Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053538

ARTICLE

Sports Events Recognition Using Multi Features and Deep Belief Network

Bayan Alabdullah[1], Muhammad Tayyab[2], Yahay AlQahtani[3], Naif Al Mudawi[4], Asaad Algarni[5], Ahmad Jalal[2] and Jeongmin Park[6,*]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

[2]Department of Computer Science, Air University, Islamabad, 44000, Pakistan

[3]Department of Computer Science, Applied College, King Khalid University, Abha, 61421, Saudi Arabia

[4]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran, 55461, Saudi Arabia

[5]Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, 91911, Saudi Arabia

[6]Department of Computer Engineering, Korea Polytechnic University, Siheung-si, Gyeonggi-do, 237, Republic of Korea

*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 03 May 2024 Accepted: 24 July 2024 Published: 15 October 2024

ABSTRACT

In the modern era of a growing population, it is arduous for humans to monitor every aspect of sports, events occurring around us, and scenarios or conditions. This recognition of different types of sports and events has increasingly incorporated the use of machine learning and artificial intelligence. This research focuses on detecting and recognizing events in sequential photos characterized by several factors, including the size, location, and position of people's body parts in those pictures, and the influence around those people. Common approaches utilized, here are feature descriptors such as MSER (Maximally Stable Extremal Regions), SIFT (Scale-Invariant Feature Transform), and DOF (degree of freedom) between the joint points are applied to the skeleton points. Moreover, for the same purposes, other features such as BRISK (Binary Robust Invariant Scalable Keypoints), ORB (Oriented FAST and Rotated BRIEF), and HOG (Histogram of Oriented Gradients) are applied on full body or silhouettes. The integration of these techniques increases the discriminative nature of characteristics retrieved in the identification process of the event, hence improving the efficiency and reliability of the entire procedure. These extracted features are passed to the early fusion and DBscan for feature fusion and optimization. Then deep belief, network is employed for recognition. Experimental results demonstrate a separate experiment's detection average recognition rate of 87% in the HMDB51 video database and 89% in the YouTube database, showing a better perspective than the current methods in sports and event identification.

KEYWORDS

Machine learning; silhouettes; extremal regions; joint points; scalable keypoints

1 Introduction

Understanding, recognizing, and locating spatiotemporal structures in succeeding video sequences is necessary to recognize events from unstructured data sources like movies and photographs. For the research community, there are several obstacles in the way of observing activities that take place inside frames because of contextual changes, body positions, illumination, and camera viewpoints. High motion and object interaction video sequences add to the difficulty of this work even further. Finding the important artifacts connected to a certain circumstance is the first step in comprehending the core of an event. It is therefore imperative to be alert for the existence of these essential components for the whole video clip.

Within computer vision, image processing, artificial intelligence, gait analysis, and human position estimation are essential fields. The necessity to increase productivity, make the best use of available resources, and distribute computing resources strategically drives these activities. The main objective is to improve the capabilities of a current system by making it capable of comprehending its surroundings and transforming it into a better set of instruments and approaches. By exact identification and monitoring of important skeletal reference points on the human body, such as the positions of joints, the movement of limbs, and the alignment of the spine, specialists can obtain useful information about the biomechanical factors that affect a person's walking style. The material offered can be used to create customized rehabilitation plans, measure and enhance sports performance, and determine the possibility of injuries.

Using two-dimensional (2D) photos or video sequences, one must estimate a person's three-dimensional (3D) pose. Inference of people's poses and motions in different situations is made possible by precise identification and monitoring of the positions and orientations of somatic joints, such as shoulders, elbows, wrists, hips, knees, and ankles. This finds important usages in virtual reality, human-computer interface, motion capture, surveillance, and action recognition.

We propose the use of evolutionary algorithms and machine learning techniques in our architecture. Our results show that the classification of sports, gait, and events can be accomplished with notable accuracy. The remaining paper is divided into different sections including a literature review in Section 2 and a proposed system architecture in Section 3. These results have been discussed in Section 4 in detail supported by general discussion in Section 5.

2 Related Work

Researchers have come up with creative answers to these problems. Several strategies have surfaced in the field of gesture event detection (GED) systems. For the analysis of sporting events, Kruk et al. [1] used body markers, recording acceleration and velocity to validate posture and identify events. Part mixture models were used by Li et al. [2] presented a video-based pose estimation method utilizing a new pose estimation network with a TCE block that performs Temporal Consistency Exploration at the feature level at a pyramid scale without requiring extra post-processing steps. Time-continuous features and Hidden Markov Models were integrated by Amft [3] to monitor dietary events using markers. Technology based on vision became more well-known; Using multi-class features, Jiang et al. [4] used a multi-modal feature technique and achieved a surprising precision. Meanwhile, by clustering the feature points, Liu et al. [5] suggested hierarchical clustering and multi-task learning for joint human event detection. Abbasnejad et al. [6] dealt with occlusion by extracting temporal and semantic data for event identification by identifying humans in unknown locations. In their convolutional neural network-based system, Seemanthini et al. [7] successfully combined machine and deep learning models. Meng et al. [8] also presented techniques for feature representation

that were only concerned with numerical and gradient characteristics. In addition, Ryoo et al. [9] investigated statistical data, co-occurrence matrices, and flow orientation histograms among other optical flow-based techniques. In Ma et al. [10], connecting the characteristics of the convolutional neural networks and Transformers, the Convolutional Transformer Network (CTN) proposes to use 3D convolutions for the extraction of tokens and depth-wise separable convolutional mapping of the multi-head self-attention layer to achieve further improvement of the performance of the fine-grained action recognition tasks with an increase in computational complexity by only several percent. These papers taken together demonstrate the range of methods and models used in the field of GED, including marker-based and vision-based approaches to tackle different problems with gesture and event detection.

3 Methods and Materials

Here we describe our work technique, which is focused on analyzing and forecasting a person's posture, movement, gait, and degree of freedom in order to predict their next moves. We propose a well-organized process that starts with the extraction of 30 frames per second from video data sets. Following pre-processing stages use methods based on Gaussian blur filtering and grayscale picture conversion, which successfully lower noise and remove distracting background elements with the rembg algorithm. To obtain a profound understanding of body movements, silhouette extraction, and human position estimate come after this methodical preparation. We apply algorithms such as BRISK, ORB, and HOG on skeleton points after extracting important feature points using MSER, SIFT, and DOF (Degree of Freedom) on whole bodies or silhouettes. We later used early fusion to fuss with these feature point values and optimized with DB Scan. Finally, we use DBN for categorization so that we may identify trends and precisely predict the person's next moves. The structural schematic of our suggested approach is shown in Fig. 1.

3.1 Image Pre-Processing

The essential first stage in machine learning, image processing, deep learning, and many other applications is picture pre-processing. The original material is videos that are framed. Subsequently, these frames are fuzzy. This image processing technique reduces the sharpness, features, or edges in a picture. To produce a smoother look, the pixel values in a neighborhood surrounding each pixel are averaged. We can quickly cut the computational cost and the unnecessary information for our system by doing this. In the domains of machine learning, image processing, and Al, the blurring method is called the Gaussian Blur Filter [11]. With this method, and with the help of Eq. (1), we apply a filter based on the following mathematical formulae to every pixel in the image:

$$G\,(x, y) = \left(\frac{1}{(2 * \pi * \sigma)}\right) * e^{\frac{(-(x^2+y^2))}{(2*\sigma^2)}} \tag{1}$$

G (x, y) is the kernel of the filter at x and y positions; π is the standard deviation of the Gaussian filter; e is the base of the natural algorithm with the values of 2.718289; and x, s are the coordinates of $(0, 0)$, $(-1, 0)$, $(0, -1)$. Applying it entails centering every pixel in the image using the Gaussian Kernel [12]. Grey scaling is the process of transforming an RGB image to a grey-scale image to lower the computing cost of 3-D images. Additionally, a well-known picture pre-processing is this method. This merely gives each pixel a greyscale value between 0–255 based on the intensity of the pixel before, converting a 3-dimensional image into a 2-dimensional image. A single channel or single-pixel layer makes processing this image simple. While RGB has several levels or channels. The average weightage

of RGB pixels is shown in Eq. (2).

$$\text{Grey} = 0.2988 * R + 0.5872 * G + 0.1137 * B \tag{2}$$

In the equation, the alphabets R, G, and B are the colored pixel values which are multiplied by their own weightage. NTPC (National Thermal Power Corporation Limited) claims this color dispersion. Perception by humans determines these values. Other ways exist to convert an RGB to greyscale, depending on the application requirement. Background noise in a picture still exists, so background reduction is necessary to separate the object of interest and improve the quality of input data for later machine learning tasks. Better performance and generalization result from the machine learning model concentrating on learning properties unique to the target object by removing unimportant background clutter and noise. Using this method, the algorithm chooses related pixels of the centered object and eliminates superfluous background information so that the object of interest is the main attraction. Fig. 2 shows the pre-processing technique step by step.

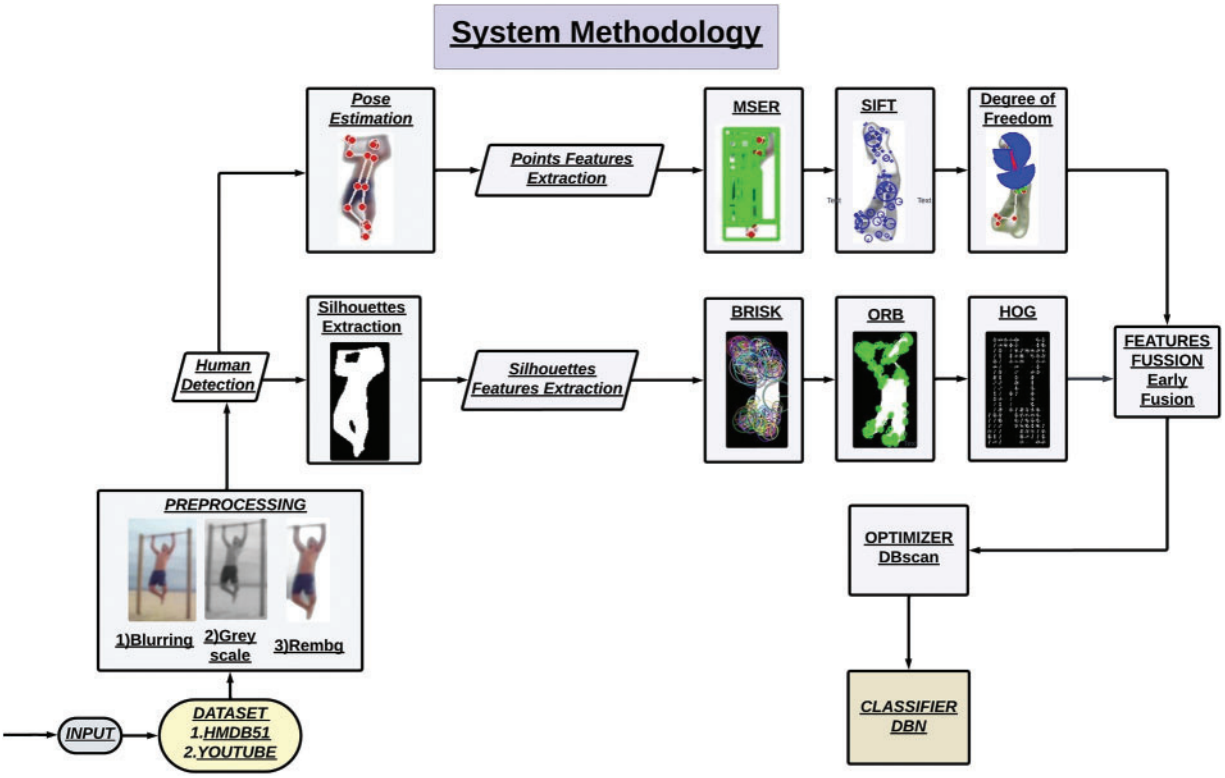


Figure 1: Proposed architecture for classification

Figure 2: Pre-processing goes like (a) raw frame extracted from video, (b) after reducing sharpness and smoothing edges, (c) converted 3-D image to 2-D, (d) background noise removed

3.2 Human Detection

Usually speaking, human detection is the act of recognizing and localizing people in a certain frame or picture. A whole-body picture is obtained by the first technique, silhouette extraction [13], and a human body skeleton is obtained by the second, position estimation. A binary mask is made to distinguish the human figure from the background. It divides the corresponding pixels of the person as foreground pixels (usually set to 1) and the remaining background pixels (usually set to 0). Many image processing systems require this step as a minimum. Another technique is to use thresholding to separate silhouettes from the background. This method sets a threshold value, as in Eq. (3), which will give one if the pixel value is higher than it and zero otherwise.

$$B.\mathrm{V} = \{255 \text{ if } \mathrm{I} \geq \mathrm{T}, 0 \text{ if } \mathrm{I} < \mathrm{T}\} \tag{3}$$

where $B.\mathrm{V}$ is the binary values, I is the intensity and T is the set threshold value. Another method consists of pose estimation. It is a method of human joint points detection. Pose estimation helps us in extracting gradient values of angles and distances which makes decision making easy. We have used it to detect the joints motion of the human body. Typically, these points are anatomical portions of the human body, such as the elbow, shoulders, wrist, hips, knees, and ankles [14] (in total 33 points). For more understanding use Eqs. (4) and (5). This determines the exact body posture, position, and orientation which is the main goal of the pose estimation. A common approach to representing human pose estimation mathematically can be represented as

$$\acute{\mathrm{K}} = f\,(I) = \{(x_1, y_1)\,, (x_2, y_2)\,, \ldots, (x_n y_n)\} \tag{4}$$

$$L = \frac{1}{N} \sum\nolimits_{i=1}^{N} Loss(K_i, \acute{\mathrm{K}}_i) \tag{5}$$

where function "f" maps an input image "I" to a set of predicted keypoints $\acute{\mathrm{K}}$, and the loss function L is used during training measures the discrepancy between the predicted points $\acute{\mathrm{K}}$. Fig. 3 here, Loss is a function (e.g., mean squared error).

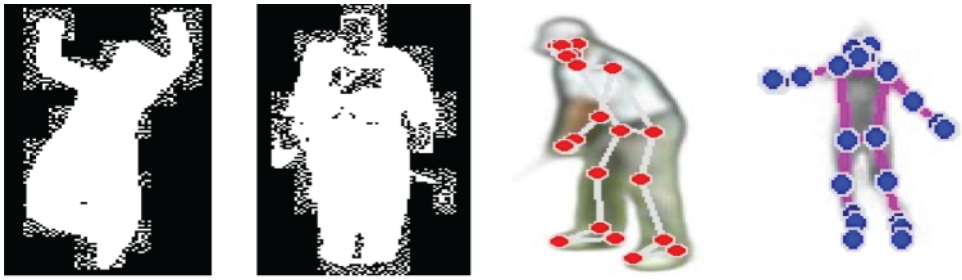


Figure 3: Binary images of human after silhouette extraction (a) HMDB51 (b) YouTube, Illustrations of human pose estimation by key points (c) HMDB51 (d) YouTube

3.3 Feature Extraction

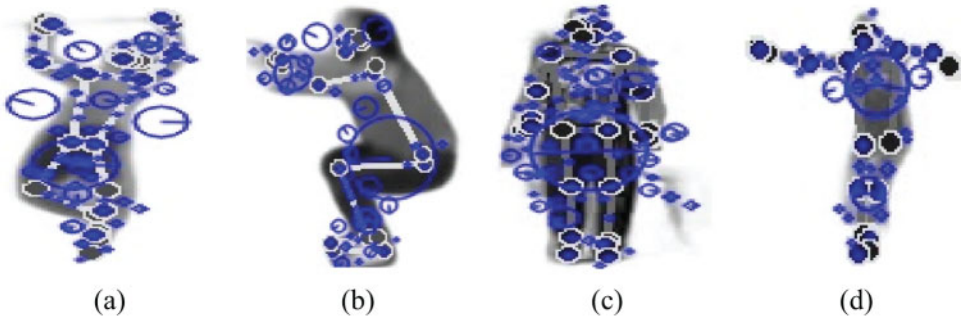
In numerous machine learning and computer vision problems, feature extraction—the process of converting raw data into a format appropriate for analysis and modeling—is an essential step. Feature extraction in the context of image processing is the process of recognizing and recording from images significant traits or patterns that may be fed into machine learning algorithms. Locating and removing significant patterns from pictures or video frames, such as edges, corners, or textures, is known as feature extraction. We employ six feature extraction techniques in our system: BRISK, ORB, and HOG for full-body (silhouette) analysis; SIFT, degree of freedom, and MSER for skeleton points (joint points). David Lowe created the robust computer vision technique known as SIFT (Scale-Invariant Feature Transform) to extract and match unique characteristics in images, enabling 3D reconstruction and object detection. It starts with Gaussian blurring a scale-space pyramid and finds scale-invariant key points by computing the Difference of Gaussians (DoG) [15]. This is also can be explained in Eqs. (6) and (7). Their placements are refined and orientations are assigned according to gradient directions by keypoint localization. A 128-dimensional vector encapsulating local gradient information describes each key point.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{6}$$

$$\theta(x,y) = \arctan\left(\frac{L(x+1,y) - L(x,y-1)}{L(x+1,y) - (x-1,y)}\right) \tag{7}$$

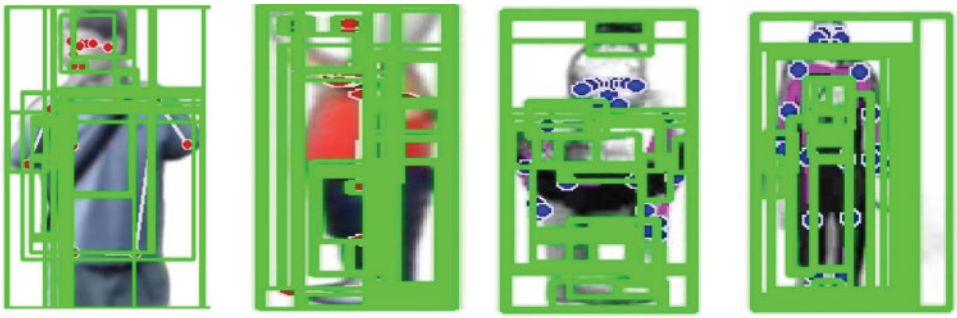
where L_x and L_y is the partial derivative of the image function with respect to x and y. The gradient orientation, represented by $\theta(x, y)$, indicates the direction in which the intensity changes most rapidly. Fig. 4 shows the SIFT points description.

MSER (Maximally Stable Extremal Regions) [16] is a computer vision algorithm designed for robust region detection in images, emphasizing stability across scales. It begins by thresholding image intensity to create a binary representation, followed by region growing to merge connected pixels into candidate regions. Stability analysis measures how these regions' areas change with varying thresholds, identifying maximally stable regions that maintain shape and size over a range of thresholds. Extremal region selection then isolates these stable regions, making MSER suitable for applications like object detection and image segmentation where consistent and robust region identification is crucial across diverse image conditions. Fig. 5 shows the examples of MESR applied on both datasets.



(a) (b) (c) (d)

Figure 4: Example of SIFT (a,b) HMDB51 and (c,d) YouTube



(a) (b) (c) (d)

Figure 5: MSER points extraction (a,b) HMDB51, (c,d) YouTube

Its mathematical representation is in the following Eq. (8), for intensity thresholding Eq. (9) region growth, and Eq. (10) stability analysis.

$$B(x,y) = \begin{cases} 1, & if\ I(x,y) \geq T \\ 0, & otherwise \end{cases} \tag{8}$$

$$R_i = \{(x,y) \in B | connected\ to\ R_i\ and\ I(x,y) \geq T \tag{9}$$

$$Stability\ (R_i) = \frac{area\ (R_i)}{area\ (R_i \oplus \Delta B)} \tag{10}$$

where $B(x,y)$ is the binary image after thresholding, T is the intensity threshold, R_i represents the ith region, and B is the binary image, $\oplus$ represents the morphological dilation operation, and ΔB is a small variation in the binary image.

The third approach is the degree of freedom in this we find the degree of freedom between the joint points and their distances. By this, we are adding a novel feature extraction technique for the automated extraction of geometric features from human pose images using computer vision and data analysis tools. The joint points serve as the basis for calculating various geometric parameters, including distances between specific joint pairs (e.g., shoulder to elbow) and angles formed by triplets of joints (e.g., shoulder-elbow-wrist). This approach enables quantitative assessment of body posture, movement dynamics, and biomechanical characteristics, making it valuable for applications in fields such as biomechanics research, sports science, physical therapy, and human motion analysis. Its further

technicalities are explained mathematically in Eq. (11) for finding angles between them. Fig. 6 is DOF of a skeleton.

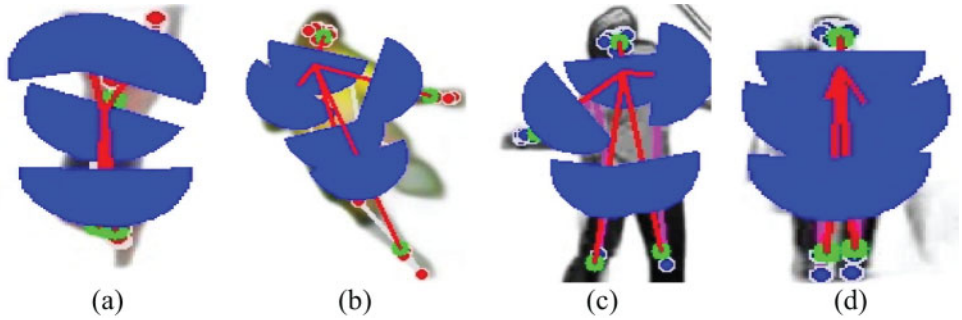


Figure 6: Finding degree points and distances (a,b) HMDB51, (c,d) YouTube

$$angle_{a-b-c} = cos^{-1}\left(\frac{dot\ product\ (\overrightarrow{AB},\overrightarrow{BC})}{\left|\overrightarrow{AB}\right| .|\overrightarrow{BC}|}\right) \tag{11}$$

where $\overrightarrow{AB}$ and $\overrightarrow{BC}$ are vectors formed by joints A–B and B–C, respectively, and dot product (·) represents the dot product of two vectors. Algorithm 1 shows the working of DOF feature.

Algorithm 1: Features points using novel system

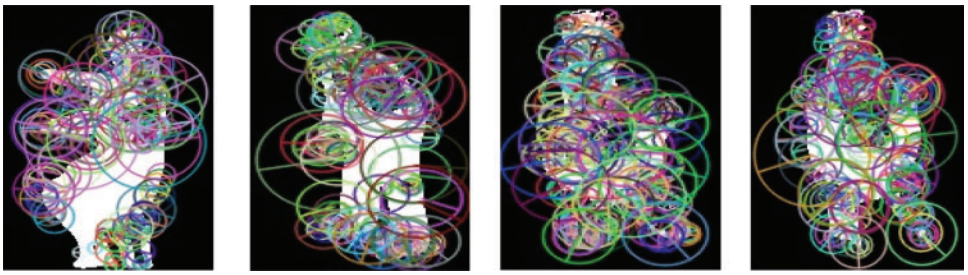
Define geometric_pose_analysis(**joint_points**) function:
Initialize features dictionary
For each specific joint pair:
Calculate distance between joint points
Store distance in features dictionary
For each joint triplet:
Calculate angle formed by joint points
Store angle in features dictionary
Input Image
Load image
Define joint points extracted from pose estimation
Perform **geometric pose analysis** using
Convert features to **Data Frame**
Append features_df to all_features DataFrame
Output:
features: feature points vector for the input image
Images
Excel file
End

The other three features that are extracted are BRISK, HOG and ORB. These features points are applied on the full human body (silhouettes). BRISK (Binary Robust Invariant Scalable Key-points) [17] is a computer vision algorithm renowned for its robust feature detection and description capabilities. Utilizing binary patterns for image patch representation, BRISK excels in scenarios with varying lighting conditions and viewpoints, ensuring robustness and reliability. BRISK constructs a

scale pyramid and detects key points using a variant of FAST, refining their positions for accuracy. It generates binary descriptors based on local intensity patterns around key points, facilitating efficient image matching and recognition in computer vision tasks. BRISK's binary descriptors enable efficient feature matching and resilience to noise, enhancing its usability in tasks such as matching of images, recognition of an object, and reconstruction of 3D objects, where robustness, accuracy, speed, and working are paramount. Its further processing is explained in Eq. (12).

$$HD\,(X,\,Y) = \sum_{i=1}^{N} x_i \oplus y_i = \sum_{i=1}^{n} (x_i, y_i) \tag{12}$$

where $b(x_i, y_i)$ denotes bit inequality, in Eq. (15), x_i and y_i are the i^{th} bits of the descriptors X and Y, respectively. Its examples of both datasets are can be seen in Fig. 7.



(a) (b) (c) (d)

Figure 7: BRISK features points (a,b) HMDB51, (c,d) YouTube

In computer vision, ORB (Oriented FAST and Rotated BRIEF) [18] is a feature detection and description algorithm. The keypoint description is made robust by BRIEF (Binary Robust Independent Elementary Features), while keypoint detection is made efficient by FAST (Features from Accelerated Segment Test). ORB is made to respond quickly, effectively, and robustly to noise and vision changes. It computes the orientations of key points and produces binary descriptors to allow quick matching between key points in various pictures, hence achieving orientation invariance. The work is described in the equation below. Eq. (13) shows ORB descriptor-Patch's moment's definition, while Eq. (14) finds the Orientation of the mass of the patch

$$m_{pq} = \sum_{x,y} x^p y^p I(x, y) \tag{13}$$

$$\theta = atan2(m_{01}, m_{02}) \tag{14}$$

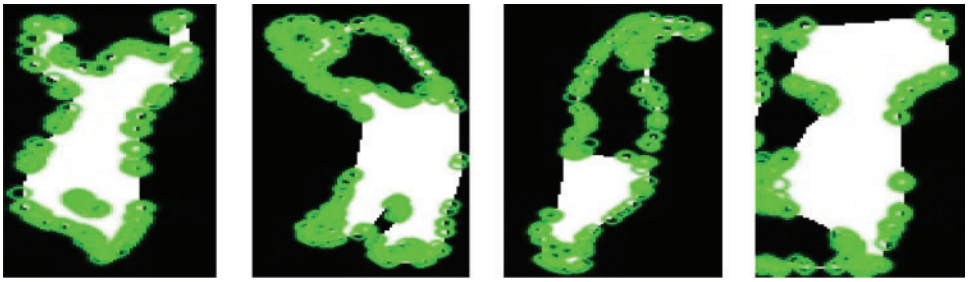
where, $I\,(x, y)$ represents the intensity of a pixel at coordinates (x, y), as extracted in Fig. 8.

HOG (Histogram of Oriented Gradients) [19] is a feature descriptor used in computer vision for object detection and recognition. It works by calculating gradient orientations and magnitudes in localized regions of an image, constructing histograms of these orientations within cells, normalizing histograms in blocks for robustness, and concatenating normalized histograms to generate a comprehensive descriptor capturing texture and edge information. HOG's ability to represent local image patterns makes it valuable for tasks like pedestrian detection, object recognition, and scene understanding, where detailed visual features are essential for accurate analysis and classification. Its working can be accomplished by using Eqs. (15) and (16) and can be seen in Fig. 9.

$$||x, y|| = \sqrt{(H_{x,y} - H_{x+1,y})^2 + (H_{x,y} - H_{x,y+1})^2} \tag{15}$$

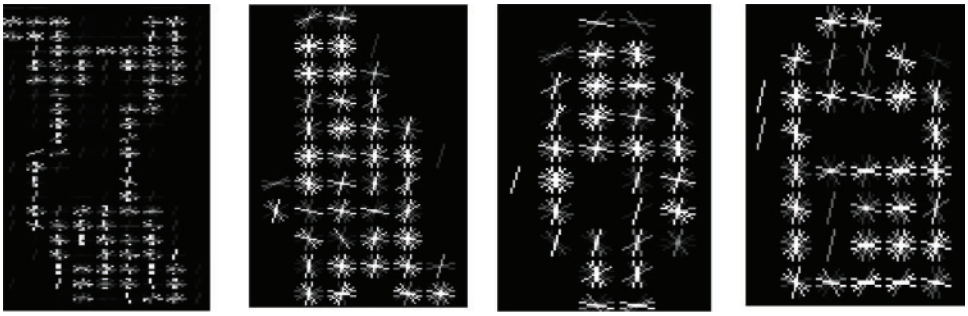
$$\theta\,(x,y) = atan2[\left(H_{x,y} - H_{x+1,y}\right)\,.\,(H_{x,y+1} - H_{x,y})] \tag{16}$$

where H_x is Sobel horizontal and H_y is Sobel vertical directions.



(a) (b) (c) (d)

Figure 8: ORB features extracted (a,b) HMDB51, (c,d) YouTube



(a) (b) (c) (d)

Figure 9: HOG gradients points (a,b) HMDB51, (c,d) YouTube

3.4 Features Fusion

Feature fusion refers to the process of combining multiple types or sources of features into a single representation. This fusion can occur at different stages of a machine learning pipeline, such as combining features from different sensors, modalities, or algorithms. The goal is to enhance the overall feature representation, potentially improving the performance of downstream tasks like optimization, classification, or clustering. The technique that we have used is Early fusion, also known as feature-level fusion, which occurs at the beginning of the pipeline, where raw or preprocessed features from different sources are combined before any further processing or analysis. The combined feature vector contains information from all sources, providing a comprehensive representation of the input data for subsequent learning algorithms. Early fusion can be beneficial when features from different sources complement each other and capture diverse aspects of the data, leading to improved model performance. For more understanding, we take Eq. (17). Rocchio's algorithm can also be applied to merge the vectors of the same feature spaces into a single vector. Fig. 10 is the comparison of graphs of both human detection techniques.

$$q_m = \alpha q_o + \beta_1 |I_r| \sum\nolimits_{ij \in I_r} I_j ij - \gamma_1 |I_{nr}| \sum\nolimits_{ij \in I_{nr}} I_j \tag{17}$$

In this equation, q_m represents the modified query, q_o is the original query, I_r denotes the collection of relevant documents or images, and I_{nr} is the collection of non-relevant documents. The weights are denoted by α, β_1, and γ_1.

Figure 10: feature fusion graphical representation, (a) full body points (silhouettes), (b) pose estimation points

3.5 Optimization

In the context of machine learning and data analysis, Feature optimization refers to the process of selecting the most relevant and informative features from the available data. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an algorithm that groups data points based on their density, making it useful for optimization tasks. By selecting parameters like ε (epsilon) and minPts (minimum number of points), DBSCAN may find clusters of densely packed data points and isolate them from noise or outliersOverall, DBSCAN's density-based technique offers a solid way to optimize parameters or features by grouping comparable solutions and spotting outliers in the data. For finding the core points Eqs. (18) and (19) can be used.

$$N_{\varepsilon}(x_i) = \left\{x_j \in X | dist\left(x_i, x_j\right) \leq \varepsilon\right\}, X = \{X_1, X_2, \ldots, X_n\}, \tag{18}$$

$$C = \{X_i \in X \,||\, N_{\varepsilon}(x_i) \geq MinPts\} \tag{19}$$

where $x_i \in X||N_{\varepsilon}(x_j)$ and $x_j \in C$; x_i is the epsilon neighborhood of x_j and x_j is the core point. Fig. 11 shows the clustered graph of optimization.

3.6 Events Classifier: Deep Belief Network

A classifier in machine learning is a model that is trained using labeled training data to identify relationships between data points and class labels. A Deep Belief Network (DBN) is a neural network with multiple layers that consists mostly of probabilistic generative models, namely Restricted Boltzmann Machines (RBMs). RBMs in a DBN capture hierarchical features from the input data through unsupervised learning, generating a hierarchy of hidden layers that learn increasingly abstract representations [20]. The process of learning layer by layer allows the network to uncover complex patterns and structures in the input. Their utilization of hierarchical representation learning enables them to extract significant characteristics, resulting in enhanced performance in comparison to shallow

networks. In general, DBNs utilize deep learning principles to independently acquire complex data representations and attain high precision in different machine learning tasks. This method can be described as consisting of two steps: the energy function, represented by Eq. (20), and the conditional probability, represented by Eqs. (21) and (22). Fig. 12 is the internal skeleton of the classifier.

$$E(v,h) = -\sum_{i=1}^{N_v}\sum_{j=1}^{N_h} W_{ij}v_i h_j - \sum_{i=1}^{N_v} b_i v_i - \sum_{j=1}^{N_h} c_j h_j \tag{20}$$

$$P\left(h_j = 1|v\right) = \frac{1}{1+\exp(-(b_{j+}\sum_{j-1}^{N_h} W_{ij}v_i))} \tag{21}$$

$$P\left(v_i = 1|h\right) = \frac{1}{1+\exp(-(c_{i+}\sum_{j-1}^{N_h} W_{ij}h_j))} \tag{22}$$

where v is the visible layer, h is the hidden layer, W_{ij} is the weight between visible unit d hidden unit j, b_i is the bias of visible unit i, c is the bias of hidden unit j, N_v and N_h are the number of units in the visible and hidden layers, respectively.

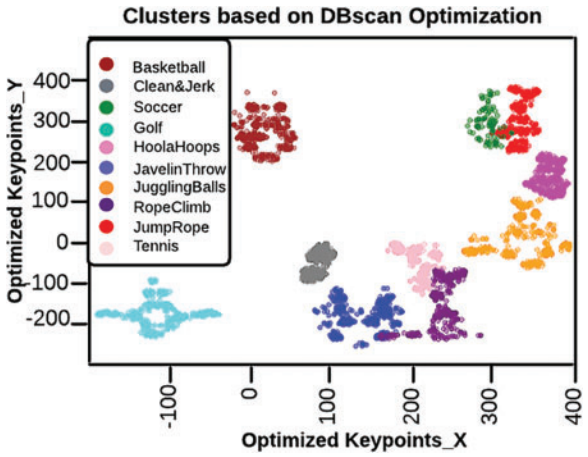


Figure 11: Structure of DBscan optimizer's clusters

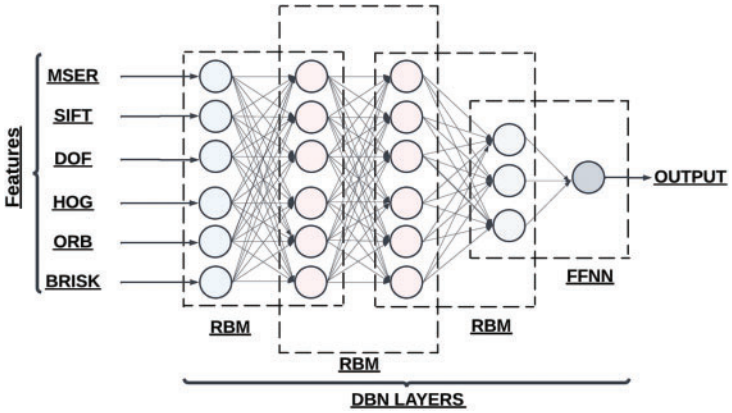


Figure 12: Exemplary structure of DBN classifier

4 Experimental Results

In this part, the experimental results of our approach and its distinctions from the other research. Its accuracy on these datasets in comparison to previous works is remarkable.

4.1 Datasets Description

For our research we have used 2 widely used benchmark datasets in the field of action recognition which are HMDB51 dataset, short for "Human Motion Database 51" and 51 shows the number of classes such as walking, running, jumping, and more. With a total of 6766 videos randomly distributed in classes. The other dataset is YouTube-8M which has more than thousands of classes, and the count of videos goes to millions. Common formats include AVI, MP4, or similar video formats. In our study, we conducted experiments using Python on a computer system equipped with an Intel Core i5 CPU and 8 GB of RAM. We used Eq. (23) for precision, Eq. (24) for recall, and Eq. (25) for accuracy to assess the performance of our recognition model. The findings showed an 86.7% accuracy rate on HMDB51 dataset and 88.9% on YouTube dataset. Figs. 13 and 14 are the confusion matrices, Tables 1 and 2 present the comparison of each class with their precision, accuracy and recall. Table 3 is comparison table. We have taken 10 classes from each dataset.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{23}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{24}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{25}$$

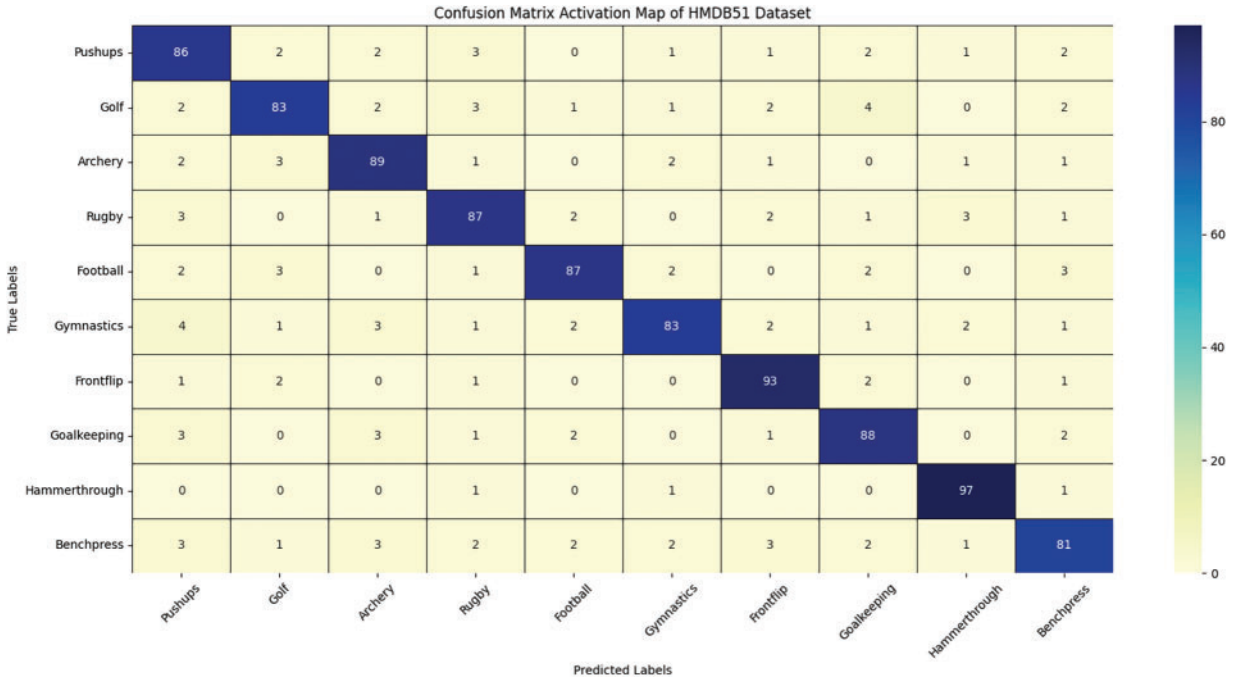


Figure 13: Confusion matrix of proposed method on HMDB51 dataset

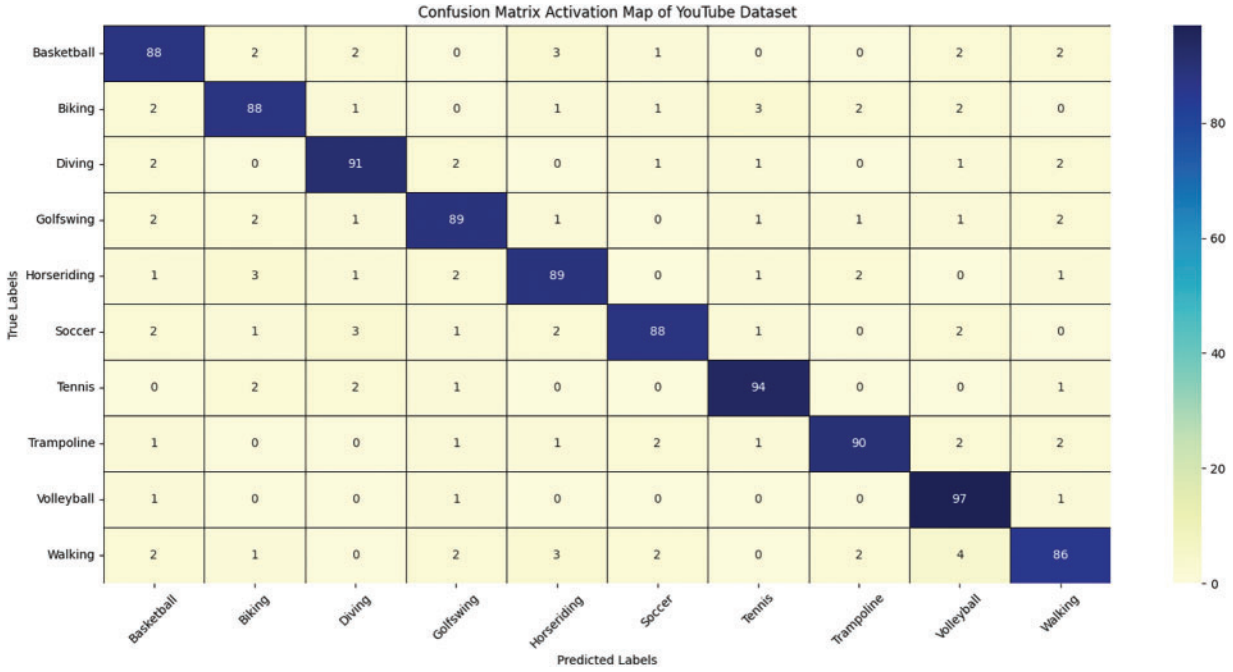


Figure 14: Confusion matrix of proposed method on YouTube dataset

Table 1: Accuracy table of each class of HMDB51

Events	Accuracy	Precision	Recall
Push up	0.86	0.77	0.94
Golf	0.83	0.83	0.81
Archery	0.89	0.80	0.84
Rugby	0.87	0.84	0.79
Foot ball	0.87	0.90	0.89
Gymnastics	0.83	0.82	0.83
Front flip	0.93	0.82	0.88
Goal keeping	0.88	0.88	0.86
Hammer throw	0.97	0.92	0.94
Bench press	0.81	0.83	0.88
Mean	**0.87**	**0.841**	**0.866**

Table 2: Accuracy table of each class of YouTube

Events	Accuracy	Precision	Recall
Basket ball	0.88	0.90	0.93
Biking	0.88	0.85	0.78
Diving	0.91	0.87	0.83
Golf swing	0.89	0.89	0.77
Horse riding	0.89	0.91	0.88
Soccer	0.88	0.87	0.84
Tennis	0.94	0.90	0.90
Trampoline	0.90	0.88	0.85
Volley ball	0.97	0.85	0.92
Walking	0.86	0.92	0.87
Mean	**0.89**	**0.88**	**0.857**

Table 3: Comparison of proposed model with state-of-the-art methods

Methods	HMDB51 (%)	Methods	YouTube (%)
Dey et al. [21]	81.9	Meng et al. [8]	82.5
Nasir et al. [22]	82.5	De Siva [24]	84.2
Alsawadi et al. [23]	83.7	Akhter et al. [25]	85.0
Proposed system	**87**	**Proposed system**	**89**

4.2 Failure Cases

In preprocessing when we remove the background the system does not subtract the noise which is present in the foreground. This is because the pixels are connected or linked with each other so the system takes it as a main subject of interest. This thing leads to the confusion of system and disturbs the accuracy of the whole system some examples of system failure are given in Fig. 15.

(a) (b)

Figure 15: Failure cases, (a) back ground removed properly, (b) case failure

5 Conclusion and Future Works

In this research work, we present an approach based on features and optimal classifiers for gait analysis and event recognition. We developed a better detection model to describe human postures and compare them frame by frame to see changes and categorize events. Experimental results demonstrate a noteworthy accuracy of our suggested approach. Adding energy and angular points in the future will increase the efficacy of our suggested features. In addition, we will use our model to interpret drone photos and analyze crowds in public spaces during the more complicated event.

Acknowledgement: The authors are thankful to Princess Nourah bint Abdulrahman University Researchers Supporting Project, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) Program (IITP-2024-RS-2022-00156326) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This research was supported by the Deanship of Scientific Research at Najran University, under the Research Group Funding program grant code (NU/RG/SERC/13/30).

Author Contributions: Study conception and design: Muhammad Tayyab and Bayan Alabdullah; data collection: Naif Al Mudawi and Yahay AlQahtani; analysis and interpretation of results: Muhammad Tayyab and Asaad Algarni; draft manuscript preparation: Muhammad Tayyab, Ahmad Jalal and Jeongmin Park. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] E. van der Kruk and M. M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review," *Eur. J. Sport Sci.*, vol. 18, no. 6, pp. 806–819, 2018. doi: 10.1080/17461391.2018.1463397.

[2] Y. Li, K. Li, X. Wang, and R. Y. D. Xu, "Exploring temporal consistency for human pose estimation in videos," *Pattern Recognit.*, vol. 103, no. 1–3, 2020, Art. no. 107258. doi: 10.1016/j.patcog.2020.107258.

[3] O. Amft, "Adaptive activity spotting based on event rates," in *IEEE Int. Conf. Sensor Netw., Ubiquitous, Trustworthy Comput.*, Newport Beach, CA, USA, 2010, pp. 169–176.

[4] Y. -G. Jiang, Q. Dai, T. Mei, Y. Rui, and S. -F. Chang, "Super fast event recognition in internet videos," *IEEE Trans. Multimed.*, vol. 17, no. 8, pp. 1174–1186, Aug. 2015. doi: 10.1109/TMM.2015.2436813.

[5] A. Liu, Y. Su, W. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 102–114, 2017. doi: 10.1109/TPAMI.2016.2537337.

[6] I. Abbasnejad, S. Sridharan, S. Denman, C. Fookes, and S. Lucey, "Complex event detection using joint max margin and semantic features," in *Int. Conf. Digit. Image Comput.: Tech. Appl.*, Gold Coast, QLD, Australia, 2016, pp. 1–6.

[7] K. Seemanthini, S. Manjunath, G. Srinivasa, B. Kiran, and P. Sowmyasree, "A cognitive semantic-based approach for human event detection in videos, smart innovation," *Syst. Technol.*, vol. 165, no. 1, pp. 243–253, 2020.

[8] Q. Meng, H. Zhu, W. Zhang, X. Piao, and A. Zhang, "Action recognition using form and motion modalities," *ACM Trans. Multimed., ACM Trans. Multimed. Comput., Commun. Appl.*, vol. 13, no. 16, pp. 1–6, Apr. 2020. doi: 10.1145/3350840.

[9] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE 12th Int. Conf. Comp. Vis. Conf.*, Kyoto, Japan, 2009, pp. 1593–1600.

[10] Y. Ma, R. Wang, M. Zong, W. Ji, Y. Wang and B. Ye, "Convolutional transformer network for fine-grained action recognition," *Neurocomput.*, vol. 569, no. 45, 2024, Art. no. 127027. doi: 10.1016/j.neucom.2023.127027.

[11] E. Pulfer, "Different approaches to blurring digital images and their effect on facial detection," in *Computer Science and Computer Engineering Undergraduate Honors Thesis*, Fayetteville: University of Arkansas, Fersity of Arkansas, 2019, pp. 6–18.

[12] I. Hayder, A. Younis, and H. Younis, "Digital image enhancement gray scale images in frequency domain," *J. Phys. Conf. Ser.*, Jul. 2019, pp. 1–6.

[13] S. Sulaiman, A. Hussain, N. Tahir, S. Samad, and M. Mustafa, "Human silhouette extraction using background modeling and subtraction techniques," *Inf. Technol. J.*, vol. 7, no. 4, pp. 155–159, 2008.

[14] N. U. Khan and W. Wan, "A review of human pose estimation from single image," in *2018 Int. Conf. Audio, Lang. Image Process. (ICALIP)*, Shanghai, China, 2018, pp. 230–236.

[15] L. Tang, S. Ma, X. Ma, and H. You, "Research on image matching of improved SIFT algorithm based on stability factor and feature descriptor simplification," Appl. Sci., vol. 12, no. 17, pp. 1–9, 2022.

[16] L. Lian, G. Li, H. Wang, H. Tian, and S. Xu, "Corresponding feature extraction algorithm between infrared and visible images using MSER," *J. Electron. Inf. Technol.*, vol. 33, no. 7, pp. 1625–1631, 2011.

[17] Y. Liu, H. Zhang, H. Guo, and N. Xiong, "A FAST-BRISK feature detector with depth information," in *School of Information Engineering*, Nanchang, China: East China Jiaotong University, 2018, vol. 18, no. 11, pp. 1–6.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 2564–2571.

[19] W. Zhou, S. Gao, L. Zhang, and X. Lou, "Histogram of oriented gradients feature extraction from raw bayer pattern images," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 67, no. 5, pp. 946–950, 2020.

[20] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *2018 IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 1–6.

[21] A. Dey, S. Biswas, and D. -N. Le, "Workout action recognition in video streams using an attention driven residual DC-GRU network," *Comput. Mater. Contin.*, vol. 79, no. 2, pp. 3067–3087, May 2024. doi: 10.32604/cmc.2024.049512.

[22] I. M. Nasir, M. Raza, J. H. Shah, M. A. Khan, and A. Rehman, "Human action recognition using machine learning in uncontrolled environment," in *2021 1st Int. Conf. Artif. Intell. Data Anal. (CAIDA)*, Riyadh, Saudi Arabia, 2021, pp. 182–187.

[23] M. S. Alsawadi and M. Rio, "Skeleton-split framework using spatial temporal graph convolutional networks for action recognition," in *4th Int. Conf. Bio-Eng. Smart Technol. (BioSMART)*, Paris/Créteil, France, 2021, pp. 1–5.

[24] N. H. T. M. De Siva and R. A. H. M. Rupasingha, "Classifying YouTube videos based on their quality: A comparative study of seven machine learning algorithms," in *IEEE 17th Int. Conf. Ind. Inf. Syst. (ICIIS)*, Peradeniya, Sri Lanka, 2023, pp. 251–256.

[25] I. Akhter, A. Jalal, and K. Kim, "Adaptive pose estimation for gait event detection using context-aware model and hierarchical optimization," *J. Electr. Eng. Technol.*, vol. 16, no. 1, pp. 2721–27292, 2020.