and finally, LRCN, which is a combination of CNN and LSTM, where the CNN is used for feature extraction and the LSTM is used for identifying the temporal relations and classifying the activity.LRCN offers several advantages for human activity recognition from video for office surveillance. Firstly, LRCN has a simpler architecture compared to convLSTM and GRCNN, which makes it faster and easier to train. Additionally, LRCN can process variable length inputs, making it more versatile in real-world scenarios.Secondly LRCN is less susceptible to overfitting and it has a smaller number of parameters than both GRCNN and ConvLSTM, which can lead to less overfitting on smaller datasets. This can make LRCN a more effective option for scenarios where limited data is available. In conclusion, the above methods have produced satisfactory accuracy and are therefore promising methods that can be improved upon.The choice of model architecture for human activity recognition depends on the specific requirements and constraints of the application. However, based on the analysis of the performance and advantages/disadvantages of the three models, it can be concluded that LRCN is a suitable choice for HAR in office surveillance scenarios. This work is not only limited to the office environment and can be further expanded to other places, such as government institutions like schools, colleges, hospitals, etc. It can also be used for abnormal activity detection, i.e., surveillance in banks, shops, etc., or even in the monitoring of the elderly in rehabilitation centers or patients in hospitals.

3. A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network

IEEE/2020 | doi:10.1109/AIPR50011.2020.9425332

- **Problem Addressed:**
  This paper presents an adaptive human activity recognition model with a two-stage learning process to recognize human activity recorded using a waist-mounted accelerometer and gyroscope sensor. In the first step, The activity is classified into static and moving, using a Random Forest (RF) binary classifier. In the second step, It

adopts a Support Vector Machine (SVM) to identify individual static activity and 1D Convolutional Neural Network (CNN)- based deep learning model for individual moving activity recognition. This makes the approach more robust and adaptive. The static activity has less frequency variation in features compared to dynamic activity waveforms for CNN to learn. On the other hand, SVM demonstrated superior performance to recognize static activities but performs poorly on moving, complex, and uncertain activity recognition. This method is similarly robust to different motion intensity and can also capture the variation of the same activity effectively. In this hybrid model, the CNN captures local dependencies of activity signals as well as preserves the scale invariance.



- **Challenges:**
  The HAR remains challenging as the sensor data is noisy in nature and the activity signal varies from person to person. To recognize different types of activity with a single classifier is often error-prone.
  One of the main challenges of activity prediction is to generalize the model for different problems, sensors, and activities. Activity signals may vary significantly for different humans, even the same person may do the same activity differently in another time. Similarly, different activities may have similar signal patterns which may confuse the learning process. Other challenges include the computational cost to implement in embedded and portable devices, accurate data annotation, variety of complex daily activities, and ensuring privacy of the subjects. Traditional pattern recognition based HAR requires to extract problem specific features to fit a machine learning model. Deep learning (DL) makes the task easy and adoptable by automatically learning the features. In a DL approach, it also can extract high-level features in deep layer that makes it appropriate for complex activity recognition.
- **Summary:**
  In this research, they have used the UCI-HAR dataset. The recording of 30 subjects having an age range from 19 to 48 years. The dataset consists of the six activity signals of daily living obtained by a waist mounted smartphone following the activity protocol. The data acquisition uses the smartphone's accelerometer and gyroscope.

The tri-axial (x, y, z) data of activities are: walking, walking-upstairs, walking-downstairs, sitting, standing, and laying. After acquisition, the data were sampled at 50Hz and separated into 128 windows with 50% overlapping. There are a total 9 channels of gyroscope and accelerometer: (i) 3 channel body accelerometer, (ii) 3 channel total accelerometer, and (iii) 3 channel body gyroscope. Each channel recorded 128 real valued vectors to depict an activity. The dataset also includes the accurate labels. The sensor data are noisy in nature. Thus, a noise removal filter removed the noise and a low pass Butterworth filter separated the gravitational and body motion components from the acceleration signal. Along with the raw sensor signals, the dataset also includes a very well-engineered set of 561 features calculated from 128 reading of each channel.

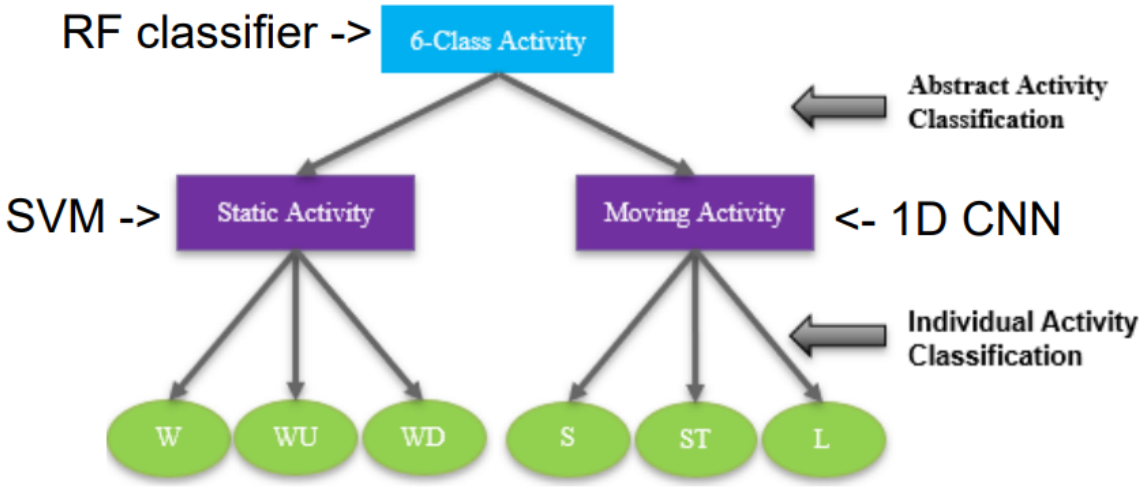


Fig. 3. Illustration of hybrid method for human activity classification. From six total activity we classify the type of activity using a binary classification. Features of each class are fed into separate classifiers to identify individual activity.

This hybrid model is a good choice because the CNN captures the spatial relation between signals and the SVM captures the spatio-temporal relationship. Together it enhances the ability to recognize different activities that have varied signal distributions. This model first identifies the static and moving activity using a Random Forest (RF) binary classifier. The RF combines many decision trees into a single model and uses the average of multiple trees or computes the majority votes to make a prediction in the terminal leaf. In this problem, RF provided better results (100% accuracy) than other binary classifiers thus they selected RF for high-level classification. After identifying the abstract level of static or moving activities, the data is sent according to static activities for SVM and moving activities for the CNN model for appropriate classification.

SVM is a one-vs-one three class classification model with degree 7 ploy svc kernel. The multi-class SVM model is trained with the 561 feature vectors extracted from the

raw activity signals. A deep learning CNN model was developed to recognize the moving activities. The CNN model has three convolution layers, each convolution layer 1D filter size is 5×5 with ReLU activation followed by max-pooling and 20% dropout.

- **Dataset used:** In this research, they have used the UCI-HAR dataset. The recording of 30 subjects having an age range from 19 to 48 years. The dataset consists of the six activity signals of daily living obtained by a waist mounted smartphone following the activity protocol. The data acquisition uses the smartphone's accelerometer and gyroscope. The tri-axial (x, y, z) data of activities are: walking, walking-upstairs, walking-downstairs, sitting, standing, and laying.
- **Results and future works:**
  In this paper, a hybrid method to effectively perform human activity recognition was presented. This method first identifies the abstract activity by using a Random Forest classifier to identify the activities type as static and moving. For static activity's specific recognition it uses a support vector machine model and for moving activities a deep 1D CNN was designed. It achieved an overall accuracy of 97.71% which is comparable to state-of-the-art performance. Future plans are to deploy implementations into low-power integrated circuits to make it well-suited for wearable sensors to identify a wide variety of activities in real-time.

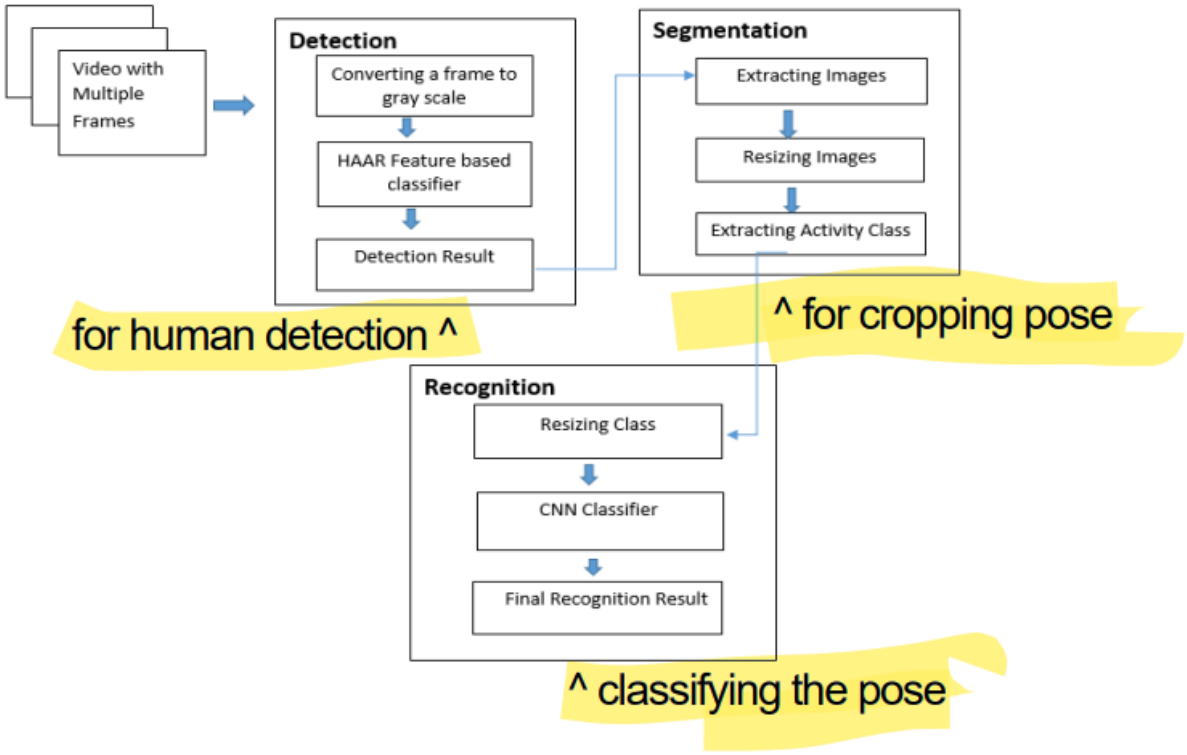
4. Human Activity Recognition System from Different Poses with CNN

IEEE/2020 | doi:10.1109/STI50764.2020.9350508

- **Problem Addressed:**
  In this paper, focused on CCTV videos and camera images to detect human poses using HAAR Feature-based Classifier and recognize the activities of the human using the Convolutional Neural Network (CNN) Classifier. The Human Activity Recognition System was trained using their own collected dataset which is composed of 5648 images. The approach accomplished an efficacious detection accuracy of 99.86% and recognition accuracy of 99.82% with approximately 22 frames/second in 20 epochs.
- **Challenges:**
  The suggested approach is outlined in this section and is divided into two categories: dataset and system design. System design is divided into three categories: human detection / localization, segmentation, and video frame recognition. Complex features and temporal information could be used simultaneously to detect and classify activity classes in order to achieve high precision for variations in environments. However, the computational load has to increase dramatically as more characteristics are measured. When detecting and recognizing the activity class from videos or images, a system that can minimize the computation time and achieve the most accuracy must be built.
  The detection intends to prepare all frames from an input video in real-time because continuous frames may hold human or not that are obliged to develop a proper human activity recognition system. Pose activities may not be recognized in delivered frames due to various reasons, for example, the human may be

hidden or partially hidden by other objects or mostly shaded when they are out of focus. They trained the HAAR Feature-based classifier using only human images to detect the human poses from the input videos or images

- **Summary:**
  The proposed human activity recognition system is separated into three subsections: human detection, segmentation, and recognition of human activity from the videos or images with high precision and lower time complexity. The system architecture of human activity recognition is shown in the below fig



2. Proposed Human Pose Detection And Recognition System Architec-

To detect the human from a video frame or images, the input image needs to be converted into gray-scale and this is done by a function. The final result of the HAAR Feature based classifier is the (X, Y,H, W) where (X, Y) is the origin coordinate of the human and H Height of the human and W is the Width human of the input image.

Human segmentation is essential for so many reasons such as extracting the two or many people in one single frame, removing the background, etc. The human segmentation is done by the result of the HAAR feature-based classifier where we get (X, Y, H, W), (X, Y) as starting coordinate and H height of the human and W is the width of the human. After deriving the human it is resized the image into 64 X 64 resolution.

For training the CNN model 80-20 split was taken. It is a deep CNN classifier with input, convolution, normalization, activation, dense, flatten, dropout layers.

To investigate human activity recognition precision, The CNN Classifier was tested with 1039 human pose images and got an accuracy rate of 99.82% with

more moderate computational time. After 20 epochs the train and validate loss of human activity recognition is 3.56% and 1.07% respectively

- **Dataset used:**
  The experiment was conducted by their own collected dataset and the collection has been done with the laptop camera and CCTV camera and image frames from different videos. They have collected over 5648 different activity images in several kinds of weather and light conditions. The dataset contains 60 human data. The data was collected from indoor and outdoor, rooftop, room, the road in low light and bright light conditions. As a result, the image resolution was not good enough which introduced a challenge to correctly detect and recognize human activities. The dataset is composed of only five different classes of human activity and the quantity of the images is not the same for all classes. The classes are: Walking, Running, Standing, Sitting, and Laying.
- **Result and Future work:**
  A Human activity recognition system was developed with the accuracy of 99.82% where the total no of images is counted. Which will help to track human activity. Since their setup had no GPU and low computing power, they were unable to train the system using RGB images and only used uniform gray images to train and evaluate the CNN classifier on a 4 GB RAM general purpose computer and a 2.4 GHz Core i5 processor. More memory is needed for using some other proponent or using a deeper CNN model. A broad CNN model on the GPU system would certainly boost the overall performance of the recognition system for human activity.
  • Due to the unavailability of various tasks, it can only identify five types of activity groups.
  • Due to a lack of a broad data collection, it could not test the proposed human activity recognition system with distinct weather and light conditions.

5. Automated Daily Human Activity Recognition for Video Surveillance Using Neural Network

IEEE/2017 | doi:10.1109/ICSIMA.2017.8312024
Conference Name : *4th IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) 28-30 November 2017, Putrajaya, Malaysia*

- **Problem addressed :**

The research paper addresses the problem of efficient human activity recognition within video surveillance systems. It highlights the limitations of traditional surveillance methods that rely heavily on human monitoring and the need for more automated and intelligent systems to detect, analyze, and recognize human activities without constant human intervention.

The key issues addressed include:

1. Dependency on Human Monitoring: Traditional surveillance systems require constant human supervision, which can be prone to errors, fatigue, or inconsistency in monitoring activities.
2. Need for Automated Recognition: The paper aims to develop a system that automates the process of identifying and categorizing human activities without the need for continuous human intervention.
3. Enhanced Surveillance for Security: The system aims to improve surveillance by enabling the detection of suspicious or abnormal activities, thereby enhancing security measures in various settings.
4. Utilizing Digital Image Processing and Neural Networks: The research aims to leverage digital image processing techniques and neural networks to extract features from video data and classify various human activities accurately.
5. Performance Evaluation: The paper presents results and analysis showing the efficiency and accuracy of the proposed system in recognizing different activities such as walking, sitting, boxing, hand waving, and lying down.

- **Sub problems addressed :**
    1. Background Subtraction : The need to separate moving objects (foreground) from the static background in surveillance video frames. Achieved through techniques like frame differencing or statistical modeling.
    2. Binarization : The process of converting grayscale images to binary (black and white) images. Helps simplify subsequent processing by segmenting objects of interest.
    3. Morphological Operations : Techniques to manipulate shapes and structures in binary images. Includes dilation (expanding object boundaries) and erosion (shrinking object boundaries).
    4. Feature Extraction : Extracting relevant features from video frames to represent human activities.These features serve as input to the neural network.
    5. Neural Network Architecture : Building a robust neural network for classification.Using a multilayer feed-forward perceptron network.Training, testing, and validating the network to achieve high performance.
    6. Activity Recognition Rate : Evaluating the system's effectiveness in recognizing human activities.The ultimate goal is to achieve promising performance in activity recognition.
- **Challenges of the problem** :
    1. Background Subtraction: Accurately subtracting the background from the video frames can be challenging, especially in complex environments with varying lighting conditions and occlusions.

2. Binarization: Choosing an appropriate threshold for binarization can be difficult, as it affects the accuracy of activity recognition. Setting the threshold too high or too low can result in false positives or false negatives.
3. Morphological Operations: Applying morphological operations to enhance the extracted features requires careful parameter tuning. Selecting the right structuring element and size can impact the accuracy of the system.
4. Training the Neural Network: Building a robust neural network for activity recognition requires a large and diverse dataset for training. Collecting and labeling such a dataset can be time-consuming and labor-intensive.
5. Similarity in Human Shape and Movement: Recognizing activities that have a high similarity in human shape and movement can be challenging. Distinguishing between similar activities, such as walking and running, requires the system to capture subtle differences in motion patterns.

- **Summary :**

  The system uses a single static video camera in an indoor environment. The solution approach involves several steps:

  1. Preprocessing and feature extraction: The first step is to preprocess the video frames to make them understandable for the computer. This includes determining the background frame, subtracting the background from the foreground image, thresholding the image using the Otsu method, applying a median filter to reduce noise, and performing morphological operations like dilation and erosion. The result is a binary image with the background represented in black and the human body in white. Blob analysis is then performed to calculate features such as bounding box, area, and centroid of the human body.

  2. Building a features database: From the extracted information, a database of features is created. This includes values for the bounding box, centroid, and area for each activity. These values differ for different activities.

  3. Constructing a neural network: A multilayer perceptron feed-forward neural network is used to train the system. The input data is the extracted features, and the output is the designated class for each activity. The system is trained using a portion of the collected samples, and the remaining samples are used for testing and validation.

  4. Recognition: After training the system, it is tested by inputting an image and running it through the neural network. The system recognizes the activity based on the input image and provides an output.

  Overall, the system achieves a satisfactory level of accuracy in detecting and recognizing human activities in the video footage.

- **Performance evaluation metrics that they have used:**

In the study, the performance of the designed system was evaluated using the recognition rate as the main metric. The recognition rate was measured by observing the validity of the system to recognize human activity. The recognition rate was analyzed through the use of a confusion matrix, which is a supervised learning terminology used to measure the performance of a classification model. The confusion matrix compares the number of correctly predicted samples from the classification set model to the number of wrongly predicted samples. The recognition rate was calculated using the equation:

*Recognition Rate (%) = (Number of Correctly Classified Samples / Total Number of Samples) * 100%*

The overall recognition rate for the designed system was reported to be 94%.

6. An Efficient Human Activity Recognition Using Hybrid Features and Transformer Model

IEEE/2023 | doi:10.1109/ACCESS.2023.3314492
Journal Name : IEEE ACCESS

- **Problem addressed :**

The problem addressed in the research study is the need to improve the efficiency and accuracy of human activity recognition (HAR) systems. While previous studies have used manual features to identify human activities, the performance of such features degrades in complex situations. Therefore, the research aims to enhance the efficiency and accuracy of HAR systems by proposing a HAR system that applies data enhancement techniques and captures robust and discriminative features from each activity instance. The study also utilizes a transformer model, known for its ability to capture long-range dependencies, to improve the recognition of complex human activity patterns.

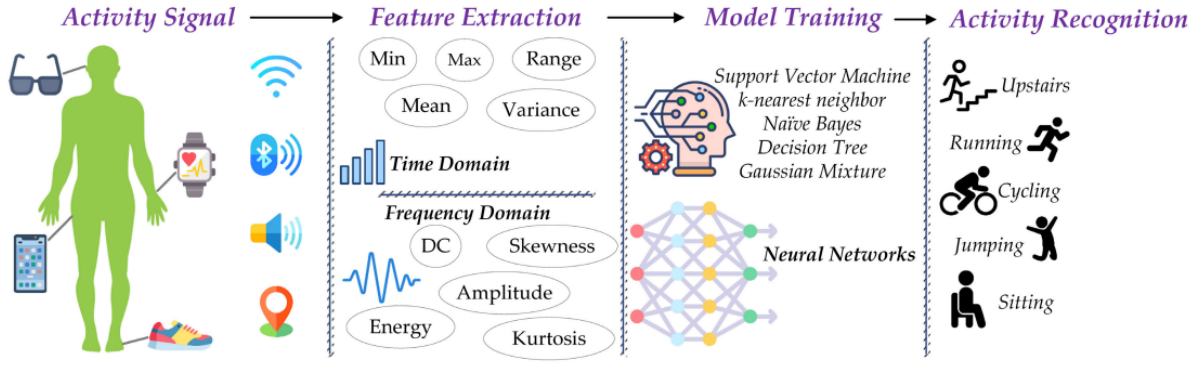


FIGURE 2. An illustration of HAR process using conventional approaches.

- **Sub problems addressed :**

1. Capturing robust and relevant features for human activity recognition (HAR).
2. Enhancing the size of training data through data enhancement methods.
3. Increasing the identification rate of human activities and reducing the training time of the model under limited computational resources using a transformer model.
4. Comparing the performance of the HAR technique using a transformer model with existing HAR techniques.

- **Challenges of the problem** :
    1. Complexity of human activities: Human activities can vary greatly in terms of duration, intensity, and context, making it difficult to build robust and accurate recognition algorithms.
    2. Choice of sensors and their placement: The performance of HAR algorithms can be impacted by the choice of sensors and their placement. Wearable sensors may provide more accurate data but may be more intrusive for the user, while non-wearable sensors may be less accurate but more convenient for the user.
    3. Unstructured and unpredictable environments: There is a growing interest in developing HAR algorithms that can work in real-world environments, where the activities may be unstructured and unpredictable.
    4. Data fusion: One approach to address the challenge of unstructured environments is to use multi-modal data fusion, where raw data from various sensors is merged to offer a more detailed view of the activity being performed.

- **Novelty/ Key contributions of the paper:**

  The key contributions and novelty of the research study on human activity recognition using hybrid features and a transformer model are as follows:

    1. Captured robust and relevant features for human activity recognition (HAR): The study extracted a diverse set of features from sensor data to capture a wide range of patterns and information, making the model more robust.
    2. Data enhancement methods: The research employed data enhancement techniques to increase the size of the training data, which can improve the performance of the HAR system.
    3. Use of a transformer model: The study utilized a transformer model for HAR, which is a type of deep learning model that can capture long-range dependencies in the data. This model was chosen to improve the identification

rate of human activities and reduce the training time under limited computational resources.

4. Improved identification results: The proposed HAR technique using a transformer model achieved better identification results compared to existing HAR techniques. The accuracy of the model was evaluated on three datasets (WISDM, PAMAP2, and UCI HAR), and it outperformed the baseline methods.

Overall, the research study contributes to the field of HAR by capturing robust features, employing data enhancement techniques, utilizing a transformer model, and achieving improved identification results compared to existing techniques.

- **Summary :**

The research study proposed an efficient approach for human activity recognition (HAR) using hybrid features and a transformer model. The goal was to improve the efficiency and accuracy of HAR systems. The approach involved the following steps:

1. Data Enhancement: Data enhancement techniques were employed to increase the size of the training data. This helped in capturing more diverse patterns and information from the sensor data.

2. Feature Extraction: Relevant and discriminative features were extracted from each activity instance. The proposed approach used hybrid features, which could capture both low-level and high-level information from the sensor data. These features were designed to enhance the discriminative power of the system.

3. Transformer Model: A transformer model, known for its ability to capture long-range dependencies, was employed for activity recognition. The extracted hybrid features were used as input to the transformer model. The transformer model was constructed with a fewer number of layers to improve efficiency and reduce training time.

4. Performance Evaluation: The proposed approach was evaluated on three datasets: WISDM, PAMAP2, and UCI HAR. The performance of the transformer model was compared to existing HAR techniques. The results showed that the proposed approach achieved better identification results compared to the baseline methods.

Overall, the research study demonstrated the effectiveness of using hybrid features and a transformer model for HAR. The approach improved the accuracy of activity recognition and showed promising results on multiple datasets.

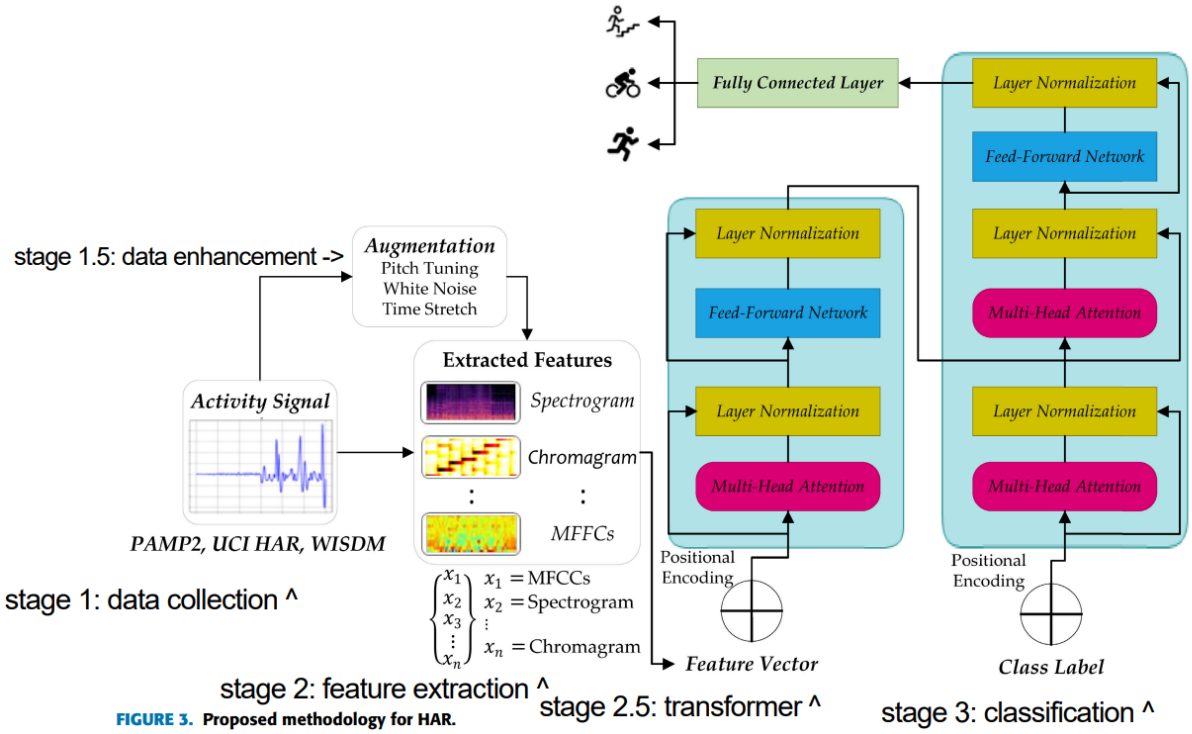


FIGURE 3. Proposed methodology for HAR.

- **Pros and cons of the method:**

Advantages:

1. Improved efficiency and accuracy: The proposed hybrid features and transformer model enhance the efficiency and accuracy of human activity recognition systems compared to existing techniques.
2. Robust feature extraction: The study captures robust and discriminative features from sensor data, allowing for a wide range of patterns and information to be captured, making the model more robust.
3. Long-range dependency capture: The transformer model used in the study is capable of capturing long-range dependencies, which is beneficial in recognizing human activities.
4. High recognition rates: The proposed model achieved high accuracy rates for different datasets, such as 98.2% for PAMAP2, 98.6% for UCI HAR, and 97.3% for WISDM, outperforming baseline methods.

Disadvantages:

1. Domain expertise required: The extraction and combination of a large number of diverse features require domain expertise, which may limit the applicability of the proposed approach to other datasets.

2. Limited evaluation on other datasets: While the proposed model showed promising results on the evaluated datasets, its effectiveness on other datasets with different characteristics remains to be seen.
3. Computational resources: The study mentions that the transformer model was designed to work under limited computational resources, but the specific limitations and resource requirements are not discussed in detail.

Overall, the research study demonstrates the advantages of using hybrid features and a transformer model for human activity recognition, but further investigation and evaluation on different datasets are needed to fully understand its potential and limitations.

- **Any future enhancement scope :**

  The research study on human activity recognition using hybrid features and a transformer model suggests several potential areas for future enhancement. These include:
  1. Exploration of other datasets: While the proposed technique showed promising results on the three datasets (WISDM, PAMAP2, and UCI HAR), it would be beneficial to evaluate its performance on other datasets. Different datasets may have unique characteristics that could affect the model's effectiveness.
  2. Comparative investigation of DL approaches: Conducting a comparative investigation of human activity recognition systems based on deep learning approaches using various datasets would provide insights into the performance and robustness of different models.
  3. Combination of device-based and device-free activity recognition: Exploring the combination of device-based activity recognition (using wearable sensors) and device-free activity recognition (using non-wearable sensors) could enhance the overall accuracy and applicability of the HAR system.
  4. Further feature extraction techniques: While the study extracted a diverse set of features, there is room for exploring additional feature extraction techniques to capture a wider range of patterns and information from sensor data.
  5. Optimization of the transformer model: The transformer model used in the study could be further optimized to improve its efficiency and training time under limited computational resources.

  Overall, these potential areas for future enhancement aim to enhance the accuracy, efficiency, and applicability of the human activity recognition system using hybrid features and a transformer model

- **Open issues not addressed:**
  The research study on human activity recognition using hybrid features and a transformer model did not address the following open issues:

1. Generalizability: The study evaluated the proposed technique on three specific datasets (WISDM, PAMAP2, and UCI HAR). It is unclear how well the model would perform on other datasets with different characteristics. Further investigation is needed to assess the generalizability of the approach.
2. Feature Extraction Complexity: The proposed technique required domain expertise to extract and combine a large number of diverse features. This process can be complex and time-consuming. Exploring more efficient and automated feature extraction methods could be beneficial.
3. Comparative Evaluation: The study did not compare the proposed technique with other state-of-the-art human activity recognition (HAR) systems based on deep learning approaches using different datasets. Conducting a comparative evaluation with a wider range of HAR systems and datasets would provide a more comprehensive understanding of the proposed technique's performance.
4. Combination of Device-based and Device-free Activity Recognition: The study focused on device-based activity recognition using sensor data. Exploring the combination of device-based and device-free activity recognition approaches could enhance the overall performance of HAR systems.

- **Performance evaluation metrics that they have used:**

  The performance evaluation metrics used in the research study on human activity recognition using hybrid features and a transformer model include recall, precision, F1-score, and recognition accuracy. These metrics were used to assess the performance of the transformer model on three datasets: WISDM, PAMAP2, and UCI HAR.

- **Dataset used :** WISDM, PAMAP2, and UCI HAR.

7. A Survey on Human Activity Recognition And Classification

IEEE/2020 | doi: 10.1109/ICCSP48568.2020.9182416
Journal Name: *2020 International Conference on Communication and Signal Processing (ICCSP)*

- **Problem addressed :**

  This paper addresses a survey on various methods on HAR (Human Activity Recognition). It provides a detailed survey on majorly used methods such as vision-based (using pose estimation), wearable devices, and smartphone sensors. This paper also discusses their uses, advantages, disadvantages and explains their

implementation. It also addresses the popularity of these methods throughout the time.

- **Challenges of the problem** :

  The challenges in Human Activity Recognition (HAR) and Classification, highlighted in the paper, include sensor movement, placement, background clustering, and inherent variability in how individuals perform activities. These complexities impact data accuracy, activity differentiation, and the need for robust methods to address individual variations. Privacy concerns, particularly in vision-based approaches, add another layer of challenge. Addressing these issues is essential for successful implementation of HAR systems in real-world contexts.

- **Novelty/ Key contributions of the paper:**

  The novelty of this paper lies in its comprehensive exploration and analysis of Human Activity Recognition (HAR) and Classification, addressing the challenges in an unrestricted environment. The paper contributes to the field by:

  1. Methodological Diversity:

  - The paper covers three distinct methods for HAR—vision-based (pose estimation), smartphone sensor-based, and wearable sensor-based. This methodological diversity allows for a thorough examination of different approaches, offering readers insights into the strengths and weaknesses of each.

  2. Recent Research Synthesis:

  - The paper conducts a survey of recent research papers from 2016-2019, providing a snapshot of the advancements in HAR during this period. This synthesis enables readers to grasp the current landscape of HAR methodologies and emerging trends.

3. Technological Comparison:

- A significant contribution is the comparative analysis of the three methods, including a discussion on their pros and cons. This comparative study aids researchers and practitioners in understanding the trade-offs and benefits associated with each approach, guiding the selection of suitable methods for specific applications.

4. Focus on Vision-Based Approach:

- The paper highlights the increasing popularity of the vision-based approach, particularly using pose estimation. This emphasis reflects the evolving trend in HAR research, shedding light on the growing importance of vision-based techniques and their potential to overcome limitations with advancements in technology.

5. Insights into Real-world Implementation:

- The discussion on challenges, findings, and future scope provides valuable insights for researchers, developers, and practitioners aiming to implement HAR systems in real-world scenarios. The paper acknowledges the practical considerations and limitations, offering a grounded perspective on the feasibility and challenges of deploying HAR technologies.

- **Summary :**

The paper adopts a comprehensive approach to Human Activity Recognition (HAR) and Classification, exploring three main methods: vision-based (pose estimation), smartphone sensors, and wearable devices. The solution approach involves a detailed review of recent research papers from 2016-2019, categorizing them into these three techniques.

In the vision-based approach, the authors discuss the Pose-Based Approach, emphasizing the significance of poses in analyzing videos containing human subjects. This method involves assessing the composition of body parts through 3D poses or 2D human pose estimation (HPE). A regression problem is formulated, modeled with a Convolutional Neural Network (CNN), to predict the pixel

coordinates of 15 key body joints. The authors mention popular datasets like MPII containing labeled images for activity recognition.

The smartphone sensor-based approach leverages the ubiquitous nature of smartphones, which are equipped with built-in sensors like accelerometers and gyroscopes. These sensors capture information from body gestures to recognize activities. Machine learning techniques such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), Bagging, and Ada Boost are employed for classification. Practical considerations, such as sensor frequency and smartphone location, are acknowledged as limitations.

The wearable sensor-based approach involves mounting sensing devices on the subject to collect data from multiple sensors, including accelerometers, gyroscopes, magnetometers, and RFID tags. These devices are designed for user accessibility, providing lightweight and comfortable options for activity monitoring. After feature extraction and modeling, machine learning algorithms are applied to recognize human activities. Despite the benefits, challenges are noted, such as the need to wear multiple sensors on different body parts, potentially causing discomfort for the subject.

The paper then conducts a comparative study, presenting a table showcasing the accuracies of different machine learning algorithms used in developing HAR systems. The findings reveal the commonality of smartphone and wearable sensor technology in recent HAR research. The authors observe that the pose-based approach, while not popular initially due to limitations in 3D space, has gained traction. They highlight the potential of vision-based approaches with advancements in technology, particularly the availability of machines with high computational power, making them a promising choice for HAR in the future. The conclusion emphasizes the emergence of wearable technology as a practical solution while acknowledging the need for further research to enhance accuracy and address limitations in recognizing certain actions.

- **Pros and cons of the method:**

  Pros:

  1. Comprehensive Review: The paper provides an extensive review of recent research papers (2016-2019) in the field of Human Activity Recognition (HAR).

  2. Diverse Methods: It categorizes activity recognition into three main methods - vision-based (pose estimation), smartphone sensors, and wearable devices, offering a broad perspective.

  3. Comparative Study: The authors conduct a comparative analysis, presenting a table showcasing the accuracies of different machine learning algorithms used in developing HAR systems.

  4. Insights into Technologies: The findings reveal the popularity of smartphone and wearable sensor technology in recent HAR research, shedding light on the current trends in the field.

  5. Identification of Trends: The paper notes the emergence of wearable technology as a practical solution and highlights the potential of vision-based approaches with advancements in technology.

  Cons:

  1. Limited Temporal Scope: The review focuses on research papers from 2016-2019, potentially missing out on more recent developments in the rapidly evolving field of HAR.

  2. Privacy Concerns: Vision-based approaches, while effective, raise privacy concerns, as individuals may feel uncomfortable or compelled to modify their behavior due to surveillance.

  3. Practical Limitations: Smartphone sensor-based approaches face practical limitations, such as variations in sensor frequency and smartphone location, impacting real-time data collection.

  4. Wearable Sensor Challenges: Despite the benefits, wearable sensor-based approaches are noted for potential discomfort and inconvenience to subjects due to the need for multiple sensors on different body parts.

  5. Lack of Unified Metric: The paper notes the absence of a specific indicator or measurement to determine whether wearable sensors are better than smartphone sensors, leaving the choice context-dependent.

- **Datasets Used:** MPII Dataset

8. Human Activity Recognition In Videos

Stanford/2012
Journal Name: CS 229 Machine Learning Final Projects, Autumn 2012

- **Problem addressed :**

  The paper tackles the challenge of recognizing human activities in diverse videos with varying backgrounds and camera motions. It addresses the limitations of models relying on low-level features, especially in scenarios where video classes share similar objects and backgrounds. The focus is on accurate object track detection, crucial for understanding the interrelation between objects and event labels. To overcome the high human effort in obtaining object tracks, the paper proposes a solution that involves extracting candidate tracks and modeling the selection of correct tracks using Latent SVM (LSVM). This joint framework enhances action recognition and weakly supervised object tracking, outperforming existing methods on the Olympic Sports Dataset.

- **Challenges of the problem** :

  The challenges in human activity recognition in videos involve coping with diverse backgrounds, camera movements, and limited performance of models relying on low-level appearance and motion features. Additionally, recognizing activities with similar objects and backgrounds poses difficulties. The paper addresses these challenges by proposing a method that leverages Latent SVM to jointly perform action recognition and weakly supervised object tracking. This approach reduces the dependence on explicit object annotations and minimizes the human effort required for obtaining accurate object tracks in videos.

- **Novelty/ Key contributions of the paper:**

  - Joint Framework: Introduces a novel joint framework for action recognition and weakly supervised object tracking.

- Latent SVM Integration: Combines Latent SVM to model the choice of correct object tracks, addressing the challenge of accurate object track extraction.

- Candidate Track Extraction: Proposes a method using Deformable Part-based Models and a tracking algorithm to extract candidate object tracks.

- Spatio-temporal Motion Modeling: Captures spatio-temporal object motion through features extracted from object tracks, enhancing discriminative power.

- Semi-Convex Optimization: Utilizes a semi-convex optimization approach for training the discriminative model, accommodating the absence of explicit object annotations.

- Performance Improvement: Demonstrates performance improvement over state-of-the-art methods on the Olympic Sports Dataset, showcasing the effectiveness of the proposed approach.

- **Summary :**

  This paper tackles the challenge of classifying real-world videos based on human activity, addressing issues arising from diverse backgrounds and camera motions. The traditional use of low-level appearance and motion features often falls short, especially when videos share similar objects and backgrounds. The paper introduces a comprehensive solution by integrating action recognition and weakly supervised object tracking within a joint framework. It acknowledges the difficulty of accurate object track extraction and proposes a Latent SVM approach to model the selection of correct object tracks.

  To overcome the challenge of reliable object track extraction, the paper suggests the use of Deformable Part-based Models and a tracking algorithm, extracting candidate object tracks. The spatio-temporal object motion is then modeled using features derived from these tracks. The proposed framework allows simultaneous action recognition and object track extraction, contributing to a more robust and discriminative selection of object tracks for event classification.

The model's performance is evaluated on the Olympic Sports Dataset, a collection of 800 sports videos. The results showcase improvement over baseline methods, including Bag of Words, Niebles et al., and Tang et al. The proposed approach outperforms state-of-the-art methods, particularly excelling in events where human motion provides crucial information.

The paper's novelty lies in its integration of action recognition and weakly supervised object tracking, addressing the challenge of accurate object track extraction in real-world videos. By treating the choice of object tracks as latent variables, the model identifies the most informative tracks, leading to more accurate event classification. The joint framework, leveraging Latent SVM, demonstrates superior performance, showcasing its effectiveness in scenarios where traditional methods may struggle.

The experimental results, presented in terms of average precision for each event class, validate the efficacy of the proposed approach. Additionally, qualitative results highlight the model's ability to select the most discriminative human track even in videos with multiple individuals.

In conclusion, the paper contributes a valuable approach to human activity recognition in videos, offering a robust solution to the challenges posed by diverse backgrounds, camera motions, and the need for accurate object track extraction in real-world scenarios.

- **Pros and cons of the method:**

  Pros:

  1. Integration of Action Recognition and Object Tracking: The paper introduces a joint framework that seamlessly combines action recognition and weakly supervised object tracking, providing a holistic solution to video analysis.

2. Latent SVM for Discriminative Object Track Selection: Leveraging Latent SVM, the model addresses the challenge of accurate object track extraction by treating the choice of object tracks as latent variables, leading to a more discriminative selection.

3. Performance Improvement: Experimental results demonstrate superior performance over baseline methods and state-of-the-art approaches, showcasing the effectiveness of the proposed framework, particularly in events where human motion is crucial.

4. Flexibility in Object Track Extraction: The paper suggests using Deformable Part-based Models and a tracking algorithm for candidate object track extraction, offering a flexible and adaptive approach to handle diverse videos.

Cons:

1. Dependence on Initial Object Detection: The paper relies on an initial object detection step, and the performance may be influenced by the accuracy of this step. Inaccuracies in object detection could affect the subsequent steps of the proposed framework.

2. Computational Complexity: The method's reliance on Stochastic Sub-Gradient descent for optimization, especially in the context of a large search space for object tubes, may introduce computational challenges, affecting the scalability of the proposed approach.

3. Performance Variability: The model's performance may vary based on the nature of events, and it acknowledges a drop in performance for events where the initial object detection faces challenges, such as significant deformation of humans.

4. Limited Quantitative Evaluation of Object Tracking: While the paper mentions the intention to quantitatively evaluate object tracking in future work, the current focus is primarily on event classification, leaving room for further exploration of tracking performance.

Despite these challenges, the proposed approach offers a valuable contribution to the field of human activity recognition in videos, demonstrating promising results and paving the way for future advancements.

- **Datasets Used:**

  The paper does not explicitly mention the name of the dataset used. However, it refers to the "Olympic Sports Dataset" containing 800 sports videos collected from YouTube. The dataset involves events with both object motion and human actions. The paper utilizes this dataset for complex event classification, focusing on extracting human objects from each video segment.