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

Advancements in Human Activity Recognition: Novel Shape Index Local Ternary Pattern and Hybrid Classifier Model for Enhanced Video Data Analysis

Milind V. Kamble[1], Rajankumar Bichkar[2]

Submitted: 16/09/2023 **Revised**: 17/11/2023 **Accepted**: 29/11/2023

Abstract: Human activity recognition using video data is a fundamental yet challenging task in computer vision. This paper proposes a novel method for spatial feature extraction and classification model design to enhance the accuracy and robustness of human activity recognition systems. The primary proposed work is the development of the Shape Index Local Ternary Pattern, a novel spatial feature extraction method aimed at capturing intricate spatial relationships within video data. This feature significantly enhances the representation of spatial information, thereby overcoming the limitations of traditional texture-based features. Moreover, a Hybrid Classifier Model is proposed, comprising an initial layer integrating a Support Vector Machine and k-Nearest Neighbour classifiers. The outputs of these classifiers are fed into an Artificial Neural Network for the final classification. This hybrid approach combines the strengths of different classifiers and artificial neural network architectures, resulting in improved classification accuracy and robustness. The efficacy of the proposed method was rigorously evaluated and validated through extensive experiments conducted on UCF101 and HACS datasets. The results showed superior performance in accurately identifying a range of human activities. The Shape Index Local Ternary Pattern feature demonstrates its effectiveness in capturing intricate spatial details, whereas the hybrid model shows substantial advancements in classification accuracy and robustness.

Keywords: *Artificial Neural Network, Human activity recognition, Hybrid Classifier, k-Nearest Neighbour classifiers, Shape Index Local Ternary Pattern, Support Vector Machine.*

1. Introduction

Human activity recognition (HAR) involves the analysis of video data to understand and categorize human actions. In HAR, spatial-temporal patterns, called features, are extracted from video data, capturing how things move in both space and time. These features are then used to train a classifier and teach the system to recognize and categorize different human actions based on these patterns. The HAR model interprets movement, gestures, and patterns in videos, enabling systems to identify and label various activities that are crucial for applications in surveillance[1], healthcare for elderly people[2][3], and more[4], [5]. The contemporary field of human activity recognition tackles persistent challenges primarily rooted in the complexities inherent to visual interpretation. Capturing the details of spatial and temporal relationships within video data presents a formidable challenge, compounded by the limitations of traditional feature extraction techniques that are primarily reliant on texture-based analyses. These methodologies often fail to effectively represent intricate spatial and temporal details intrinsic to diverse human actions within video streams.

In pattern-based HAR, spatiotemporal features are extracted and used to train the classifier[6]. One limitation of relying solely on spatiotemporal features for human activity recognition is the potential loss of discriminative information related to finer details and surface properties of the objects involved. Spatial-temporal features capture the dynamic aspects and context of activities but may not provide sufficient information about the specific actions or interactions within the activity. Consequently, the accuracy and robustness of activity recognition systems may be compromised when dealing with complex or subtle activities that require more detailed analysis. Incorporating additional features, such as texture-based features, can help address this limitation by providing complementary information regarding the visual characteristics and appearance of the activities [7], [8]. Activity recognition systems can enhance accuracy and performance by integrating diverse types of features, thereby enabling a more comprehensive representation. A new feature called the Shape Index Local Ternary Pattern (SILTP), a spatial feature extraction technique is proposed that specifically targets and captures intricate spatial relationships within video data. The SILTP represents a variation from conventional texture-based features aimed at overcoming

[1] *Research Scholar, Dept. of E&TC, G. H. Raisoni College of Engineering and Management, Pune, Maharashtra, India*
(e-mail: milind.kamble@vit.edu)
ORCID ID: 0000-0002-9237-184X
[2] *Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India*
** Corresponding Author Email: milind.kamble@vit.edu*

their limitations and accurately representing detailed spatial information.

Traditional pattern-based methods for HAR using single classifiers have shown limitations in terms of accuracy, adaptability, and scalability. In response, introducing hybrid classifiers within HAR significantly enhances accuracy and robustness [9]. By merging multiple classifiers, this hybrid strategy attempts to capture a broader spectrum of activity patterns and improve the overall classification performance [10], [11],[12].

Hybrid classifiers fuse diverse techniques, including deep learning, genetic algorithms, linear discriminant analysis, k-nearest neighbors (k-NN), and Support Vector Machines (SVM) [13], [14], [15]. These methods employ varied feature extraction and selection procedures, rectify class imbalances, and optimize the use of available information. By leveraging the strengths of these various classifiers, the hybrid approach is equipped to achieve better accuracy and robustness.

In the proposed work, an addition involves integrating a hybrid classifier model, which combines the robust capabilities of support vector machines (SVM), k-nearest neighbor (KNN) classifiers, and Artificial Neural Network architectures (ANN). The primary goal of this integration is to effectively utilize features generated by the SILTP, aiming to elevate the accuracy and robustness of classification within the human activity recognition framework.

This study attempts to address the limitations of human activity recognition using video data by introducing new methods for spatial feature extraction and classification. The primary purpose is to enhance the accuracy and robustness of human activity recognition systems, thereby enabling them to accurately interpret and classify a diverse range of human actions depicted in video sequences.

The main contribution of this research is:

1. **Introduce the Shape Index Local Ternary Pattern (SILTP):** Develop and implement SILTP as a novel spatial feature extraction technique to capture intricate spatial relationships within video data. The objective was to overcome the limitations of traditional texture-based features and effectively represent detailed spatial information for improved recognition systems.
2. **Integrate a Hybrid Classifier Model:** Propose and implement a hybrid classification model by leveraging the strengths of the SVM, KNN classifiers, and Artificial Neural Networks. The aim is to utilize SILTP-generated features to effectively enhance classification accuracy and robustness.

In summary, the proposed study aims to address the limitations of traditional pattern-based human activity recognition by utilizing a hybrid classifier. This approach seeks to improve the performance of a HAR system by combining multiple classifiers.

By leveraging advancements in computer vision, sensing technologies, and machine-learning techniques, this study aims to contribute to the field of HAR and enable the development of more effective and reliable systems in various domains. The remainder of this paper is organized as follows. Section 2 discusses methodologies, feature extraction, and proposed SILTP features for HAR. Section 3 discusses the proposed hybrid classifier model, and the results are presented in Section 4.

2. Methodology

This study presents a novel strategy for enhancing human activity recognition using real-world video sequences. The proposed method comprises three main stages: preprocessing, feature extraction, and activity classification. The architecture of the proposed HAR system is shown in Figure (1), where the input to the system is a video sequence. Each video sequence frame underwent pre-processing, including data filtering and background subtraction, to isolate the Region of Interest (ROI). The pre-processed frames, denoted as img^{pre} were then used for feature extraction. Various features, such as SMltp (f^{SMltp}), CS-Mltp ($f^{CS-Mltp}$), MoBSIFT ($f^{MoBSIFT}$), extended SURF (f^{E-SURF}), and SILTP, (f^{SILTP}) were extracted and concatenated as $F = f^{SMltp} + f^{CS-Mltp} + f^{MoBSIFT} + f^{E-SURF} + f^{SILTP}$, are extracted and concatenated to form one complete feature vector, $F = f^{SMltp} + f^{CS-Mltp} + f^{\mathrm{MOBSIFT}} + f^{E-SURF} + f^{\mathrm{SILTP}}$, to represent a single input video. A feature set is created by extracting the features of each video in the dataset. The feature set is divided into two sets: one for training the classifier and the other for testing the trained classifier.

In the proposed work, for activity recognition, a hybrid classifier model combining Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN) is introduced. First, the SVM and k-NN classifiers were trained using training feature sets. Subsequently, the outputs from the SVM and k-NN, denoted as out^{SVM} and $out^{k-NN,}$ respectively, were used to train the ANN classifier, which provided the final decision regarding human activity.

2.1 Pre-processing of the video signal and Region of Interest (ROI)

In the first stage of HAR, the input video signal V^{inp} is sent for video-to-frame conversion, where the individual frames are extracted from V^{inp}. The extracted image frame was denoted as *img*. The extracted image frame *img* was filtered via Gaussian filtering [16] to reduce noise using Equation (1).

In Equation (1), a and b denote the image coordinates and σ indicates the standard deviation. The filtered image frame was denoted as img^{filter} . The filtered image frame, img^{filter}, is then subjected to Background Subtraction (BS) to isolate the ROI from the background (Non-Region of Interest).

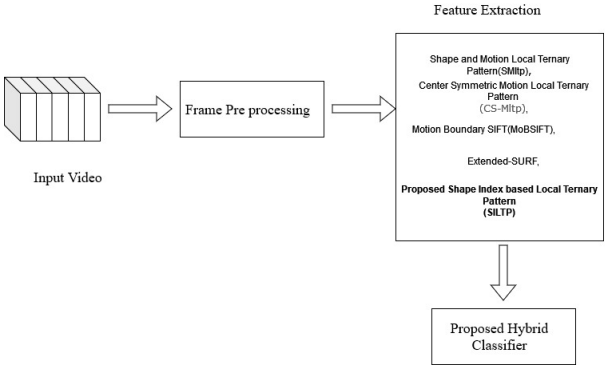

Fig 1: Block diagram of the proposed HAR system.

2.2 Feature Extraction

The following features are extracted from the processed video frames.

2.2.1: Shape and Motion Local Ternary Pattern (SMltp):

The Shape and Motion Local Ternary Pattern (SMltp) [17] is a spatio-temporal feature. It combines two local descriptors: the shape local ternary pattern (Sltp) and the motion local ternarypattern (Mltp).

2.2.2: Center Symmetric Motion Local Ternary Pattern (CS-Mltp):

The Center Symmetric Motion Local Ternary Pattern (CS-Mltp) [6] combines the local ternary pattern with intensity consistency across frames to capture the motion informationin the video signal.

2.2.3: Motion Boundary SIFT:

Motion Boundary SIFT (MoBSIFT) [18] is a feature descriptor that effectively captures temporal variations and motion patterns in a video sequence. It combines Motion SIFT (MoSIFT) and Motion Boundary Histogram (MBH) techniques.

2.2.4: Extended-SURF :

Extended-speeded-up robust features(E-SURF) [19], [20] are designed to be invariant to various image transformations, such as scale changes, geometric variations, and lighting variations.

2.2.5: Shape Index-based Local Ternary

Pattern (SILTP):

For spatial relations, a new feature called the shape-index-based local ternary pattern (SILTP) is proposed in this

$$img^{\text{filter}} = G(a,b) = \frac{1}{2\pi\sigma^2}\cdot e^{\left(-\frac{a^2+b^2}{2\sigma^2}\right)} \qquad (1)$$

study. It incorporates shape index information and creates a ternary pattern representation to capture and recognize finer details and surface properties of objects involved in human activity. The surface topology of an image is described by the shape index, which is a single-value measure[21]. Based on their apparent local shapes, they can be used to locate structures. The proposed feature used a shape index instead of the original image in the binary threshold function.The shape index can be computed differentially from the local second-order derivatives of a textured image. The shape index, S, was calculated using Equation (2) [22]:

$$S = \frac{1}{2} - \frac{1}{\pi}\left[\frac{(-L_{xx} - L_{yy})}{\sqrt{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}}\right] \qquad (2)$$

where L_{xx}, L_{yy} and L_{xy} denote the second derivatives of the texture image in horizontal, vertical, and diagonal directions, respectively.

The local ternary pattern (LTP) descriptor was applied to Shape Index images, which represent the orientation of the local surface normal of an image. The descriptor is computed by dividing the Shape Index image into small regions. For each region, the LTP code for each pixel was calculated using Equation (3):

$$f^{\text{SILTP}}(x_c, y_c) = \sum_{p=0}^{P-1} f(g_p - g_c)3^p \qquad (3)$$

where P is the number of neighbors; x_c and y_c are the coordinates of the central pixel; g_c is the value of the centralpixel; g_p is the value of the p^{th} neighbor; and $f(x)$ is the step function defined in Equation (4).

$$f(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

The resulting LTP codes are then used to form a histogram for each region. The LTP descriptor was normalized to reduce the effects of illumination variation. The SILTP captures shape-related cues that are not fully represented by the existing features (SMltp, CS-Mltp, MoBSIFT, and E-SURF). By combining SILTP with existing features, we can leverage the complementary information provided by each feature to create a more comprehensive representation of human activity. By incorporating shape-related information, the feature set becomes more capable

of recognizing a wide range of human activities, including those with subtle variations in the human body shape and appearance. This holistic representation can capture both motion and shape characteristics, resulting in a richer and more accurate understanding of the activities being performed.

All extracted features are concatenated to form a complete feature vector F as $F = f^{SMltp} + f^{CS-Mltp} + f^{SILTP} + f^{MoBSLBT} + f^{E-SURF}$.

3. Proposed Hybrid Classifier Model

Traditional Human Activity Recognition (HAR) systems utilize off-the-shelf classifier algorithms, such as decision trees, Support Vector Machines (SVM), Hidden Markov Models (HMM), and various machine learning methods to categorize activities. However, these methods have limitations including reduced adaptability, susceptibility to noise and variations, and restricted scalability. To overcome these constraints, the proposed solution suggests employing a hybrid classifier to broaden the capture of activity patterns and enhance the overall classification accuracy.

The proposed novel approach integrates SVM, k-NN, and ANN models into a hybrid classifier system. The proposed hybrid model, outlined in Figure (2), combines the outcomes of SVM and k-NN, trained with the training feature set to train the ANN for human activity classification subsequently.

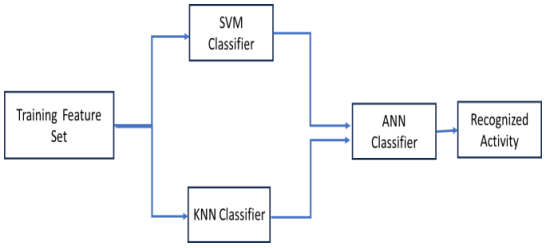

Fig 2. Block diagram of proposed hybrid classifier for HAR.

3.1. Support Vector Machine (SVM)

The training feature set is used to train the SVM classifier. It aims to create a classification function that forms a hyperplane to segregate an N-dimensional space into distinct classes. It uses support vectors to determine the optimal hyperplane with the maximum margin, which is crucial for accurate classification. A non-linear SVM using an RBF kernel function was applied in the proposed system.

3.2. k-Nearest Neighbors (k-NN):

The k-NN was trained using the same training feature set for the SVM classifier. k-NN is a non-parametric algorithm that relies on proximity to classify or predict data points. Predictions were provided based on majority voting by selecting the K-nearest neighbors and using their class labels.

3.3. Artificial Neural Network (ANN):

The outputs from the SVM and k-NN classifiers were combined and used to train the ANN. The ANN, comprising input, hidden, and output layers, combines the outputs from the k-NN and SVM classifiers. The information flows through these layers via synaptic weights, determines the connection strengths between nodes, and provides the final output. The synaptic weight associated with each link offers information on the relative strength of the connections between nodes. The output from the nodes is expressed by Equation (5).

$$out^{ANN} = f\left(\sum_{i=1}^{m} W_i \cdot F_i^* + \text{bias}_i\right) \quad (5)$$

here, F_i^* is the input and W_i is the weight function. In addition, bias $_i$ denotes the network's bias and f is the activation function represented by Equation (6).

$$f(x) = \frac{1}{1 - e^{-x}} \quad (6)$$

The output generated by the ANN model was compared with the target output, and the error E was computed using Equation (7).

$$E = \sum_P \sum_P \left(Out^{ANN} - Out_{\text{tar}}\right) \quad (7)$$

where P denotes the number of training patterns, Out_{tar} denotes the desired output and Out^{ANN} denotes the outcome generated by the ANN. An ANN provides the final outcome for human activity classification, and the error that occurs during classification must be reduced. The reduction in the ANN error was defined as this study's primary objective. Mathematically, the objective function is expressed by Equation (8).

$$Obj = \min(E) \quad (8)$$

This hybrid model, which combines SVM, k-NN, and ANN, presents a more comprehensive and accurate approach to human activity classification, addressing the limitations of traditional pattern-based HAR systems.

4. Result and Discussion

4.1. Experimental Setup

The performance of the proposed hybrid classifier with the SILTP feature was validated using data collected from

datasets 1 and 2. The experimeantal setup to validate the performance of the proposed model is done in three steps:

1. Proposed feature set extraction from dataset 1 and dataset 2 and splting the feature set into training and testing feature sets.
2. Evaluate the effect of proposed SILTP feature on the performance of the HAR system.
3. Evaluate the performance of proposed the Hybrid Classifier by comparing its perfomance with the standard calssifiers.

4.2. Features set from Dataset1 and Dataset 2

The features from datasets 1 and 2 were extracted to create separate feature sets for each dataset. We obtained the proposed features from every video and made a feature set. This set was divided into two parts: training and testing feature set.

Dataset 1: Dataset 1 was obtained from [23], https://This dataset is known as the UCF101-Action Recognition dataset and contains 101 action categories divided into 25 groups. Human-object interaction, body motion only, human-human interaction, playing musical instruments, and sports are the different action types.

Dataset 2: Dataset 2 was collected from [24], http://hacs.csail.mit.edu/. The dataset is called Human Action Clips and Segments (HACS) and has 200 action categories. Each video was less than 4 min long, with a 2.6-minute average.

4.3. Effect of proposed SILTP feature on theperformance of the HAR System

The efficacy of SILTP and the hybrid classifier model was rigorously assessed using the UCF101-Action Recognition and Human Action Clips and Segments (HACS) datasets. The objective is to demonstrate an improved performance in accurately recognizing a diverse variety of human activities within video data. The performance of the proposed SILTP features was assessed to capture intricate spatial relationships within the video data.

A statistical t-test was performed to evaluate whether the addition of the SILTP features improved the classification accuracy of the HAR system. Algorithm (1) shows how a t-test was conducted to compare the performance of the HAR system using two sets of features: one formed by the original collection of features and another created by adding the proposed SILTP feature to the original group of features. These two groups of feature sets were used to train the SVM Classifier model for the HAR.

Algorithm 1 Statistical t-test to evaluate the performance of proposed SILTP feature

Require: Training data, Testing data

Ensure: T-test results

1: Train SVM classifier using the original set of features.

2: The proposed SILTP feature is added to the original feature set, and the SVM classifier is trained again.

3: Split the data into training and testing sets.

4: The classification accuracy of the SVM model is calculated using the original set of features and the original set of features with the proposed SILTP feature.

5: Calculate the difference in classification accuracy between the two models.

6: Calculate the Standard Error (SE) of the difference in accuracy using $SE = \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$, where s_1 and s_2 are the standard deviations of the accuracies of the two models, and n_1, n_2 are the sample sizes.

7: Calculate the tvalue using the formula $t = \frac{x_1 - x_2}{SE}$, where x_1 and x_2 are the accuracies of the two models.

8: Determine the degrees of freedom (df) using $\mathrm{df} = n_1 + n_2 - 2$.

9: The p-value of the t-test was determined using a t-distribution table.

10: **return:** T-test results

According to the t-test results presented in Table (1), a t-value of 3.03 indicates a significant difference between the compared groups of features. This suggests that the variation in classification accuracy between the HAR model using the original set of features and that incorporating the SILTP feature is not due to chance. With a p-value of 0.0025 below the significance threshold of 0.05, there is strong evidence to reject the null hypothesis, which states that there is no significant difference in classification accuracy between the two models. Thus, we can conclude that adding the SILTP feature significantly improved the performance of the HAR model.

TABLE 1: Values obtained after performing t-test on and without proposed SILTP features.

Parameter	Value
Difference in accuracy with and without proposed SILTP features	0.06
Standard deviation with the proposed feature set	0.481
Standard deviation without proposed feature set	0.461
t value	3.03
P value	0.0025
Significance Level (α)	0.05

4.4. Hybrid Classifier Performance Analysis

The proposed hybrid model was compared with the standard classifiers to evaluate the performance of the proposed system. Performance was compared based on parameters including Accuracy, Sensitivity, Specificity, F-measure, Matthews correlation coefficient (MCC), False Positive Rate (FPR), Negative Predictive Value (NPV), Precision, and False Negative Rate (FNR).

4.4.1. Result Analysis on dataset1:

The results are presented in Table (2).

- **Accuracy:** The proposed hybrid model consistently outperformed DBN, RF, SVM, and NB, with 5.9%, 3.7%, 8.18%, and 5.9% higher accuracies, respectively. This indicates a better overall predictive power.
- **Sensitivity:** The proposed hybrid model also exhibits sensitivity, which measures the ability of the model to identify positive instances correctly. Its value of 0.719 surpasses the sensitivities of all other models by 48.5%, 47.1%, 56.8%, and 52.7%, respectively.
- **Specificity:** Again, the hybrid model demonstrated strong performance in specificity, which measures the model's ability to correctly identify negative instances, with a value of 0.84, surpassing all other models by 13%, 15.4%, 14.1%, and 13.09%, respectively.
- The proposed hybrid model maintains a precision, F-measure, of 0.72 and high NPV, indicating its robustness in correctly classifying positive and negative instances.
- **MCC (Matthews Correlation Coefficient):** The proposed hybrid model consistently outperforms other models, indicating robust performance in balancing true positive and true negative rates.

4.4.2. Result analysis on dataset2:

The results are listed in Table (3).

- **Accuracy:** The proposed hybrid model showed 7.3%, 8%, 11%, and 6.5% higher accuracy than DBN, RF, SVM, and NB, respectively.
- **Sensitivity:** The proposed hybrid model demonstrated significantly higher sensitivities of 54.1%, 52.77%, 54.1%, and 48.6%, respectively, indicating a better ability to identify positive instances correctly.
- **Specificity:** It maintained a high specificity (0.832), showing its capability to identify negative instances accurately. It is 9.8%, 11.6%, 9.3%, and 10.9% higher than those of DBN, RF, SVM, and NB, respectively.
- **Precision and F-Measure:** The proposed hybrid model outperformed the other models regarding precision (0.729) and F-measure (0.72), indicating a better balance between precision and recall.
- **Mattews Correlation Coefficient (MCC):** The MCC for the proposed hybrid model in Dataset 2 was notably high (0.88), showing a strong correlation between the predicted and observed classifications.
- **False Negative Rate (FNR):** The FNR for the proposed hybrid model is considerably lower (0.28) than that of the other models, suggesting fewer actual positive instances being misclassified as negative.

4.4.3. Robustness:

The proposed hybrid model exhibits robustness in various respects. The proposed hybrid classifier outperformed the standardclassifiers in terms of the F-measure, showing the best balance between precision and recall for HAR. It also had low FPR and FNR, high MCC, and high accuracy and specificity, indicating its ability to identify true positives and negatives while avoiding false positives. This can reduce the cost of false predictions and provide reliable HAR for various applications in different fields. This model maintains a balance between sensitivity, specificity, precision, and accuracy, making it a robust choice for multiple applications.

TABLE 2: Performance comparison of the proposed hybrid classifier with DBN, RF, SVM, NB classifiers on Dataset 1.

Measures	Deep Belief Network	Random Forest	Support Vector Machine	Naïve Bayes	Proposed Hybrid Model
Accuracy	0.85	0.87	0.83	0.85	0.904
Sensitivity	0.37	0.38	0.31	0.34	0.719
Specificity	0.73	0.71	0.721	0.73	0.84
Precision	0.37	0.38	0.34	0.35	0.72
F- Measure	0.37	0.38	0.32	0.35	0.72
MCC	0.35	0.32	0.34	0.35	0.824
NPV	0.93	0.912	0.92	0.93	0.94
FPR	0.27	0.29	0.279	0.27	0.16
FNR	0.63	0.62	0.69	0.66	0.281

.

TABLE 3: Performance comparison of the proposed hybrid classifier with DBN, RF, SVM, NB classifiers on Dataset 2

Measures	Deep Belief Network	Random Forest	Support Vector Machine	Naïve Bayes	Proposed Hybrid Model
Accuracy	0.843	0.87	0.84	0.85	0.91
Sensitivity	0.33	0.34	0.33	0.37	0.72
Specificity	0.74	0.723	0.745	0.73	0.832
Precision	0.33	0.34	0.327	0.32	0.729
F- Measure	0.33	0.34	0.33	0.34	0.72
MCC	0.34	0.32	0.33	0.33	0.88
NPV	0.904	0.93	0.94	0.93	0.932
FPR	0.26	0.277	0.255	0.27	0.168
FNR	0.67	0.66	0.67	0.63	0.28

5. Conclusion

In this study, we introduced the Shape Index Local Ternary Pattern (SILTP) feature and a Hybrid Classifier Model to enhance human activity recognition system accuracy using videos in practical scenarios. Our investigation of diverse datasets highlights the robustness of the proposed model. Comparative analysis with individual classifiers—Deep Belief Network, Random Forest, SVM, and Naïve Bayes—highlighted the superior performance of our hybrid model

with SILTP. The proposed hybrid model consistently outperformed standalone classifiers across all key performance metrics, notably accuracy, sensitivity, specificity, and the Matthews Correlation Coefficient (MCC). Incorporating the SILTP within the HAR System, as supported by the t-test outcomes and the model's performance metrics, confirmed its pivotal role. With a substantial t-value and p-value below the significance threshold, rejecting the null hypothesis showed a significant accuracy improvement attributed to SILTP's inclusion of SILTP. In conclusion, our study emphasizes the importance of innovative features, such as SILTP and the power of hybrid models, in improving the accuracy and reliability of HAR systems. These findings offer promising prospects for practical application in this vital domain. Future research on HAR should focus on refining the hybrid model, exploring additional features, and practical implementation in surveillance and healthcare settings.

Author contributions

Milind Kamble: Conceptualization, Methodology, Software, Writing-Original draft preparation, Software, Validation.

Rajankumar Bichkar: Conceptualization, Methodology, and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from UAV- captured video sequences," *Pattern Recognit.*, vol. 100, p. 107140, 2020, doi: https://doi.org/10.1016/j.patcog.2019.107140.

[2] Z. A. Khan and W. Sohn, "Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1843–1850, 2011, doi: 10.1109/TCE.2011.6131162.

[3] Y. Chen, L. Yu, K. Ota, and M. Dong, "Robust activity recognition for aging society," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 6, pp. 1754–1764, 2018, doi: 10.1109/JBHI.2018.2819182.

[4] K. Viard, M. P. Fanti, G. Faraut, and J.-J. Lesage, "Human Activity Discovery and Recognition Using Probabilistic Finite-State Automata," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 2085–2096, 2020, doi: 10.1109/TASE.2020.2989226.

[5] J. Ye, G. J. Qi, N. Zhuang, H. Hu, and K. A. Hua, "Learning Compact Features for Human Activity Recognition Via Probabilistic First-Take-All," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 126–139, 2020, doi: 10.1109/TPAMI.2018.2874455.

[6] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for RGB-D human action recognition," *Pattern Recognit. Lett.*, vol. 50, pp. 139–148, 2014, doi: https://doi.org/10.1016/j.patrec.2014.03.024.

[7] V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human Activity Recognition Using a Dynamic Texture Based Method," 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:1280572

[8] S. Rahman, J. See, and C. C. Ho, "Exploiting textures for better action recognition in low-quality videos," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 74, 2017, doi: 10.1186/s13640-017-0221-2.

[9] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, and M. Lindén, "A Comparative Analysis of Hybrid Deep Learning Models for Human Activity Recognition," *Sensors*, vol. 20, 2020, doi: 10.3390/s20195707.

[10] M. Arshad *et al.*, "Hybrid Machine Learning Techniques to detect Real-Time Human Activity using UCI Dataset," *Int. J. Sci. Eng. Res.*, vol. 4, p. 6, 2021.

[11] D. K. Vishwakarma and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouette and cells," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6957–6965, 2015, doi: https://doi.org/10.1016/j.eswa.2015.04.039.

[12] R. R. Subramanian and V. Vasudevan, "A deep genetic algorithm for human activity recognition leveraging fog computing frameworks," *J. Vis. Commun. Image Represent.*, vol. 77, p. 103132, 2021, doi: https://doi.org/10.1016/j.jvcir.2021.103132.

[13] M. M. H. Shuvo, N. Ahmed, K. Nouduri, and K. Palaniappan, "A Hybrid Approach for Human Activity Recognition with Support Vector Machine and 1D Convolutional Neural Network," 2020. doi: 10.1109/AIPR50011.2020.9425332.

[14] T. R. Mim *et al.*, "GRU-INC: An inception-attention based approach using GRU for human activity recognition," *Expert Syst. Appl.*, vol. 216, p. 119419, 2023, doi: https://doi.org/10.1016/j.eswa.2022.119419.

[15] X. Yin, Z. Liu, D. Liu, and X. Ren, "A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-11880-8.

[16] R. J. Nemati and M. Y. Javed, "Fingerprint verification using filter-bank of Gabor and Log Gabor filters," in *2008 15th International Conference on Systems, Signals and Image Processing*, 2008, pp. 363–366. doi: 10.1109/IWSSIP.2008.4604442.

[17] J. Luo, "Feature Extraction and Recognition for Human Action Recognition," University of Tennessee, Knoxville, 2014.

[18] P. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Anal. Appl.*, vol. 23, no. 2, pp. 611–623, 2020, doi: 10.1007/s10044-019-00821-3.

[19] S. Megrhi, W. Mseddi, and A. Beghdadi, "Spatio-temporal SURF for Human Action Recognition," 2013. doi: 10.1007/978-3-319-03731-8_47.

[20] M. Niu, X. Mao, J. Liang, and B. Niu, "Object Tracking Based on Extended SURF and Particle

Filter," in *Intelligent Computing Theories and Technology*, 2013, pp. 649–657.

[21] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image Vis. Comput.*, vol. 10, no. 8, pp. 557–564, 1992, doi: https://doi.org/10.1016/0262-8856(92)90076-F.

[22] N. Alpaslan and K. Hanbay, "Multi-Scale Shape Index-Based Local Binary Patterns for Texture Classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 660–664, 2020, doi: 10.1109/LSP.2020.2987474.

[23] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," no. November, 2012, [Online]. Available: http://arxiv.org/abs/1212.0402

[24] H. Zhao, Z. Yan, L. Torresani, and A. Torralba, "{HACS}: Human Action Clips and Segments Dataset for Recognition and Temporal Localization," *arXiv Prepr. arXiv1712. 09374*, 2019.