A Hierarchical Framework for Video-Based Human Activity Recognition Using Body Part Interactions

Original Scientific Paper

Milind Kamble
Department of Electronics and Telecommunication Engg.,
G. H. Raisoni College of Engineering and Management
Pune, Maharashtra, India
Email: milind.kamble@vit.edu

Rajankumar S. Bichkar
Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology
Baramati, Maharashtra, India
Email: bichkar@yahoo.com

Abstract *–Human Activity Recognition (HAR) is an important field with diverse applications. However, video-based HAR is challenging because of various factors, such as noise, multiple people, and obscured body parts. Moreover, it is difficult to identify similar activities within and across classes. This study presents a novel approach that utilizes body region relationships as features and a two-level hierarchical model for classification to address these challenges. The proposed system uses a Hidden Markov Model (HMM) at the first level to model human activity, and similar activities are then grouped and classified using a Support Vector Machine (SVM) at the second level. The performance of the proposed system was evaluated on four datasets, with superior results observed for the KTH and Basic Kitchen Activity (BKA) datasets. Promising results were obtained for the HMDB-51 and UCF101 datasets. Improvements of 25%, 25%, 4%, 22%, 24%, and 30% in accuracy, recall, specificity, Precision, F1-score, and MCC, respectively, are achieved for the KTH dataset. On the BKA dataset, the second level of the system shows improvements of 8.6%, 8.6%, 0.85%, 8.2%, 8.4%, and 9.5% for the same metrics compared to the first level. These findings demonstrate the potential of the proposed two-level hierarchical system for human activity recognition applications.*

Keywords*: Human Activity Recognition(HAR);Hierarchical Model, Hidden Markov Model(HMM); Support Vector Machine(SVM)*

1. INTRODUCTION

Human activity recognition using video signals is a rapidly evolving field with applications in various domains, such as surveillance systems [1], human-computer interaction, and healthcare monitoring [2]. The ability to automatically analyze and understand human activities from video data has significant implications for improving safety, enhancing user experiences, and enabling intelligent systems [3]. In this study, we present a comprehensive approach for human activity recognition that leverages spatial-temporal features and a two-level hierarchical method, integrating hidden Markov models (HMM) [4] and support vector machines (SVM) to achieve accurate and robust activity classification.

Our research aims to develop a practical framework that captures the dynamic nature of human activities by extracting meaningful features from video frames and modeling the temporal dependencies between different activity states. We employ a step-by-step process from extracting spatial-temporal features to the final classification of human activities to accomplish this.

The first step in our approach involves extracting spatial-temporal features from the video frames. Recognizing that human appearance and motion play crucial roles in activity recognition, we begin by identifying humans in each frame using the method proposed in [5]. This technique combines a support vector machine (SVM) classifier and a histogram of oriented gradients (HOG) features to detect human regions accurately.

The SVM classifier effectively learns the decision boundaries between human and non-human regions, while the HOG features capture local shape and appearance information. By leveraging these techniques, we can robustly identify humans in the video frames, forming the basis for subsequent analysis.

To capture the spatial characteristics of human activities, we further divide the human region into six segments: the head region, the torso region, and the lower body part, as shown in Figure 3. This segmentation allows us to focus on specific body regions and extract region-specific features. For each body region, we compute the histogram of optical flow, which captures the motion information between consecutive video frames. The correlation between the histograms of optical flow for different body regions is then calculated, encoding the relationships and coordinated movements between these regions. This correlation-based approach results in a comprehensive feature vector that effectively represents the spatial interactions within the human body.

In addition to the spatial features, we recognize that temporal dynamics are essential for accurate activity recognition. We adopt a two-level hierarchical method to model and classify human activities accurately. In the first level, we employ hidden Markov models (HMMs) to capture the temporal dependencies and transitions between different activity states. HMMs are widely used in activity recognition tasks because they model sequential data well. HMMs can capture the underlying dynamics and temporal patterns by representing activities as a sequence of hidden states. In our approach, we train HMMs on labeled training sequences, allowing them to learn the emission and transition probabilities. We use these probabilities to recognize unseen activities in test sequences, as shown in Figure 1.

From the confusion matrix obtained through the HMM classification, similar activities are grouped based on their patterns and transitions. This grouping enhances the discriminability of the activity classes and provides a foundation for the second level of our hierarchical approach.

In the second level, as shown in Figure 2, we employ support vector machines (SVMs) for activity classification. SVMs are well-established supervised learning algorithms known for handling high-dimensional feature vectors and effectively separating data into different classes. The grouped activities from the first level serve as input to the SVM, allowing it to learn the discriminative patterns and decision boundaries between different activity classes. Training the SVM on the grouped activities can make fine-grained distinctions between similar activities and generalize well to unseen data. The SVM classification stage serves as a refinement step, further improving the accuracy and robustness of activity recognition.

To evaluate the effectiveness of our approach, we conducted extensive experiments on benchmark datasets in human activity recognition. We compared the performance of our framework against state-of-the-art methods, including those based on deep learning approaches. The results demonstrate that our approach achieves competitive or superior performance in terms of accuracy, robustness, and computational efficiency.

The key contribution of the proposed work are:

1. Two-level Hierarchical Structure: The paper introduces a novel hierarchical model to enhance system accuracy in human activity recognition, particularly in video-based scenarios.
2. Body Part Relationships: The research explores and leverages the interconnections among various body parts, enriching the feature set and leading to better recognition and classification of human activities.

The following is a general breakdown of the structure of this paper: Section 2 presents the methodology. Each component of the proposed system was elaborated from human detection in the video frame through activity modelling. The training of level one using HMM and level two using SVM is also described. The experiment and results analysis are described in Section 3. Finally, Section 4 concludes the study.

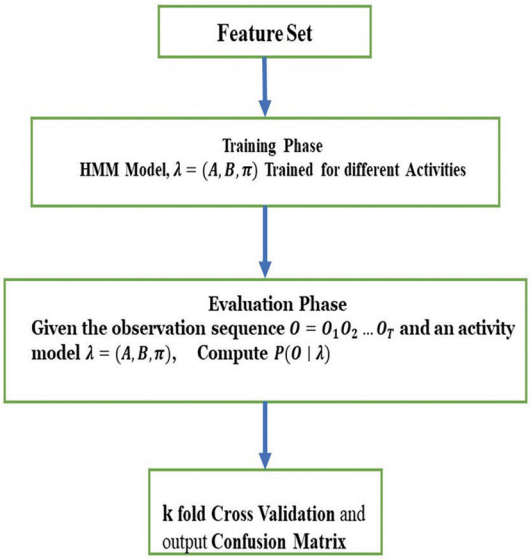


Fig. 1. Level 1, where activities are recognized based on HMM models

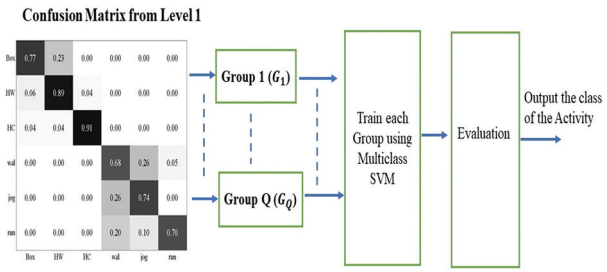


Fig. 2. At level 2, we group activities based on the confusion matrix of level 1 and use SVM as a classifier

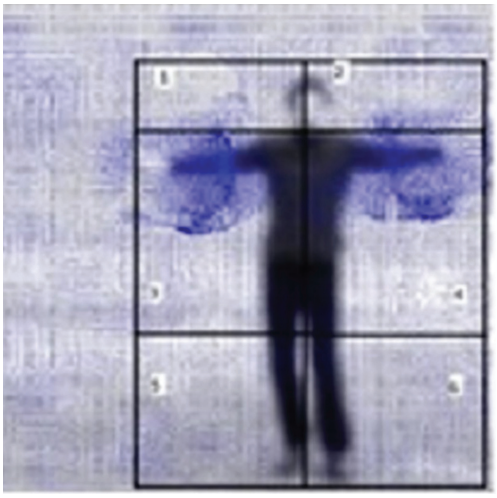


Fig. 3. The Region of Interest (ROI) in a video frame

1.1. RELATED WORK

Human activity recognition (HAR) using video signals is a rapidly evolving field, with numerous studies proposing innovative approaches and methodologies. This literature survey reviews influential works in this area, focusing on different aspects of human activity recognition.

In early research, spatial-temporal features played a significant role in human activity recognition, and researchers have explored various techniques to leverage these features effectively. In [6], which highlighted the inadequacy of direct 3D counterparts to commonly used 2D interest point detectors and proposed an alternative approach. In [7], the researcher presented a novel methodology that utilized local features to overcome the limitations of previous approaches of not detecting and localization of non-periodic activities. In [8], dense trajectories as features are used to enhance recognition accuracy. In [9], the authors proposed a novel method that uses local spatiotemporal color-depth information to enhance the robustness and accuracy of human action recognition in RGB-D videos. [10] proposed a sensing model capturing discriminative spatio-temporal features of human motion. In [11], they introduced a new paradigm for recognizing aggressive human behaviors, such as boxing action, using a fusion of Spatio Temporal Interest Point (STIP) and Histogram of Oriented Gradient (HoG) features. In [12], it aimed to improve the efficiency of optical flow feature extraction and explored spatio-temporal feature fusion methods.

To improve the recognition of human activities in real-world scenarios, researchers have proposed hierarchical models to address the challenges of activity recognition. In [13], the authors introduced the switching hidden semi-Markov model (S-HSMM). In [14], the authors applied the hierarchical hidden Markov model (HHMM) to capture the hierarchical nature of activities. In [15], they focused on detecting unstructured human activities in unstructured environments. In [16], authors proposed a spectral divisive clustering algorithm to extract hierarchies from tracklets. In [17], researchers developed a hierarchical sequence summarization approach for multi-temporal feature representations. In [18], deep hybrid models and active learning are combined in a continuous activity learning framework. [19] addressed group activity recognition with deep LSTM and 2-stage temporal models.

Recent advancements in HAR include the Progressive Skeleton-to-sensor Knowledge Distillation (PSKD) model [20] for wearable sensor-based HAR using smartwatch accelerometer data. In [21], the authors developed a methodology for automatic accident detection in surveillance videos to enhance safety and security systems. In [22], authors provided a comprehensive summary of deep neural network architectures in HAR, particularly convolutional neural networks (CNNs), offering insights into their advancements and applications.

Addressing key challenges in HAR, [23] introduced a rank-based fuzzy approach to capture transitional relationships between postures in temporal sequences. In [24], they introduced keypoint-MoSeq, an unsupervised machine-learning platform for identifying behavioral modules from keypoint data. In [25], the authors proposed SparseFormer, a method inspired by human sparse visual recognition, which focuses on key visual elements.

In conclusion, human activity recognition has witnessed significant advancements through exploring spatial-temporal features and hierarchical models. Recent research has addressed vital challenges and introduced innovative approaches for improved recognition, accuracy, and understanding of human actions. The integration of wearable sensors, deep neural networks, and human expertise shows promising directions for future research in this field.

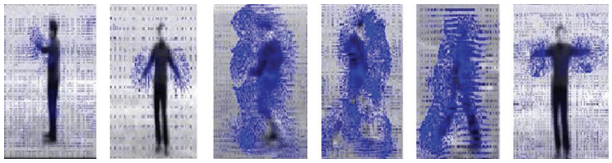
2. METHODOLOGY

The initial steps in the proposed HAR approach involve pre-processing the video data and segmenting the region where the action occurs. Next, human detection is accomplished using HOG features calculated using Equation 1 and SVM as a classifier. We identify an active region by drawing a bounding box around the detected human in the frame, as shown in Figure 3. We subsequently partition the active region into six sub-blocks organized in a 3×2 grid, as depicted in Figure 3, and features are extracted from each sub-block. Feature extraction focuses on identifying the connection between a pixel and its surrounding pixels in space and time. We achieve this by computing the relative motion between the pixels and the Optical Flow (OF) for each pixel.

$$|\mathrm{GradientofI}| = \sqrt{x^2 + y^2}$$
$$\emptyset = \tan^{-1}\frac{I_y}{I_x} \quad (1)$$

2.1. OPTICAL FLOW

To extract features, a rectangular Region of Interest (ROI) is defined by a bounding box, which is further subdivided into two columns and three rows, as depicted in Figure 3. The Optical Flow (OF) is computed for each subblock using the Lucas-Kanade method [26], as illustrated in Figure 4. This method is robust to changes in lighting and background clutter and is computationally efficient. According to the Lucas-Kanade technique, the displacement of the visual content between subsequent frames in the vicinity of point p is low and roughly constant; thus, all pixels within a window with a center of p are expected to follow the OF equation, as given in Equation 2.



(a) (b) (c) (d) (e) (f)

Fig. 4. Optical flow of person who carries out the subsequent action a) Boxing b) Hand Waving c) Jogging d) Running e) Walking f) Hand Clapping

$$\begin{aligned} &I_x(q_1)V_x + I_y(q_1)V_y = -I_t(q_1) \\ &I_x(q_2)V_x + I_y(q_2)V_y = -I_t(q_2) \\ &\vdots \\ &I_x(q_n)V_x + I_y(q_n)V_y = -I_t(q_n) \end{aligned} \tag{2}$$

where $I_x(q_i)$, $I_y(q_i)$, $I_t(q_i)$ are the partial derivatives of image I with respect to x, y, and t, respectively, evaluated at pixel q_{-i}, (Vx, Vy) is the local image flow vector and q_1, q_2,…q_n represent the pixels inside the window.

Equation 2 can be expressed in matrix form $Av=b$, as shown in equation 3.

$$A = \begin{bmatrix} I_x(q_1) & I_y(q_1) \\ I_x(q_2) & I_y(q_2) \\ \vdots & \\ I_x(q_n) & I_y(q_n) \end{bmatrix}, v = \begin{bmatrix} V_x \\ V_y \end{bmatrix}, \quad b = \begin{bmatrix} -I_t(q_1) \\ -I_t(q_2) \\ \vdots \\ -I_t(q_n) \end{bmatrix} \tag{3}$$

We obtain the solution to equation 3 using the least-square method.

$$v = (A^T A)^{-1} A^T b \tag{4}$$

where, v is a 2×1 dimension vector. The magnitude and phase components of v are given by Equation 5.

$$\begin{aligned} |v| &= \sqrt{Vx^2 + V_y^2} \\ \phi &= \tan^{-1}\frac{V_y}{V_x} \end{aligned} \tag{5}$$

where the components of v represent the velocities in x and y directions. From Equation 5, the optical flow histograms were calculated for each body segment using computer-vision techniques. Optical flow represents the motion of pixels between consecutive frames of a video, and the histogram summarizes the distribution of this motion in different directions and magnitudes. The algorithm can capture distinctive motion patterns associated with various human activities by calculating the optical flow histograms for each body segment. After calculating the histogram for all blocks in the ROI of a frame, they are concatenated to form a complete feature vector of size 1×54 for the video frame.

2.2. FEATURES BASED ON BODY PART INTERACTIONS

The proposed approach employs an autocorrelation strategy to capture the relationship and interaction between various body parts. The optical flow histograms for each body segment were cross-correlated with the histograms for all the other body segments to measure the similarity of their motion patterns. The resulting correlation matrix represents the inter-segment relationships and is used to build a hierarchical model that captures the overall structure of the human body and the relationships between its parts.

Cross-correlation was utilized to create feature vectors that link the six sub-blocks computed histograms. In addition, the cross-correlation determines the similarities between two series based on their separation from one another. Equation 6 defines the cross-correlation function for the two histograms $x[k]$ and $y[k]$, which correspond to the two body regions.

$$\begin{aligned} R_{xy} &= \sum_{m=-\infty}^{m=\infty} x[m]y[m-k] \\ R_{yx} &= \sum_{m=-\infty}^{m=\infty} y[m]x[m-k] \end{aligned} \tag{6}$$

where the value of k is $-\infty \le k \le \infty$. At $k=0$, $R_{xy}=R_{yx}$. The cross-correlation between one sub-block histogram and the remaining five sub-block histograms was calculated at $k=0$ to determine the similarity of each sub-histogram block to one another. The results are stored in a 1×6 vector. An identical approach was used for the remaining sub-blocks. A 1×6 vector is extracted from each sub-block, and this vector is combined with other vectors to form a 1×36 vector, which represents the feature vector of a single frame. The entire feature vector for a video sequence is created by applying this procedure to each video sequence frame and concatenating them. Once the feature vectors for all videos in a dataset are obtained, they are used to train the two-level proposed hierarchical model for HAR. In level one, each human activity is modelled using HMM. These feature vectors are used to train the HMM for each activity.

2.3. LEVEL ONE OF THE PROPOSED HIERARCHICAL MODEL

The proposed system utilizes the HMM to model each human action in the dataset at level one, as shown in Fig. 1. HMMs are generative probabilistic models that are used to generate hidden states based on observable data [26].

The proposed system employed an HMM in two stages. In the first stage, known as the training stage, each human activity is modelled by learning the model parameters $\lambda=(A,B,\pi)$ from the training data such that $P(O|\lambda)$, the probability of the observation sequence $O=O_1, O_2...O_T$ given the model λ, is maximized. The parameters of the HMM were A,B,π, where A represents the state transition probability matrix, B represents the observation symbol probability matrix, and π represents the initial state. The feature vectors derived from the video sequence frames create an observation sequence O.

In the second stage, referred to as the classification stage, the model that best captures the activity class for a particular observation sequence $O=O_1, O_2...O_T$ is chosen. We calculate the probability of the observation sequence given the activity model λ, denoted as $P(O|\lambda)$. An activity model that maximizes the probability $P(O|\lambda)$is selected.

Equation 7 illustrates the $P(O|\lambda)$ by adding the joint probabilities of all conceivable state sequences, q.

$$P(O|\lambda) = \sum_{allQ} P(O|Q,\lambda)P(Q|\lambda)$$
$$= \sum_{q_1 q_2 \ldots q_T} b_{q_1}(O_1)a_{q_1 q_2}b_{q_2}(O_2) \quad \ldots a_{q_{T-1} q_T}b_{q_T}(O_T) \tag{7}$$

where, $q_1,q_2 \ldots q_N$ represents the N number of states, where q_i represents the state at time step i, $a_{q_i q_j}$ represents the probability of transition from state q_i to state q_j, and $b_{q_j}(O_k)$ represents the probability of observing the symbol (O_k) in state q_j.

HMM uses a Gaussian Mixture Model(GMM) to identify human behavior in the proposed system. The GMM models the probability distribution of features extracted from video frames that capture human action, as shown in Equation 8, as a combination of M Gaussian distribution.

$$b_j(O) = \sum_{m=i}^{M} c_{jm}\mathfrak{N}[O, \mu_{jm}, U_{jm}], \quad 1 \le j \le N \tag{8}$$

Where O denotes the observation being modeled, c_{jm} represents the mixture weight for the m^{th} mixture in state j, and $\mathfrak{N}$ indicates the Gaussian density with the mean and covariance matrices μ_{jm} and U_{jm} associated with state j and the m_{th} mixture, respectively. Equation 9 represents the constraint to be satisfied by mixing the weights, c_{jm}.

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1 \le j \le N$$
$$c_{jm} \ge 0, \quad 1 \le j \le N, 1 \le m \le M \tag{9}$$

After extracting feature vectors from all video signals in the dataset, human activities were modelled using the HMM. We evaluated the HMM models using the test signals to generate a confusion matrix. Finally, based on the confusion matrix, we clustered similar human activities in the first level. We provided them as input to the second level, where we used SVM as a classifier.

2.4. LEVEL TWO OF THE PROPOSED HIERARCHICAL MODEL

Based on the output from level one, activities are now grouped together using the confusion matrix for level one. The confusion matrix helps us understand the performance of the classification model by showing the number of correct and incorrect predictions for each activity. By analyzing this matrix, we can identify patterns and similarities between activities, allowing us to group them based on their classification results. This grouping will help us gain better insights and improve the accuracy of our classification process. At level two, the activities inside the group G_i, created after the first level, were classified using the SVM classifier, as shown in Figure 2. The SVM creates a hyperplane to classify data. It seeks to identify the appropriate hyperplane for classifying the data into distinct groups. We selected the hyperplane to have the largest possible distance between it and the nearest data points for each class. The data points closest to the hyperplane, also known as the maximum margin hyperplane, are called the support vectors. The given training set consists of n data points of the form $(\bar{x}_1,\bar{y}_1),\cdots,(\bar{x}_n,\bar{y}_n)$, where $\bar{x}_i \in R^p$, p- dimensional input feature vector, and $\bar{y}_i$ is the target label for a binary classifier with values, {1,-1} are used for training the SVM. The values of $\bar{y}_i$, show the class to which point $\bar{x}_i$ belongs to. The SVM algorithm searches for the maximum-margin hyperplane that separates the data points belonging to the class, $\bar{y}_i$=1 from $\bar{y}_i$=-1. The hyperplane constraint, which requires data to be on the proper side of the margin, is represented by Equation 10.

$$\overline{y_i}(x_i \cdot w - b) \ge 1 \tag{10}$$

where b represents the offset of the hyperplane from the origin, and vector w depicts the orientation of the hyperplane. We can formulate the optimization problem as shown in Equation 11.

$$\text{Minimize} \parallel w \parallel^2 \text{ subject to } y_i(x_i \cdot w - b) \ge 1, \quad \text{for } i = 1, \ldots, n. \tag{11}$$

The classifier is determined by the w and b that solve the problem $x^{yields} \longrightarrow sgn(w \cdot x\text{-}b)$.

Let $R=\{R_1, R_2,\cdots,R_P\}$ be the P numbers of rows from the confusion matrix and let $G=\{G_1, G_2,\cdots,G_Q\}$ be the set of elements, where each element G_i is a set of similar activities.

Once we group identical activities from the output of the first level, we use the multi-class SVM, which is a collection of binary classifiers that distinguish between one of the classes and all the others (one-versus-all) or between every pair of classes (one-versus-one) to classify activities within each element G_i from the set G.

3. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Four datasets were used to evaluate the proposed method: KTH [27], SFB588 Basic Kitchen Activity (BKA) [28], HMDB-51 [29], and UCF101 [30]. This section presents the results and findings of our Human Activity Recognition (HAR) study. We analyze the performance of our proposed model through the following subtopics:

1. Feature Extraction and Cross-Validation: We discuss the feature extraction process and the robust cross-validation techniques to ensure reliable results.
2. Selection of HMM Parameters: We discuss the empirical section of the number of states and the number of Gaussian mixers used in GMM for modeling the activities.
3. Confusion Matrix from Level 1 and Level 2: We present and interpret the confusion matrices from Level 1 and Level 2 classifications, revealing improved prediction accuracy.
4. Effect of Feature Length on the Model: We discuss how varying feature lengths influence the model's accuracy and efficiency.
5. Comparison with Previous Work: Our model is comprehensively compared with existing approaches, showcasing advancements and improvements.
6. Evaluation on Real-World Video-Based Dataset: We assess our model's real-world applicability and performance using video-based data, offering valuable insights for practical implementation.
7. Statistical Analysis: Key metrics, including precision and recall, are presented for quantitative evaluation.

By thoroughly examining these subtopics, we aim to provide a comprehensive understanding of our model's strengths and limitations, contributing to the advancement of HAR.

3.1. FEATURE EXTRACTION AND CROSS-VALIDATION

We extracted features from each frame of the input video sequence at 25 fps. Each frame yielded a feature vector of size 1×36, which was obtained by concatenating the extracted features. The feature matrix was created by storing features from all videos in the dataset. The feature matrix is divided into two parts for training and testing the activity model. We trained the HMM using the first portion of the feature set, whereas we used the second portion for the evaluation and cross-validation. For k-fold cross-validation, the feature set was randomly divided into k equally sized sub-parts.

During each iteration of the k-fold cross-validation, one sub-part was reserved as validation data, and the remaining k-1 sub-parts were utilized for training. This operation was performed k times, using each sub-part as validation data precisely once. Finally, the results from each k-fold were averaged or combined to estimate the model's performance.

3.2. SELECTION OF HMM PARAMETERS

The first level of the system utilizes the HMM with the GMM. HMM represents the activity being performed, and GMMs model the probability distribution of the features extracted from the video sequence. The parameters of the HMM with GMM are A, B, π, number of states (N), and Gaussians in the mixture model (M). An empirical analysis was performed to determine the number of states N and Gaussian in the mixture model (M) by experimenting with different parameter values and evaluating the model's performance. The optimal values of N and M are determined by analyzing the system's accuracy for various combinations of N and M, as shown in Fig. 5.

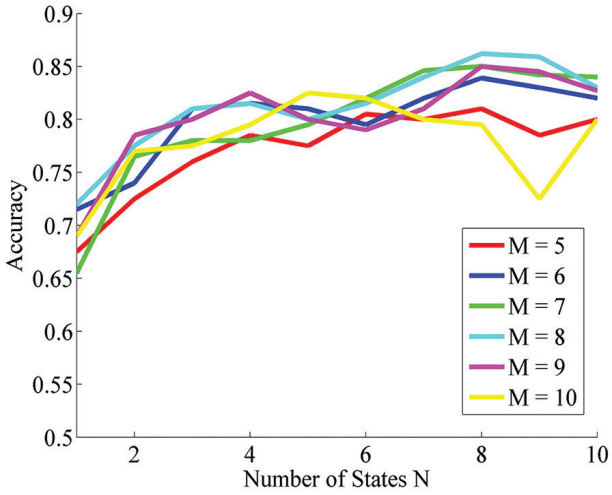
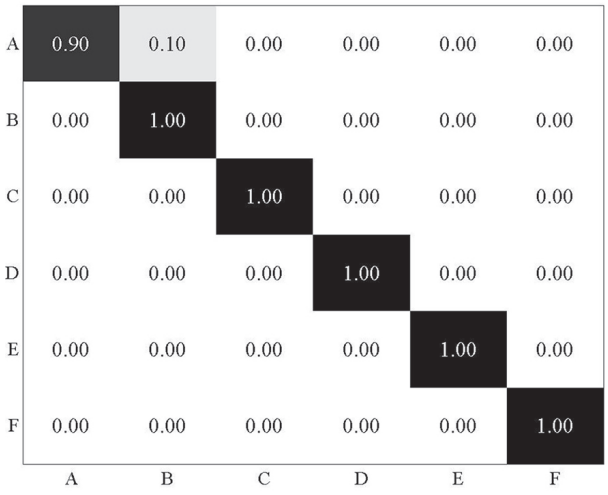


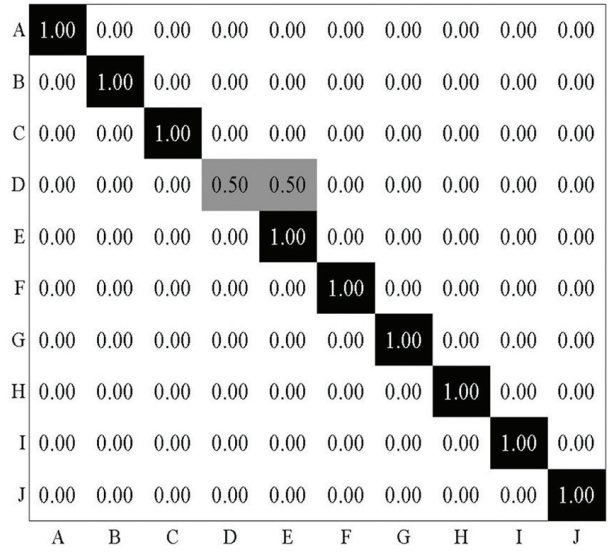
Fig. 5. Relationship between the accuracy of the system and the number of states.

3.3. CONFUSION MATRIX FROM LEVEL 1 AND LEVEL 2

Fig. 6 shows the confusion matrix from level one for the KTH and BKA datasets. In level one, activities are modelled using HMM, and then the probability of the observation sequence O given the model λ, P(O|λ) is calculated. Fig. 7 shows the confusion matrix from level two for the KTH and BKA datasets. In Level 2, the activities are grouped based on the confusion matrix obtained from Level one. For the KTH dataset, from the confusion matrix of Level one, as shown in Figure 6 (a), the activities related to the upper body's involvement were included in group one, and the activities involving the lower body were grouped in group two. Each group's activities were classified using a multi-class SVM at level two.

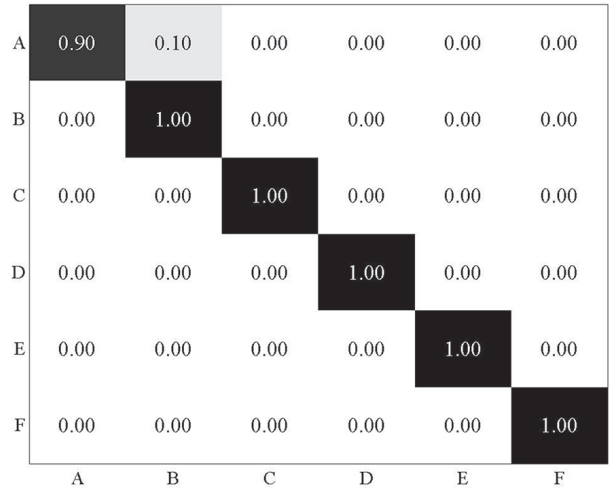


(a)

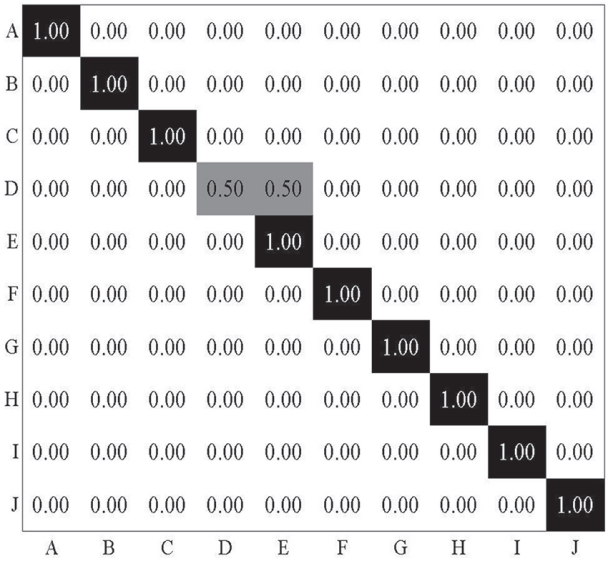


(b)

Fig. 6. Confusion matrix generated by Level 1 classification (a) KTH Dataset, where A: Boxing, B: Handwaving, C: Hand Clapping, D: Walking, E: Jogging, F: Running.(b) BKA Dataset, where 1. Chop, 2. Grate, 3. Mash, 4. Mill, 5. Pour, 6. Roll 7. Saw, 8. Slice, 9. Stir, 10. Sweep



(a)



(b)

Fig. 7. Confusion matrix generated by Level 2 classification. (a) KTH Dataset, where A: Boxing, B: Handwaving, C: Hand Clapping, D: Walking, E: Jogging, F: Running. (b) BKA Dataset, where 1. Chop, 2. Grate, 3. Mash, 4. Mill, 5. Pour, 6. Roll 7. Saw, 8. Slice, 9. Stir, 10. Sweep

3.4. EFFECT OF FEATURE-LENGTH ON THE PROPOSED MODEL

Features are created by concatenating the features obtained from one frame, and the number of frames required to generate a feature determines its length. The length of the feature vector affected the overall effectiveness of the system. Table 1 shows the number of frames for the accuracy of the proposed system for the KTH dataset. The proposed system's accuracy increased with the feature vector's length. The accuracy of the proposed method decreases when fewer frames are utilized for the feature vector.

Table 1. Relationship between the number of frames and the accuracy of the system on the KTH dataset

Number of Frames	Accuracy
50	87.3
75	91.2
90	95.6
100	98.3

3.5. COMPARISON WITH PREVIOUS WORK

The proposed system was compared with a group of previously published studies. The results reveal that the proposed approach increases the accuracy rate for the KTH dataset by 0.4% compared to [31], as shown in Table 2. For the BKA dataset, Table 3 compares the proposed method with activity recognition using the Histogram of Oriented Optical Flow (HOOF) and Histogram of Feature Flow (HoFF) [28]. The accuracy of the proposed method is 97.1%.

Table 2. Performance comparison of the proposed method and other approaches on the KTH Dataset

Method	KTH Dataset
Schuldt [27]	71.72
Liu and Shah [32]	94.16
Bregonzio et al.[33]	94.33
Lin [34]	95.77
Moussa [31]	97.89
Proposed Method	98.3

Table 3. Comparison between Histogram of Optical Flow (HoOF) and Histogram of Feature Flow (HoFF) on the BKA Dataset

BKA dataset I	HoOF	HoFF
Unit recog. [27]	96.7	96.6
Proposed Method	97.1	

3.6. EVALUATION OF THE PROPOSED WORK ON REAL-WORLD VIDEO-BASED DATASET

The experimental results obtained for various realistic datasets, namely UCF101 and HMDB-51, are listed in Table 4 and Table 5, respectively. The accuracies obtained for UCF101 and HMDB-51 are 84.6% and 69.2%, respectively.

Table 4. Performance comparison of the proposed method and other approaches on the UCF101 Dataset

Model	Accuracy
3-dimensional (3D) Residual ConvNet [35]	85.9
Multi-region Two-Stream R-CNN [36]	91.1
Optical Flow Guided Feature [37]	96
Two-stream+LSTM [38]	88.6
Proposed Method	84.6

Table 5. Performance comparison of the proposed method and other approaches on the HMDB-51 Dataset

Model	Accuracy
Two-stream I3D [39]	80.9
Multi-stream I3D [40]	80.92
TVNet + IDT [41]	72.6
Proposed Method	69.2

3.7. STATISTICAL ANALYSIS OF THE PROPOSED SYSTEM

Table 6 presents a statistical analysis of the proposed two-level HAR system, which evaluates its performance using six metrics: accuracy, recall, specificity, precision, F1-score, and Matthews Correlation Coefficient (MCC). The first level of the model employed an HMM for classification, whereas the second level used an SVM for further classification.

The second level of the proposed system significantly outperformed the first level when tested on the KTH dataset, with improvements of 25%, 25%, 4%, 22%, 24%, and 30% in accuracy, recall, specificity, Precision, F1-score, and MCC, respectively. On the BKA dataset, the second level of the system shows improvements of 8.6%, 8.6%, 0.85%, 8.2%, 8.4%, and 9.5% for the same metrics compared to the first level.

The proposed system exhibits high precision and recalls to effectively identify activities and avoid false positives and negatives. The F1 score is a widely used metric for evaluating the performance of classification models because it considers both precision and recall. The F1 scores for the KTH and BKA datasets were 98.33% and 94.67%, respectively, representing increases of 24% and 8.4% compared to the first level. The results suggest that the proposed hierarchical system improves the classification ability of the system.

Table 6. Performance metrics of Level 1 and Level 2 classifiers on KTH and BKA datasets.

Performance Metrics	KTH Dataset		Basic Kitchen Activities	
	Level 1	Level 2	Level 1	Level 2
Accuracy	0.78	0.98	0.87	0.95
Error	0.21	0.02	0.126	0.05
Recall	0.79	0.98	0.874	0.95
Specificity	0.96	0.99	0.99	0.99
Precision	0.80	0.98	0.89	0.97
False Positive Rate	0.043	0.003	0.014	0.0056
F1-score	0.78	0.98	0.87	0.95
Matthew's Correlation Coefficient	0.75	0.98	0.87	0.95

4. CONCLUSION

The proposed system consists of a two-level hierarchical framework for activity recognition, with the first level using HMM to categorize activities and the second using SVM for classification. The system can recognize similar activities, and experiments on four datasets showed that the hierarchical model outperformed the HMM and SVM applied separately, resulting in higher accuracy. The number of frames utilized for modeling affects the system's precision, with fewer frames resulting in poorer accuracy. The F1 scores for the KTH and BKA datasets were 98.33% and 94.67%, respectively, which is 24%and 8.4% rise compared to level one. The increase in the F1 score indicates that the proposed model has low false-positive and false-negative values, and the system correctly identifies the classes. However, the system's performance on datasets such as HMDB-51 and UCF101 was lower because of factors such as the short video duration, cluttered backgrounds, multiple people in the frame, and the occlusion of body parts. The proposed method relies on hand-crafted features and pre-processing of video signals. Future work could explore using deep learning algorithms for feature extraction from real-world video signals.

5. REFERENCES

[1] W. Lin, M. Sun, R. Poovandran, Z. Zhang, "Human Activity Recognition for Video Surveillance", Proceedings of the IEEE International Symposium on Circuits and Systems, Seattle, WA, USA, 18-21 May 2008, pp. 2737-2740.

[2] P. Khatiwada, A. Chatterjee and M. Subedi, "Automated Human Activity Recognition by Colliding Bodies Optimization (CBO) -based Optimal Feature Selection with RNN", Proceedings of the IEEE 23rd Int Conf on High-Performance Computing & Communications, Haikou, Hainan, China, 2021, pp. 1219-1228.

[3] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, K.-H. Choi, "A Review on Video-Based Human Activity Recognition", Computers, Vol. 2, No. 2, 2013, pp. 88–131.

[4] H. S. Mojidra, V. H. Borisagar, "Article: A Literature Survey on Human Activity Recognition via Hidden Markov Model", IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science 2012, No. 6, 2013, pp. 1–5.

[5] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20-25 June 2005, pp. 886-893.

[6] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features", Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15-16 October 2005, pp. 65-72.

[7] M. S. Ryoo, J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities", Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 1593-1600.

[8] H. Kataoka, Y. Aoki, K. Iwata, Y. Satoh, "Evaluation of Vision-Based Human Activity Recognition in Dense Trajectory Framework", Advances in Visual Computing, 2015, pp. 634–646.

[9] H. Zhang, L. E. Parker, "CoDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition From RGB-D Videos", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, No. 3, 2016, pp. 541-555.

[10] X. Luo, T. Liu, B. Shen, J. Hong, Q. Chen, H. Chen, "Human Daily Activity Recognition Using Ceiling Mounted PIR Sensors", in Proceedings of the 2nd International Conference on Advances in Mechanical Engineering and Industrial Informatics, Hangzhou, Zhejiang, China, 2016, pp. 872–877.

[11] S. Ghabri, W. Ouarda, A. M. Alimi, "Towards human behavior recognition based on spatiotemporal features and support vector machines", Proceedings of the Ninth International Conference on Machine Vision, 2017, Vol. 10341, pp. 67–72.

[12] Z. Xiao, X. Xu, H. Xing, F. Song, X. Wang, B. Zhao, "A federated learning system with enhanced feature extraction for human activity recognition", Knowledge-Based Systems, Vol. 229, 2021.

[13] T. V. Duong, H. H. Bui, D. Q. Phung, S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20-25 June2005, pp. 838-845.

[14] N. T. Nguyen, D. Q. Phung, S. Venkatesh, H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20-25 June 2005, pp. 955-960 Vol. 2.

[15] H. Zhang, W. Zhou, L. E. Parker, "Fuzzy Temporal Segmentation and Probabilistic Recognition of Continuous Human Daily Activities", IEEE Transactions on Human-Machine Systems, Vol. 45, No. 5, pp. 598–611, 2015.

[16] A. Gaidon, Z. Harchaoui, C. Schmid, "Activity representation with motion hierarchies", International Journal of Computer Vision, Vol. 107, No. 3, 2014, pp. 219–238.

[17] Y. Song, L.-P. Morency, R. Davis, "Action Recognition by Hierarchical Sequence Summarization",

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23-28 June 2013, pp. 3562-3569.

[18] M. Hasan, A. K. Roy-Chowdhury, "A Continuous Learning Framework for Activity Recognition Using Deep Hybrid Feature Models", IEEE Transactions on Multimedia, Vol. 17, No. 11, 2015, pp. 1909-1922.

[19] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1971-1980.

[20] J. Ni, A. H. Ngu, Y. Yan, "Progressive Cross-modal Knowledge Distillation for Human Action Recognition", Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5903–5912.

[21] S. W. Khan et al. "Anomaly Detection in Traffic Surveillance Videos Using Deep Learning", Sensors, Vol. 22, No. 17, 2022.

[22] S. Abbaspour, F. Fotouhi, A. Sedaghatbaf, H. Fotouhi, M. Vahabi, M. Linden, "A Comparative Analysis of Hybrid Deep Learning Models for Human Activity Recognition", Sensors, Vol. 20, No. 19, 2020, pp. 1–14.

[23] F. A. Dharejo et al. "FuzzyAct: A Fuzzy-Based Framework for Temporal Activity Recognition in IoT Applications Using RNN and 3D-DWT", IEEE Transactions on Fuzzy Systems, Vol. 30, No. 11, 2022, pp. 4578–4592.

[24] C. Weinreb et al. "Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics", bioRxiv, 2023.

[25] Z. Gao, L. Wang, M. Z. Shou, "SparseFormer : Sparse Visual Recognition via Limited Latent Tokens", arXiv:2304.03768, 2023.

[26] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol. 77, No. 2, 1989, pp. 257–286.

[27] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach", Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004, pp. 32-36

[28] H. Kuehne, D. Gehrig, T. Schultz, R. Stiefelhagen, "Online action recognition from sparse feature flow", Proceedings of the International Conference on Computer Vision Theory and Applications, Vol. 1, 2012, pp. 634–639.

[29] K. Soomro, A. R. Zamir, M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", arXiv:1212.0402, 2012.

[30] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, "HMDB: A large video database for human motion recognition", Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6-13 November 2011, pp. 2556–2563,.

[31] M. M. Moussa, E. Hamayed, M. B. Fayek, H. A. El Nemr, "An enhanced method for human action recognition", Journal of Advanced Research, Vol. 6, No. 2, 2015, pp. 163–169.

[32] J. Liu, M. Shah, "Learning human actions via information maximization", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23-28 June 2008.

[33] M. Bregonzio, T. Xiang, S. Gong, "Fusing appearance and distribution information of interest points for action recognition", Pattern Recognition, Vol. 45, No. 3, 2012, pp. 1220–1234.

[34] Z. Jiang, Z. Lin, L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 3, 2012, pp. 533–547.

[35] D. Tran, J. Ray, Z. Shou, S.-F. Chang, M. Paluri, "ConvNet Architecture Search for Spatiotemporal Feature Learning", arXiv:1708.05038, 2017.

[36] X. Peng, C. Schmid, "Multi-region two-stream R-CNN for Action Detection", Proceedings of the European Conference on Computer Vision, 2016, pp. 744–759.

[37] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, W. Zhang, "Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition", Proceedings of the I IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, pp. 1390–1399.

[38] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, "Beyond short snippets: Deep networks for video classification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7-12-June, 2015, pp. 4694–4702.

[39] K. Simonyan, A. Zisserman, "Two-stream convolutional networks for action recognition in videos", Advances in Neural Information Processing Systems, Vol. 1, No. January 2014, pp. 568–576.

[40] J. Hong, B. Cho, Y. W. Hong, H. Byun, "Contextual action cues from camera sensor for multi-stream action recognition", Sensors, Vol. 19, No. 6, pp. 1–13, 2019.

[41] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, J. Huang, "End-to-End Learning of Motion Representation for Video Understanding", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018, pp. 6016–6025.