Human Activity Recognition using Support Vector Machines

Anonymous Author(s)
Affiliation
Address
email

Abstract

Human Activity Recognition (HAR) is a fundamental discipline in the field of computer vision and machine learning. It plays a pivotal role in many applications, from intelligent surveillance systems to healthcare monitoring and beyond. This report explores the application of Kernelized Support Vector Machines for classifying human activities from image datasets. The study covers a diverse dataset of 15 classes of activities, focusing on robust preprocessing and augmentation techniques. Kernelized SVMs, employing various kernels, are assessed for their ability to capture complex relationships in images. This study also compares and analyzes the performance with other techniques of performing human activity recognition like smartphone sensors, time-series data, wearable sensors, etc. The findings highlight Kernelized SVMs competitive performance in human activity recognition, making them a valuable tool for real-world applications.

1 Introduction

Human Activity Recognition (HAR) is a transformative field that allows machines to understand and classify human actions and behaviors. This report focuses on using Support Vector Machines (SVMs) for HAR, as they offer the advantage of interpretability, which is often overlooked in the more commonly used deep learning methods. The report analyzes an image dataset that encodes the intricate details of human actions pixel by pixel, with the goal of uncovering hidden insights and patterns that may not be immediately apparent. Exploratory Data Analysis (EDA) is emphasized as a crucial tool for uncovering these insights, and the report builds upon existing analyses to drive toward a more profound understanding of HAR using SVMs on image data. The potential applications of this technology are vast, ranging from improving healthcare to enhancing security and athletic performance. Overall, this report is a commitment to contributing to the ever-evolving landscape of technology and science by exploring the uncharted facets of human activity recognition using SVMs on image data.

2 Related work

Human Activity Recognition (HAR) is a field of study that has seen the exploration of several machine learning algorithms and techniques to improve activity prediction accuracy and efficiency. The applications of HAR are vast, including healthcare. Researchers have made notable contributions to this area, revealing insights into different aspects of HAR.

One significant finding is the effectiveness of the Random Forest (RF) algorithm in activity prediction, even though it is relatively slow during model construction. Research comparing various machine learning classifiers supports this observation [2]. The K-Nearest Neighbor (KNN) algorithm has demonstrated impressive accuracy and performance in recognizing human activities, particularly

in Industry 4.0 and smart factory settings [1]. Similarly, the Gaussian Naive Bayes (GNB) algorithm has shown promise in achieving high accuracy rates for activity recognition, outperforming traditional Naive Bayes (NB) classifiers [3]. For smartphone sensor-based HAR, decision tree-based models have been proposed, revealing that behavior-based models can offer high accuracy and efficiency in identifying human activities [**?**]. Furthermore, HAR has been comprehensively explored, delving into time-series data, data acquisition, preprocessing, and the various approaches. This comprehensive survey discusses machine learning algorithms, their respective datasets, and the potential challenges in HAR [**?**].

In the context of our work, which focuses on Human Activity Recognition using kernelized Support Vector Machines (SVMs) on image datasets, these prior studies offer valuable insights and methodological comparisons. Our research will contribute to the growing knowledge in HAR and provide a foundation for further advancements in activity recognition using image data. By leveraging SVMs with kernelization, we aim to bring a new perspective to this field, potentially addressing challenges that arise when dealing with image datasets. This work will likely contribute to the ongoing efforts to improve the accuracy, efficiency, and versatility of HAR systems.

Table 1: Related Work

Paper and Dataset	**Methods/Algorithms Used**	**Results**
Random Forest for Human Daily Activity Recognition[2]	Random Forest (RF), ANN, KNN, LDA, Naïve Bayes, SVM	RF achieved the highest accuracy but was slower on large datasets. Accuracy: 87.16%
Human Activity Recognition Using K-Nearest Neighbor Algorithm[1]	K-Nearest Neighbor (KNN)	Impressive testing accuracy of 90.46% with k=20 for smart factory activities.
Human Activity Recognition Using Gaussian Naïve Bayes Algorithm on sensor data[3]	Naive Bayes(NB) and Gaussian Naïve Bayes (GNB)	Achieved an accuracy rate of 89.5%, outperforming traditional Naïve Bayes.

3 Dataset Description

The dataset contains over 12k labeled images of 15 human activity classes and validation images. It has separate folders for each class, with the train directory for model training and the test directory for predictions. The dataset also has two CSV files, testing_set.csv, and training_set.csv, providing further information. The dataset includes 12,600 training files, 5,410 testing files, and two CSV files.

4 Exploratory Data Analysis (EDA)

4.1 Class Imbalance

There are 15 classes of labels in the dataset and each image has a label out of these 15 classes. All the 15 classes are balanced with 840 images under each class.

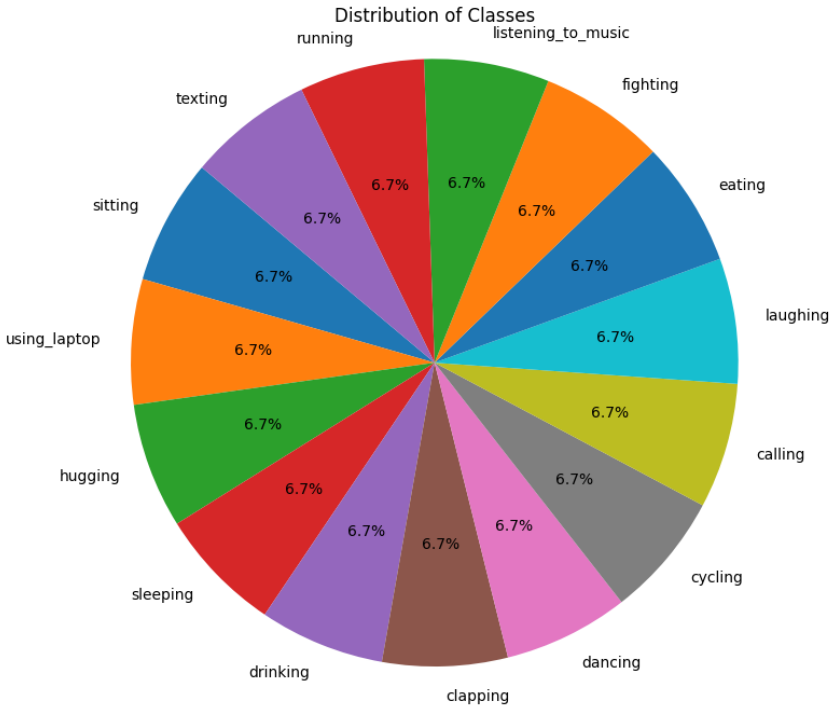

Figure 1: Pie Chart of Different classes

4.2 Image Dimensions

The Image dimensions of all the 12600 images are different and hence the average aspect ratio(width/height) of each class is also different 2.

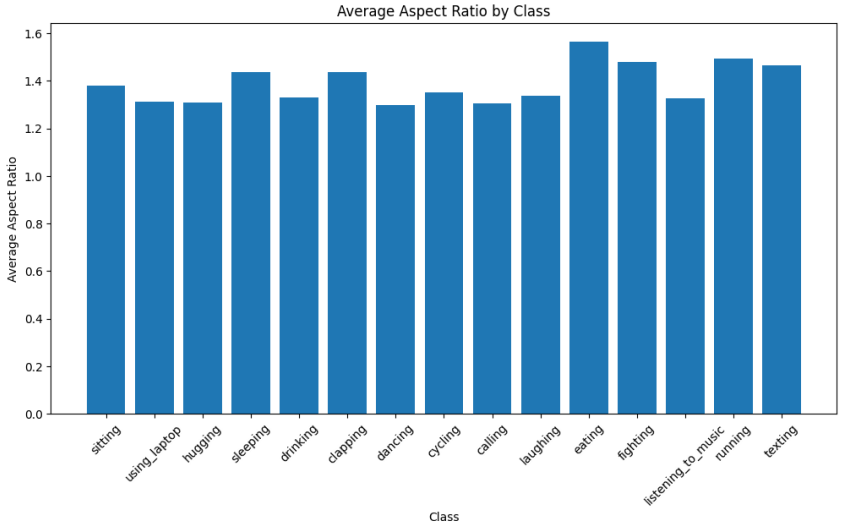

Figure 2: Average Aspect Ratio by Class

4.3 Pixel Intensities

Each of the images has different pixel intensities. Hence, the mean pixel intensity 3 and the standard deviation of the pixel intensities 4 for each class are also different. The difference in mean and standard deviation of the pixel intensities indicate the:- - various environments in which the activity is being performed, for example, indoor vs outdoor environments. - style or the manner in which

the activities are being performed, for example, fast or dynamic activities vs slow activities. - the lighting conditions of the area from which the image is taken. Low pixel intensities could represent low lighting and vice versa.

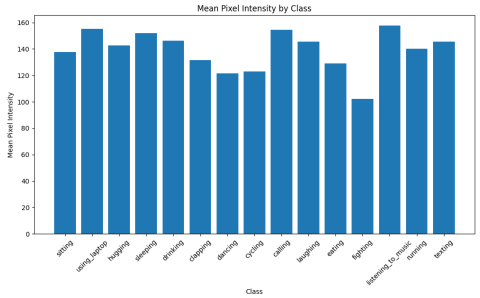

Figure 3: Mean pixel intensity by class

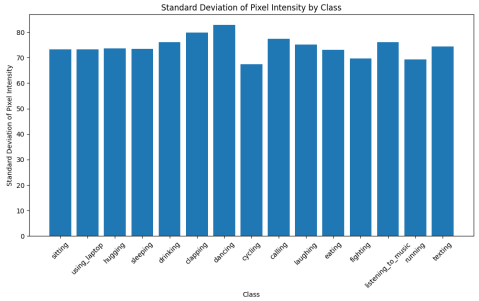

Figure 4: Std deviation of the pixel intensity

4.4 Canny Edge Detection

Canny Edge detection has also been performed to extract information about the edges in the images. It highlights significant transitions or boundaries between different objects in an image. Edges serve as important features for image analysis, and performing Canny edge detection shows that it can detect edges fairly well on our dataset.

Figure 5: Sample Image (Grayscale)

Figure 6: Sample Image after Edge Detection

5 Pre-processing of Data

From our analysis, we find that the Dataset is balanced, and there are no missing/null values. There is one unclassified, duplicate image in the train set that has been dropped. We perform various Pre-processing steps to ensure that the input data is well-suited for the algorithm we are performing. This helps enhance the model performance and facilitates the extraction of relevant information from images for accurate classification or recognition tasks.

5.1 Standardization

Standardization is a technique used to scale the values of all the features between 0 and 1 to bring them to a consistent scale. This technique has been applied to all the images in the dataset to ensure that the extensive range of pixel values does not negatively impact the model's training. By standardizing the images, the model can converge faster and focus on more important features, leading to better performance.

5.2 Image Resizing

Our experimentation to resize the image included dimensions such as 200x200, 180x180, 160x160, and 64x64. Notably, the model's performance varied significantly across these different sizes. We

observed that an image dimension of 160x160 yielded the highest accuracy. This improvement can be attributed to a more optimal balance between retaining critical image details, which are essential for effective feature extraction, and computational efficiency. Feature extraction methods such as Histogram of Oriented Gradients (HOG), Canny edge detection, and Local Binary Patterns (LBP) benefitted from this dimension size, as it provided sufficient resolution to capture relevant features without introducing excessive noise or computational complexity.

5.3 Histogram Equalization

The whole dataset has undergone histogram equalization, a technique used to redistribute the intensity values of an image to increase the contrast between different regions and objects within the image. This technique helps make relevant features and patterns in the image more distinct and easier for the SVM to recognize. By performing histogram equalization, the feature extraction methods can more effectively capture the relevant information in the images, ultimately leading to better performance.

Figure 7: Sample Image

Figure 8: Sample Image after Histogram Equalization

5.4 Dimensionality Reduction

There are 12600 images in total, with each image having dimensions as 160x160, i.e. 25600 features for every image. This slows down the model training, and hence, we try out the dimensionality reduction. To do this, we will use the concept of Principal Component Analysis (PCA). The scree plot in figure 6 shows us that 20 Principal components explain about 69% of the variation in data and 50 Principal components explain about 77% of the variation in data.

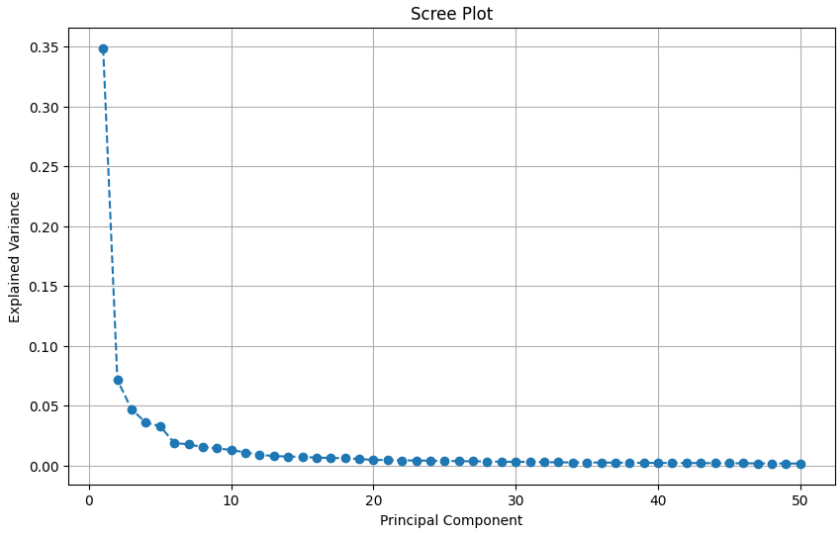

Figure 9: Scree Plot

5.5 Conversion to Grayscale and Resizing

All images in the dataset have been converted to grayscale to reduce complexity. Grayscale images have a single channel of pixel intensity values, whereas color images have three channels (RGB). This conversion is useful for the purpose of HAR, as grayscale images are less sensitive to lighting and allow the model to focus more on the shape and texture, which are relevant for recognizing the activity. Additionally, all images have been resized to the exact dimensions (160*160) to ease model training. Each image had different dimensions, which resulted in a different number of features for each image. Resizing the images to a common dimension ensures that the number of features is the same for every image, simplifying the training process and helping improve the model's performance.

5.6 Noise Removal

Gaussian Blur has been applied to reduce noise in the images. Noise caused due to the random variations in pixel values is smoothened by blurring. It reduces detail but preserves the overall structure of the image and helps in conducting enhanced feature extraction.

Figure 10: Sample Image (grayscale)

Figure 11: Sample Image after Gaussian Blur

6 Methodology

6.1 Feature Extraction

Feature extraction in the context of image datasets is a crucial step in computer vision and image processing. It involves transforming raw image data into a set of representative features that can be used for various analytical tasks like classification, detection, or recognition. Proper feature extraction is crucial for capturing the essential characteristics of human activities in images, such as motion, posture, and interaction with objects.

6.1.1 HOG (Histogram of Oriented Gradients)

Histogram of Oriented Gradients (HOG) is a popular feature descriptor used in computer vision for object detection and image classification tasks. It captures local spatial patterns and structures by representing the distribution of gradient orientations in an image. As part of our project, we utilized the HOG (Histogram of Oriented Gradients) technique to extract features from images. This technique transforms raw image data into a compact and informative representation by encoding gradient information, which is then used as input for our SVM classifier. We developed a function to implement HOG for color images. The HOG method involves two steps. Firstly, to capture the intensity changes in both horizontal and vertical directions, gradients of images are computed by convolving the image with gradient filters. Secondly, the image is divided into small cells, and for each cell, a histogram of gradient orientation is computed. To enhance the robustness of the descriptor to changes in lighting and contrast, we applied Block Normalization. This involved grouping the cell

histograms into blocks and applying normalization. Integrating HOG features into the classification pipeline produced better activity recognition results for our image dataset.

6.1.2 Image Segmentation

Image segmentation is a fundamental process in computer vision and image processing where an image is partitioned into multiple segments. The goal of image segmentation is to change the representation of an image into something that is more meaningful and easier to analyze. The segmentation can help in focusing on the regions of interest (like human figures, specific body parts, or objects interacting with the human) and reducing background noise, which enhances the quality of the features fed into the classification model.

K means clustering K means is a clustering algorithm used for image segmentation by forming clusters of pixels. It is an unsupervised learning algorithm that identifies different clusters in the data based on how similar the data points are. Here, K denotes the number of clusters to be formed. K=16 worked well for the available dataset and is the preferred choice.

Figure 12:

Mean shift clustering Mean shift clustering is a non-parametric, density-based clustering algorithm that can be used to identify clusters in a dataset. Here, unlike the K-means algorithm, the value of K need not be specified. Only a bandwidth parameter which is the size of the window used to compute the mean in the mean shift procedure.

6.1.3 LBP (Local Binary Pattern)

LBP's ability to efficiently encode local textures by comparing each pixel with its neighbors adds valuable information to the feature set. Moreover, its relative invariance to changes in lighting and its computational simplicity made it an attractive option for augmenting our model's capabilities.
When used independently, LBP achieved a moderate accuracy of about 25 percent, but its performance did not significantly improve the model when combined with other methods like Histogram of Oriented Gradients (HOG) and Canny edge detection, yielding a similar accuracy range of 20-25 percent. This outcome led us to reconsider the complementary nature of these techniques and highlighted the complexities involved in effectively integrating diverse feature extraction methods. It underscored that while certain methods are powerful on their own, their effectiveness in combination can vary depending on the specific characteristics and nuances of the dataset and the task at hand.

Figure 13:

6.1.4 Gabor Filter

Gabor filters are mathematical functions that are used to access textures in images and detect edges. Based on signal processing, they are modeled on human vision as they can analyze local spatial and frequency data of the image simultaneously. They are useful when the texture is clear, but they may not perform where there is a large variation in lighting or complex non-uniform textures. They also are highly sensitive to parameter tuning.
Although they are very useful for detecting features related to human expressions, and facial features like eyes, and mouth, which are essential in face recognition, they may not be directly useful for Human Activity Recognition because that is not texture dependent.
We applied the Gabor filter to the dataset, but it led to very poor results, hence the idea was dropped.

6.1.5 SIFT (Scale-Invariant Feature Transform)

SIFT is a computer vision algorithm that detects features in an image that are robust to changes in scale, rotation, and lighting conditions. It does this by identifying significant points, or keypoints, in the image and assigning each one a unique orientation based on the dominant gradient direction in its local neighborhood. SIFT then generates a descriptor for each keypoint that captures its distinctive features. To match keypoints between different images, SIFT uses a nearest-neighbor approach based on the Euclidean distance between descriptors. A ratio test is often used to verify the quality of matches and discard any ambiguous correspondences.
In our feature extraction pipeline, incorporating SIFT did not improve performance, suggesting that SIFT's distinctive keypoints may not be discriminative enough for our image set.

6.1.6 Data Augmentation

It is a technique to create variations of the dataset to expand the data set by adding noise, scaling, rotating, changing illumination, cropping, etc. This is done to prevent overfitting and expose the model to a wider range of data points. However, this can sometimes be very computationally heavy and modelling after that takes a lot of time and resources. We applied data augmentation to create 4 images for every single image by applying variations, but it resulted in a very huge data set. This led to marginal improvements but a lot of increase in resources and time taken.

7 Models

Kernelized Support Vector Machines (SVMs) play a crucial role in Human Activity Recognition on image datasets due to their ability to capture non-linearities in the data, a common characteristic

of image datasets. SVMs capture nonlinear characteristics of data by using the kernel trick which transforms the data into higher dimensional space where it is linearly separable.

7.1 OvR classifier

OvR technique, also known as the One-vs-Rest technique in SVM is a technique that is used in multi-class classification. This technique decomposes the problem into multiple binary classification problems. In this, the number of models is equal to the number of classes to be identified. Each model is specific to a particular class and identifies whether the data point belongs to that class or not by confidence matching. The model which predicts the data point as its own with the highest confidence is the final classification of the overall model.
However, one negative aspect of this technique is that this is very computationally expensive as the number of models is proportional to the number of classes. Hence, it is not a very good approach where the number of classes is many. In this case, since the number of classes was limited to 15, it was very useful and helped in increasing accuracy considerably overall in relative terms.

7.2 Kernel Trick

The kernel trick is a fundamental concept in machine learning, especially in SVMs. It is a technique that allows SVMs to solve complex, non-linear classification problems efficiently by using linear methods. The idea is to transform the non-linearly separable data in the original space into a higher-dimensional space where it becomes linearly separable. This transformation is known as feature mapping.

7.2.1 Gaussian Kernel

The Gaussian Kernel, also known as the Radial Basis Function (RBF) kernel, is a widely used kernel in Support Vector Machines for handling non-linear data separations. It is defined by the equation:

$$K(x, x') = \exp\left(-\gamma \cdot ||x - x'||^2\right) \tag{1}$$

where x and x' are two feature vectors, γ is a parameter that controls the width of the kernel, and $||x - x'||^2$ represents the squared Euclidean distance between the feature vectors.

The Gaussian Kernel is particularly powerful as it transforms the input feature space into an infinite-dimensional space, thereby enabling the SVM to construct a rich and flexible decision boundary. This kernel exhibits a localized behavior, meaning that the similarity between two points decreases as the distance between them increases, with the kernel value ranging from 0 (for points that are far apart) to 1 (for similar points). This property allows the SVM to create complex, non-linear decision boundaries, capturing subtle patterns in the data. In our experiments, the combination of SVM with the Gaussian kernel achieved the highest accuracy among all the kernels tested, highlighting its effectiveness in capturing complex relationships in the dataset.

7.2.2 Linear Kernel

The Linear kernel is the simplest form of SVM kernel and is particularly used in scenarios where the data is linearly separable. It is essentially defined as the product of the input vectors of two data points. This kernel is predominantly effective for data sets where the data can be separated using a straight line or a hyperplane in higher dimensions).

The Linear kernel is represented mathematically as follows:

$$K(x, x') = x^T x' \tag{2}$$

where x and x' are feature vectors.

7.2.3 Polynomial Kernel

The Polynomial kernel is a more complex kernel used in Support Vector Machines to allow for non-linear decision boundaries. It is particularly effective in data sets where the relationship between the features and the target is more intricate than a simple linear relationship. The Polynomial kernel takes the linear kernel and raises it to the power of a specified degree, thus enabling the model to fit more complex patterns.

Mathematically, the Polynomial kernel is expressed as:

$$K(x, x') = (1 + x^T x')^d \tag{3}$$

where x and x' are feature vectors, and d is the degree of the polynomial.

7.2.4 Laplacian Kernel

The Laplacian kernel is an alternative to the Gaussian kernel used in Support Vector Machines for handling non-linear data. It is particularly effective for data sets that require a focus on locality and finer details in the feature space. Unlike the Gaussian kernel, which uses the squared Euclidean distance, the Laplacian kernel is based on the absolute distance between the feature vectors.

The Laplacian kernel is mathematically formulated as:

$$K(x, x') = \exp\left(-\gamma \, ||x - x'||_1\right) \tag{4}$$

where x and x' are feature vectors, $|| \cdot ||_1$ denotes the L1 norm (Manhattan distance), and γ is a parameter that determines the width of the kernel.

8 Result

The overall accuracy obtained on the validation set was 37.5% after applying Canny edge detection and HOG features along with PCA and standardization.

Applying Image segmentation techniques like K-means clustering and Mean-Shift clustering along with canny edge and HOG features obtained a competitive accuracy of 36.6%

We have used an OVR (One-Vs-Rest) classifier. It is a method used for multi-class classification tasks. Instead of classifying all classes simultaneously, it decomposes the problem into multiple binary classification tasks.

A standard multi-model SVM classifier, on the other hand, does not break the problem into binary classification tasks.

OVR breaks down a multi-class classification problem into multiple binary classification problems. For each class, it creates a separate classifier that distinguishes that class from all other classes combined. By dealing with binary decisions, it can simplify the learning process for each classifier.

Hence, OVR produces better results than a single model in general when we apply the same pre-processing techniques to them. The best accuracy that we have produced involves the use of OVR. Without OVR, the accuracy was 35.26%, while, with OVR, it was 37.54% on a validation set of 3780 images.

The better performance of the OvR SVM classifier as compared to the multi-model SVM classifier can be accounted to the application of multiple binary classifications in the OvR classifier. The training of multiple models in the OvR classifier makes that model specialised in classifying that specific class, which helps the final class predictions to be better.

The OvR classifier, along with the Gaussian Kernel, produced the highest accuracy as compared to other kernels such as the linear, the polynomial and the laplacian kernels. The Gaussian kernel maps the feature space into an infinite dimensional feature space, thus enabling the SVM to construct a rich and flexible decision boundary.

9 Inference

In our exploration of Human Activity Recognition (HAR) using Support Vector Machines (SVMs) on image datasets, we found that feature extraction techniques like HOG, LBP, and Canny edge detection played a crucial role. While HOG showed promise in capturing gradient orientation information beneficial for classification, we observed that other methods like LBP didn't significantly contribute. This highlighted the variability in effectiveness among feature extraction techniques based on the dataset characteristics.

Throughout our data preprocessing phase, including standardization, histogram equalization, noise removal, and resizing, we realised the importance of preparing data for efficient pattern learning,

ultimately enhancing classification accuracy. However, the impact of dimensionality reduction through PCA suggested the necessity of a balance between dimension reduction and retaining essential information to avoid losing discriminative features.

Kernel selection significantly influenced classification accuracy. The Gaussian kernel's ability to capture complex relationships in data outperformed linear, polynomial, and Laplacian kernels. Its flexibility to transform data into a higher-dimensional space contributed to a richer decision boundary, showcasing its relevance in SVM-based HAR.

Using the OvR (One-vs-Rest) technique in SVMs, decomposing multi-class classification into binary tasks proved advantageous. This technique allowed us to create specialized classifiers for each class, outperforming a single-model approach.

Our study highlighted the impact of image segmentation techniques like K-means and Mean-Shift clustering alongside feature extraction, indicating their relevance in focusing on regions of interest and reducing background noise for better feature extraction.

While data augmentation expanded the dataset and exposed the model to varied data points, we observed its marginal improvement in accuracy and increased computational demands, thereby inducing us not to use it for our model.

Despite not achieving high accuracy, our study demonstrated the potential of SVMs in HAR using image data. It suggests further exploration into feature extraction methods, advanced deep learning techniques, and addressing class-specific challenges for improved accuracy in real-world applications across healthcare, surveillance, and sports analytics.

References

[1] Saeed Mohsen, Ahmed Elkaseer, and Steffen Scholz. Human activity recognition using k-nearest neighbor machine learning algorithm. 09 2021.

[2] Nurul Retno Nurwulan and Gjergji Selamaj. Random forest for human daily activity recognition. *Journal of Physics: Conference Series*, 1655(1):012087, oct 2020.

[3] Jiajie Shen and Hongqing Fang. Human activity recognition using gaussian naïve bayes algorithm in smart home. *Journal of Physics: Conference Series*, 1631(1):012059, sep 2020.