Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.053547

ARTICLE

# Robust Human Interaction Recognition Using Extended Kalman Filter

**Tanvir Fatima Naik Bukht**[1], **Abdulwahab Alazeb**[2], **Naif Al Mudawi**[2], **Bayan Alabdullah**[3], **Khaled Alnowaiser**[4], **Ahmad Jalal**[1] **and Hui Liu**[5,*]

[1]Faculty of Computing & Artificial Intelligence, Air University, Islamabad, 44000, Pakistan

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

[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11671, Saudi Arabia

[4]Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia

[5]Cognitive Systems Lab, University of Bremen, Bremen, 28359, Germany

*Corresponding Author: Hui Liu. Email: hui.liu@uni-bremen.de

Received: 03 May 2024 Accepted: 04 August 2024 Published: 18 November 2024

**ABSTRACT**

In the field of computer vision and pattern recognition, knowledge based on images of human activity has gained popularity as a research topic. Activity recognition is the process of determining human behavior based on an image. We implemented an Extended Kalman filter to create an activity recognition system here. The proposed method applies an HSI color transformation in its initial stages to improve the clarity of the frame of the image. To minimize noise, we use Gaussian filters. Extraction of silhouette using the statistical method. We use Binary Robust Invariant Scalable Keypoints (BRISK) and SIFT for feature extraction. The next step is to perform feature discrimination using Gray Wolf. After that, the features are input into the Extended Kalman filter and classified into relevant human activities according to their definitive characteristics. The experimental procedure uses the SUB-Interaction and HMDB51 datasets to a 0.88% and 0.86% recognition rate.

**KEYWORDS**

Pattern recognition; geometric features; activity recognition; full-body texture

## 1 Introduction

Human activity recognition (HAR) from video frames is a complicated problem in computer vision due to its scarcity of data and noisy background. However, many static images are available on the web which can be very useful for developing and efficient image-based activity recognition methods aimed at analyzing visual content [1]. Most of the research on HAR has been done in the Extended Kalman filter [2]. Thus, HAR is gradually becoming essential in many applications such as biophysics as human–machine interaction, surveillance, environmental intelligence, living assistance, and human–computer interaction [3].

HAR is a complex computer vision task that tries to understand human behavior through visual data analysis including pictures and movies. The objective is to detect challenging human-human interactions, but, this is hard to achieve due to issues like viewpoint fluctuation, occlusion, ambiguity, data inefficiency, and interaction difficulty. Therefore, the execution and usage of most of the approaches for HAR are rather scoped. Future HAR developments might enable improved video/image surveillance, better human–machine interaction, and safer intelligent modes of transport [4]. The objective is to detect challenging human-human interactions, which is hard to achieve due to issues like viewpoint fluctuation, occlusion, ambiguity, data inefficiency, and interaction difficulty. We have created a human-activity recognition model to recognize complex human activities from SUB-Interaction and HMDB51 datasets. The dispute over HAR research is still on, The proposed system comprises the following key contributions:

- The HSI transformation should be used in the proposed system in combination with the Gaussian filter to enhance the quality of the frames and extract crucial information.
- The statistical approach is applied to retrieve the silhouettes from already processed frames accurately.
- Sophisticated techniques to extract features such as BRISK and SIFT are used for relevant feature extraction from the obtained silhouettes.
- To differentiate between the aspects, a Gray Wolf approach is adopted. This approach improves the feature separation process demonstrated by the EKF method.

Collectively, these contributions are significant for the target system which helps it deliver better and more precise results in the given problem domain. The framework mentioned is light and thus usable in any edge device, almost doing operations real-time and consuming minimal processing overheads. Therefore, it is best for real-time applications. Since it is small, the integration process is simple and has no impact on performance and functionality.

The article is organized as follows: Section 2 is a literature review, while Section 3 is the framework which consists of preprocessing, silhouette extraction, feature extraction and discrimination. The findings of the HAR system experiments and comparisons are discussed in the following Section 4. This information proves the hypothesis and provides the proper data.

## 2 Related Work

Activity Recognition (HAR) has been a prominent research area, and both traditional and machine learning-based approaches have been explored in this field. This section reviews the relevant literature on traditional approaches and their limitations, followed by an overview of machine learning-based techniques for HAR.

### *2.1 Traditional Human Activity Recognition Approaches*

Classical HAR approaches feature rule-based systems, hand-crafted feature extraction, and shallow machine-learning algorithms. Rule-based systems operate using a set of predefined rules and logical conditions to detect particular events. While they enjoy expert knowledge and domain expertise, limitations include rule specification, difficulties in dealing with complex activities, and inadaptability to new or changing activities [5].

Handcrafted feature extraction is about manually creating features from sensor data that represent activity patterns. Although these methods are popular, they are domain-dependent and may not incorporate enough information for complex activity recognition [6].

Also, one can use shallow learning algorithms in classification automation systems such as the k-nearest neighbours (KNN) method, support vector machines (SVM) algorithm, and decision trees algorithm. Nevertheless, they use human-crafted features and may find it challenging to represent high-dimensionality and complex activity patterns, resulting in overfitting or underfitting in the case of various datasets [7,8].

### *2.2 Machine Learning-Based Human Activity Recognition Approaches*

ML models in particular are considered to be the beacon of hope within the scope of HAR as conventional strategies do not solve all those problems. Types of deep learning such as CNNs and RNNs have emerged as some of the best in identifying spatial and temporal characteristics from data collected by sensors [9]. CNNs are very efficient in spatial pattern extraction, while RNNs, including long shortterm memory (LSTM) networks and gated recurrent units (GRUs), adequately model the temporal dependencies [10]. The performance of these models in complex activity recognition and classification accuracy is auspicious [11].

The transfer learning methodologies are becoming popular in HAR where the pretrained models installed on massive datasets are used. Refinements of these models on smaller activity recognition datasets lead to better performance, less training time and overcoming the problem of scarce labelled data [12,13].

The existing sensor data fusion techniques such as fusing data from several sensors has shown potential in enhancing the accuracy of the activity recognition. Fusion could be done at different levels, which are the feature level fusion, decision level fusion, and the sensor level fusion. Combining data from different sensor like accelerometers, gyroscopes and cameras has enhanced the performance and the generalization capability [14,15].

## 3 Proposed Model

Interactive recognition based on images is more complex than video action detection as a result of the limited data, blurred background of the images, as well as vague attributes such as Similar looking the semantics of interactions may be varied and training data includes large difficult-to-code amounts of human interaction and Complex interaction marked-up data. This research aims to propose a novel image-based approach for human interaction recognition (HIR), which outperforms the existing approaches. The suggested HHI system's architectural flow is depicted in Fig. 1.

The architecture flow diagram explains the system's numerous components and their linkages. EKF is an outstanding estimation algorithm especially when integrating image features. While traditional Kalman Filters cannot make non-linear dynamics and measurement equations, EKF has that capability and therefore is used in the context of image-based problems. The use of the EKF to linearize non-linear functions allows for state estimation and tracking based on image features even in noise and uncertainties. The EKF is similar to the traditional Kalman Filters but better suited for systems that yield non-linear functions through the Jacobian linearization techniques. In HAR, the EKF is employed to estimate and track activity states as well as performing various measurements from sensors. This approach of linearizing non-linear functions with the help of the Jacobian matrix defines the true dynamics of the system as was approximated by the EKF. This makes it possible to estimate and track states accurately, irrespective of non-linearity. The fact that the EKF can handle non-linear relationships between the measured or derived data and the activities improves the HAR systems' accuracy and performance. Conclusively, by addressing the non-linearities through linearization in certain sections of the EKF, state estimation and tracking in HAR tasks is enhanced.

This functionality enables the EKF to deliver resilient and precise estimates, including human activity detection and image-based navigation. When the power of EKF is combined with image features, researchers and practitioners can open the door to a new world in computer vision, image processing, robotics, augmented reality, and autonomous systems. The EKF's ability to deal with non-linear systems and its combination with image features make it an essential instrument for state estimation in image-based research and applications.

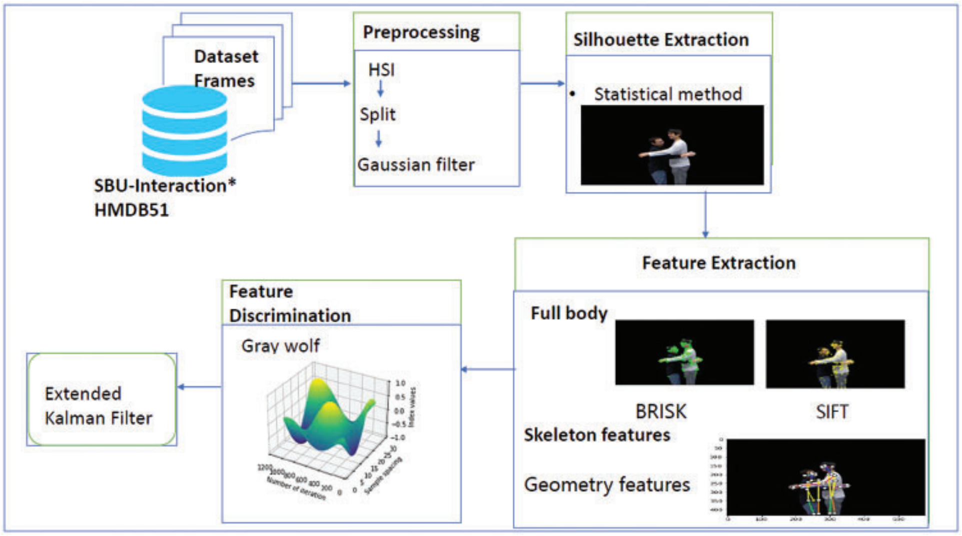


**Figure 1:** Architecture flow of our proposed HAR system involves preprocessing data frames, silhouette and feature extraction, feature discrimination, and classification with XGBoost

### *3.1 Preprocessing*

To avoid any incorrect human behavior estimation, some noise-removing preprocessing techniques need to be included in input frames. This operation is essential in the accurate extraction of important features. In this study, we propose a simplified preprocessing approach comprising two main steps: To solve this problem the following two steps are suggested: (a) color space transformation and (b) channel selection and the use of a Gaussian filter. The choice of preprocessing where the image data is converted into another color space, the number of channels is decreased and applying a Gaussian filter, is based on the enhancement of the process. These techniques are intended to mature recognized information further and exclude noise to get higher-quality data for subsequent analysis or recognition. Color space transformation enhances the feature extraction process and facilitates the extraction of more relevant features than previously extracted features; channel selection/elimination leads to the exclusion of irrelevant channels that would otherwise increase the system's complexity and noise. The Gaussian filter is used to blur the image, but while doing so, it tends to maintain better details of the image and suppresses noise. These methods enhance the representations and classification steps, providing the best predictive outcomes regarding computer vision and image-processing plans.

### *3.2 Optimal Channel Selection*

We conduct a color space transform on an input video frame, which changes it into another color representation. The transformed color space divides the frame into channels, capturing certain image features. If the transformation produces a three-channel image, they are represented by $C_1(x, y)$ as red color, $C_2(x, y)$ as green color intensity, and $C_3(x, y)$ blue color. To normalize these channels, one has to divide them by the sum of the three channels. The overlying channels (C1, C2, C3) create a certain type of importance to each channel. This is particularly useful in the case of images with relatively low contrast because normalization will bring the intensities to the same range and hence we can go to the next step. It eliminates biases that initial distributions may cause and helps in bringing an algorithm to the right convergence. Normalizing the channels makes the images easily comparable because it equalizes the data, which helps to put the best features of the image in order and thus increases the reliability of the whole process. The transformed representation is calculated using the following equations:

$$V\_1 = 1/2\,(C_{1(x,y)} - C_{2(x,y)} + C_{1(x,y)} - C_{3(x,y)}) \tag{1}$$

$$G_1 = \frac{\left\{C_{1(x,y)} - C_{2(x,y)}\right\}}{2} \tag{2}$$

$$G_2 = \left(C_{1(x,y)} - C_{3(x,y)}\right)\left(C_{2(x,y)} - C_{3(x,y)}\right) \tag{3}$$

Fig. 2 shows the transformed input image channels. Each channel is seen as a grayscale image, giving insights into the source image's color data, intensity, and other characteristics. This figure acts as a demonstration of a multi-modal theme in the area of image analysis and processing. It provides descriptions of the special characteristics shown by each channel and the role of the changed color space in image analysis.

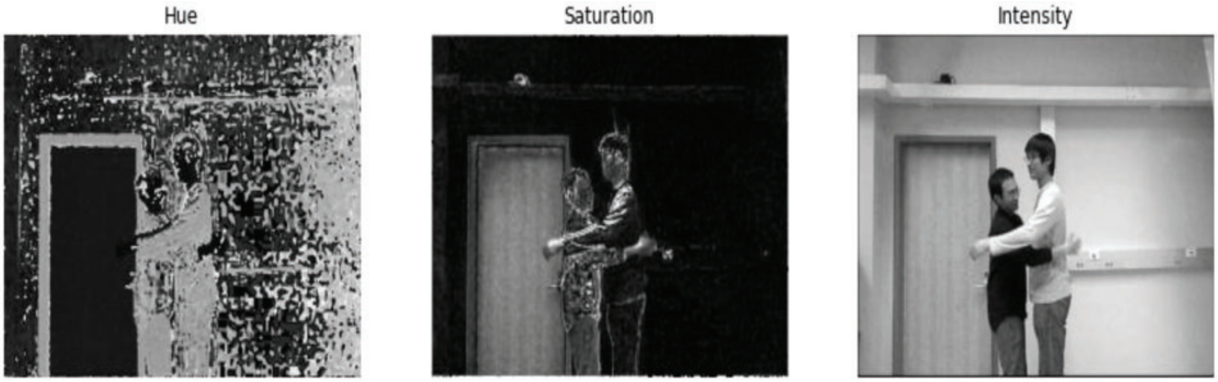


**Figure 2:** HSI Transformed representation of the original image

### *3.3 Gaussian Filter*

Among image processing and computer vision applications, the Gaussian filter is a widely used method to blur or smoother images [16,17]. It relies on the Gaussian function, which weighs the pixels in an image. Our research uses a Gaussian filter in conjunction with transformed color space. The main reason for using a Gaussian filter is that it can reduce the noise effectively, has linear shift invariance, spatial localization and filter size selectable, and is optimal among the other filters.

The Gaussian filter is mathematically defined as follows:

$$G(x, y) = \frac{1}{(2\pi\sigma\,\hat{}\,2)} e- ^{\frac{((x\hat{}2+y\hat{}2))}{(2\sigma\hat{}2))}} \tag{4}$$

Here, $G(x, y)$ is given as Eq. (4) which is the Gaussian filter, $e$ is the mathematical constant known as Euler's number, and $\sigma$ is the standard deviation of the Gaussian function. The width of the Gaussian function is defined by the standard deviation which controls the smoothening or blurring of the image indeed. The outcome of the Gaussian filtering works is presented in the next figure, Fig. 3.

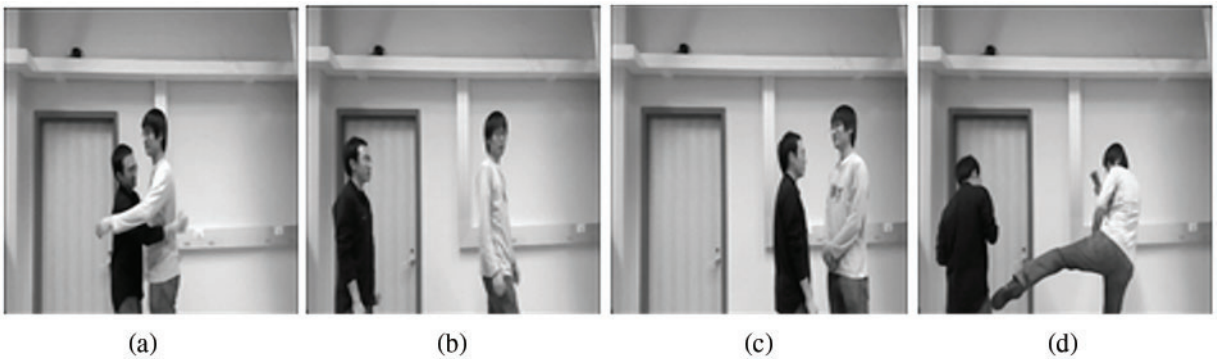


**Figure 3:** Gaussian filtering applied to highlighting the improved contrast and reduced noise in the processed image as (a) hug (b) depart (c) approch and (d) kick

We have chosen the Gaussian filter for its simplicity and ability to remove image noise while keeping the important details. Its implementation is essential in our study for correctly detecting human activities. Our image-based human interaction recognition method performs much better when we use a Gaussian filter on each channel after the color space transformation separately.

### *3.4 Silhouette Extraction*

In addition, silhouette extraction is critical in computer vision tasks such as object recognition, tracking, and segmentation [18]. This makes the silhouette extraction accurate and immune to whatever statistical-based technique is used. Such models use statistical information to learn and accomplish several tasks. The algorithm chosen for implementation is the GMM due to its capability of modeling complex distributions of data, the flexibility of modeling multiple modes, probability density function, the provision by which data points are probabilistically assigned, the procedure of locating model parameters, and generative characteristics that make the method suitable for application in clustering, density estimation and detection of anomalous data.

The approach recommended in this study is to use the Gaussian Mixture Model (GMM) for silhouette extraction. We get the binary mask of the input image by employing two types of thresholding methods and inverse thresholding methods. After that, GMM is used to segment the image to obtain the silhouettes. The silhouette is a binary image where the foreground pixels represent the subject and the background pixels are the parts of the background. The outcome is the second monochromic image placed on top of the first coloured image but on the black background. Fig. 4 shows the results of the precision and the performance obtained by our approach in silhouette extraction.

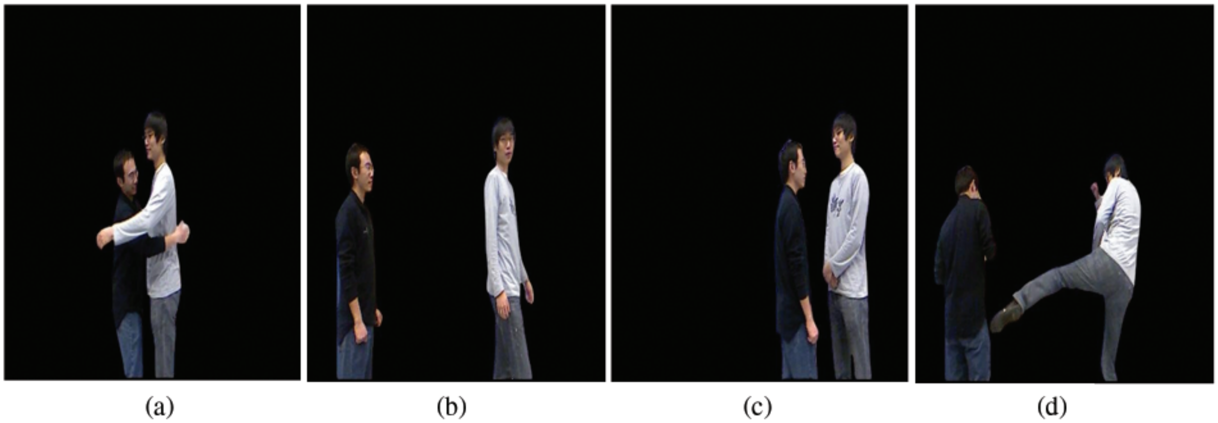

(a) (b) (c) (d)

**Figure 4:** Silhouette extraction using statistical method in the processed image as (a) hug (b) depart (c) approch and (d) kick

The GMM formulation employed in the present study is defined by Eq. (5), where $p(x)$ is probability density function, $w_i$ is the weight of the $i$th Gaussian component, is the Gaussian distribution function, $k$ is the number of Gaussian components, is the mean vector and is the covariance matrix.

$$p(x) = \sum\nolimits_{\{i=1\}\{k\}} w_i \phi (x; \mu_i, \Sigma_i) \tag{5}$$

**Algorithm 1 :** Silhouette extraction using statistical method (GMM)

**Require:** Frames

**Ensure:** A silhouette image on a black background and the original image with a silhouette overlay

```
function SILHOUETTEEXTRACTION
    $O_I$ ← readInputImage()
    $G_I$   convertToGrayscale($O_I$)
    backgroundSubtractor   initializeGMMBackgroundSubtractor()
    while True do
        $FG_I$ ← backgroundSubtractor.apply($G_I$)
        meanValue   mean($FG_I$)   ▷ Compute the mean value of the foreground mask
        if meanValue > T then
            binaryMask ← threshold($FG_I$, T, 255, THRESH   BINARY)
            inverseMask ← bitwiseNot(binaryMask)
            silhouetteImage   bitwiseAnd($O_I$, $O_I$, mask = inverseMask)
            showImageOnBlackBackground(silhouetteImage)
            showOriginalImageWithSilhouette($O_I$, silhouetteImage)
        else
            Continue   ▷ No silhouette detected, continue to the next iteration
        end if
    end while
end function
```

The silhouette extraction process by the statistical method (GMM) is presented in Algorithm 1, where $O_I$ represents the input image, $G_I$ illustrates the grey image and $FG_I$ is the foreground image. This algorithm appropriately isolates the silhouette of an object from an input image, rendering a silhouette image on a black background and the original image with the silhouette overlaid. To begin with, the image is changed to grayscale. After that, GMM background subtractor is used and the foreground mask is to be thresholded. When the mean of the values exceeds the limit, the mask is turned into an inversed mask to binarize. The silhouette obtained is presented on a black background and the original image is superimposed with the silhouette results shown in Fig. 4.

### *3.5 Process of Feature Extraction*

I used a mix of BRISK, and SIFT methods to extract features, which helped in effectively representing and characterizing visuals and features in the data.

#### *3.5.1 Binary Robust Invariant Scalable Keypoints (BRISK)*

BRISK is one of the best feature extraction and matching strategies used in many research studies onomputer vision and image processing [19]. Leutenegger et al. [20] introduced BRI, which provides a strong and efficient method for detecting and describing local image features. It is characterized by using a binary descriptor, significantly reducing memory and computational requirements. The main strength of BRISK is its robustness to all kinds of image transformations such as rotation, scaling, and viewpoint changes. This is done using a scale-space pyramid and a multi-scale feature detection approach. BRISK is found to be optimal for highspeed applications where efficiency and correctness are necessary. The effectiveness of BRISK in applications like image matching, object recognition, and visual tracking has been consistently proven in detailed evaluations performed on benchmark datasets. Versatility, robustness, and computational efficiency among other qualities of BRISK have positioned it as one of the most beneficial and frequently used tools in the dynamic field of computer vision.

The BRISK scale-space pyramid construction equation:

$$Li(x, y, \sigma) = G(x, y, \sigma)\ and\ I(x, y) \tag{6}$$

In Eq. (6), $L_i(x, y, \sigma)$ represents the level $i$ of the scale space pyramid at the spatial coordinates $(x, y)$ and scale $\sigma$. Where $G(x, y, \sigma)$ is the Gaussian kernel at position $(x, y)$ and $I(x, y)$ is the image to be smoothed. The next equation illustrates how the scale space pyramid is produced by convolving the input image with the Gaussian kernels at varying scales. As a result, the obtained pyramid has several levels, with each capturing image features at various scales, which play the role of the scale invariance in the BRISK algorithm and results shown in Fig. 5.

#### *3.5.2 Scale-Invariant Feature Transform (SIFT)*

SIFT is a popular computer vision algorithm for finding and describing local image features. As David Lowe proposed, SIFT offers tolerance to changes in scale, rotation, and affine transformations [21]. The critical step in the SIFT algorithm is developing a scale-space representation using the Difference of Gaussian (DoG) filters. This is performed by convolving the input image with a set of Gaussian filters at different scales and subtracting the blurred images to generate the DoG pyramid, undefined

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \tag{7}$$

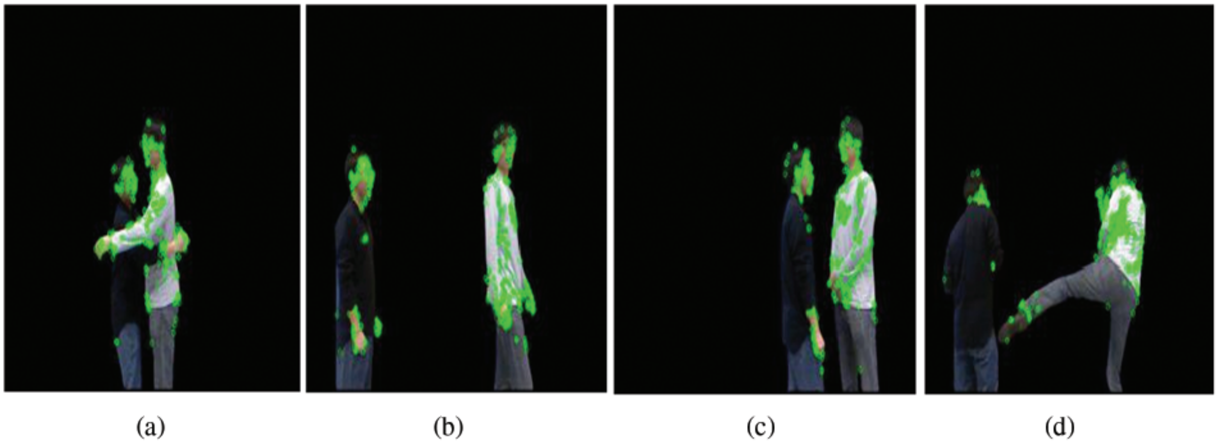


**Figure 5:** BRISK result shown as (a) hug (b) depart (c) approch and (d) kick

In this case, $D(x, y, \sigma)$ is the DoG response at the pixel points $(x, y)$ and scale $\sigma$, and $G(x, y, \sigma)$ is the Gaussian kernel at position $(x, y)$ with standard deviation $\sigma$. $k$ is a scaling factor that determines the magnitude difference between neighbouring levels in the scale space. The DoG pyramid represents important changes in pixel intensities at various scales, upon which the algorithm of extracting stable and distinctive features in SIFT is based. The key points are then represented using orientation histograms, yielding invariant and distinctive features that can be compared across different images or used for various computer vision tasks. SIFT results shown in Fig. 6.

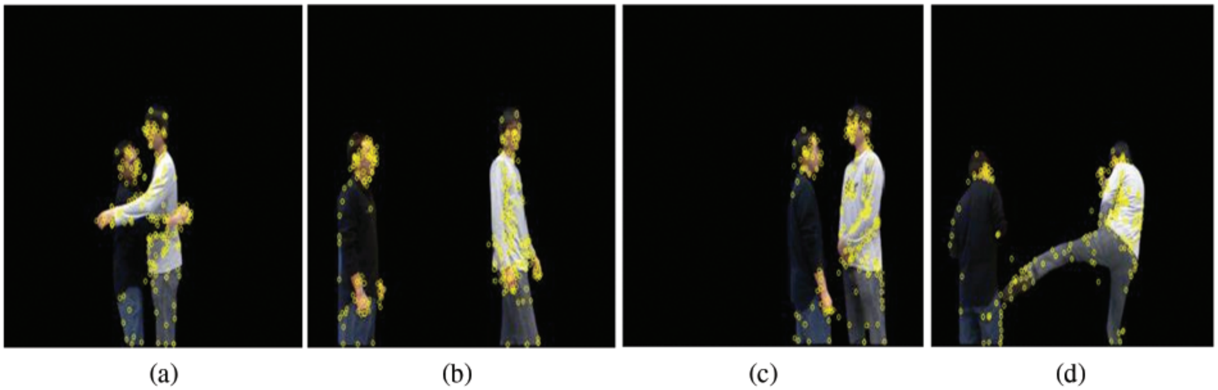


**Figure 6:** SIFT result shown as (a) hug (b) depart (c) approch and (d) kick

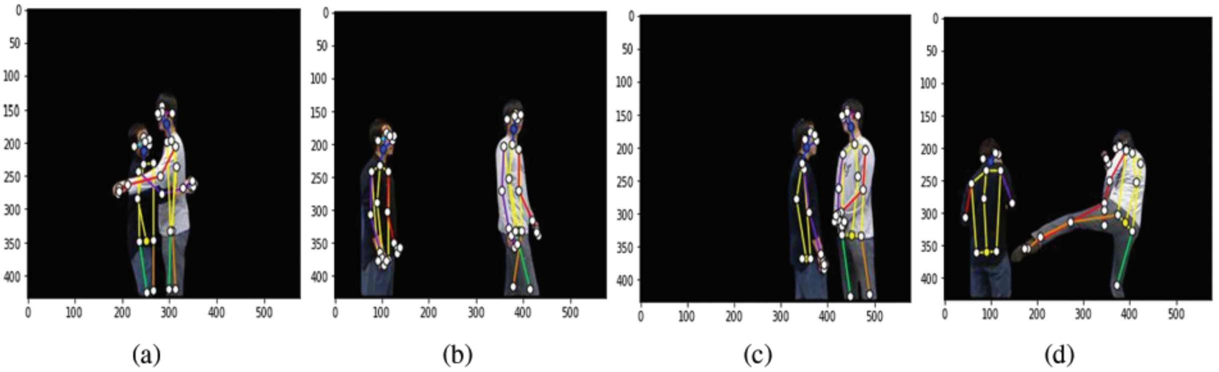
### *3.6 Skeleton Geometry features*

The characteristics of skeleton geometry are important in the analysis of the human skeleton, where information is extracted from the skeletal structure to understand human motion and behavior [22]. These attributes represent the spatial relations and geometrical characteristics of important points of the human skeleton, like the position of junctures and the length of bones. A popular method for extracting skeleton geometry features is derived from the Euclidean distance between pairs of skeleton joints. Distance calculations between particular joint combinations offer important information about the human pose and movements.

An equation commonly employed to calculate the Euclidean distance between two skeleton joints:

$$d_{\{ij\}} = \sqrt{\left\{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2\right\}} \tag{8}$$

In Eq. (8), $d_{\{i,j\}}$ denotes the Euclidean distance between the $i$-th and $j$-th skeleton joints. The coordinates of the $i$-th joint is written as $(x_i, y_i, z_i)$, the j-th joint is written as $(\mathrm{x}_j, \mathrm{y}_j, \mathrm{z}_j)$. The formula determines the three dimensional distance by adding the squares of the differences and then taking the square root of the resultant value differences between the $x$, $y$, and $z$ coordinates. Using Eq. (8), features can be derived from the skeleton geometry, thus producing much information about the skeleton structure and motion of the human being. Their results are shown in Fig. 7.



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

**Figure 7:** Skeleton Geometry features result shown as (a) hug (b) depart (c) approch and (d) kick

### *3.7 Feature Discrimination*

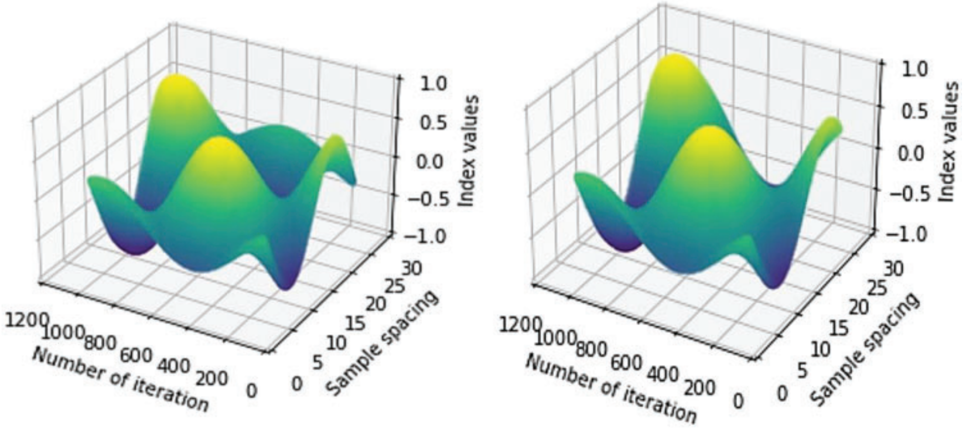
Grey Wolf Optimization (GWO) can be tailored to deal with image features by representing each image as a high-dimensional feature vector. These feature vectors reflect different attributes and details that exist in the images. The optimization process of GWO next selects iteratively the most significant features by changing the positions of the grey wolves in the search space. Every grey wolf represents one of the potential feature subsets, and the position of a wolf gives information on whether a particular feature is included in or out of the subset. The fitness of each grey wolf is assessed during the optimization process by defining an objective function. This performance function measures the discriminating power of the chosen features or how effectively they can solve an image processing task, like image classification, object detection, and image retrieval. It is an algorithm derived from the social interaction of the grey wolves. The fitness function is used to determine the quality of a solution in terms of its performance or objective value. The fitness function is used to control the optimization process and calculate the fitness values of the wolves in order to select the best solutions for further evolution. This aids in arriving at the best or near-best solution to the optimization problem, such as feature optimization.

$$\vec{X}i\,(t+1) = \vec{X}i\,(t) + \vec{V}i\,(t+1) \tag{9}$$

In this Eq. (9), $\vec{X}i\,(t+1)$ represents the new position of the $i$-th grey wolf at time $t+1$. $\vec{X}i\,(t)$ is the current position of the grey wolf and $\vec{V}i\,(t+1)$ represents the velocity term, which controls the movement of the grey wolf towards a new location. The velocity term $\vec{V}i\,(t+1)$ is calculated as follows:

$$\vec{V}i\,(t+1) = \vec{A} \odot \vec{D}_{\text{rand}} - \vec{X}i\,(t) \tag{10}$$

In Eq. (10), $\vec{A}$ is a uniformly chosen coefficient vector, and $\vec{D}$ rand is a randomly selected vector from the population of grey wolves. The equations control the motion and updating of the locations of grey wolves in the GWO algorithm in image feature extraction. Iterating through these equations enables the algorithm to navigate the feature space in search of the most informative and discriminative image features, with the results shown in Fig. 8.



(a)

(b)

**Figure 8:** Discrimination of features over the (a) HMDB51 and (b) SUB-Interaction dataset

Using the grey wolves' social hierarchy and hunting behaviour, the GWO algorithm searches the feature space to reveal the most significant and discriminative image features. Reshaped by the movements of the grey wolves which in turn are incepted by alpha, beta, delta, and omega wolves, the search is directed to the regions of the feature space exhibiting probabilistic discriminative power. You have utilized the GWO algorithm to extract features from the image; therefore, you have enlivened collective wisdom and the hierarchy feature of grey wolves, improving the performance of image processing tasks. The feature subset resulting from GWO optimization can enhance the efficiency of many image analysis applications.

## 4 Experimental Analysis

The Extended Kalman Filter (EKF) is a famous estimation algorithm that couples the principles of the Kalman Filter with non-linear system models [23]. The approach is especially useful for systems that demonstrate non-linear dynamics and measurement equations. With the current estimated state being used, the EKF linearizes the system's non-linear functions, making the state estimation and tracking somewhat effective. The EKF is a powerful and effective method of estimating the true state of a system, even when noisy measurements are present, as it iteratively updates the estimated state according to the incoming measurements. Due to its flexibility and efficiency, it is widely used in numerous application areas including robotics, control systems, navigation, and others.

### *4.1 SUB-Interaction Dataset*

The SUB-Interaction dataset is a rich source for research on human interactions in video. This dataset is a group of numerous small video fragments that are well described and reveal various aspects of human communication. The SUB-Interaction dataset is a set of videos as RGB-D data representing human interactions. 282 videos are filmed in different internal and external locations. As shown in Fig. 8, these videos are categorized into 8 types of interactions: The activities which are made up of pull, crawl, push, approach, handshake, hug, kick, pass object, punch and depart, are provided with the people positions and orientations within the videos. This dataset has been widely used in computer vision and machine learning to build models that automatically detecting and classifying human activities from video clips.

### *4.2 HMDB51 Dataset*

The set of video clips based HMDB51 dataset is common and extensively used libraries for human action recognition in research. It consists of up to 51 action classes with people performing walking, running, soccer play, and cycling activities. Action being done: Tapping every video clip of the database helps surveyors develop and test algorithms for action recognition tasks. HMDB51 dataset is one of the most famous activity detection datasets in the computer vision community as it is widely used to evaluate various techniques. It represents a potent tool for creating algorithms that would automatically follow and classify human activities in video data, helping the areas of video surveillance, sports analysis, and gesture recognition, among others. The availability of the HMDB51 dataset has led to the development of the field of human activity study and control from the visual data subfield.

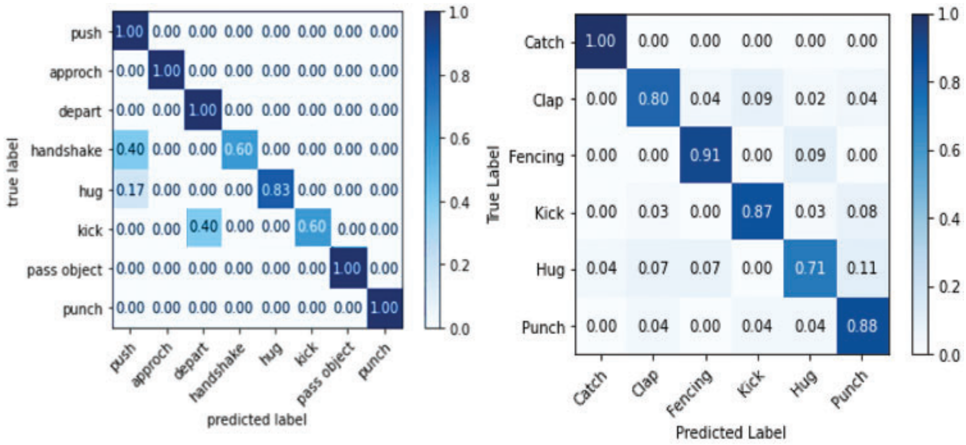
### *4.3 Performance Evaluation*

The suggested method was tested on the SUB-Interaction dataset, where it achieved 88% and HMDB51 86% recognition accuracy using EKF. The conclusion is that the variations or instances that do not appear in the training data may pose difficulties for the model when exposed to unseen problems. One possibility of its failure might be the shortcomings of the training set, that is, its small size and lack of balance between different classes; thus, increasing the size of the training dataset and balancing the classes might solve this problem and minimize misclassifications. Figure confusion matrix verifies the EKF-validated classification performance.

We analyzed the performance of our system on the SUB-Interaction dataset and attained an overall accuracy of 88%. The model shows strong performance and attained excellent recognition rates for all eight interaction classes, with F1-scores varying from 0.40 to 1.00. Our method successfully attained a high recognition rate for various interactions. The suggested approach for identifying human action in practical environments is shown in Table 1a and Fig. 9. We evaluated the performance of our system to the interaction dataset of HBDB51, obtaining an excellent overall accuracy of 86%. Our model showed a strong performance, being able to recognize all eight interaction classes well, with its F1-scores of 0.77 to 0.97. Our method effectively gained a high level of accuracy in realizing many interactions. The results provided show the efficiency and reliability of our proposed approach for human action recognition in real-world environments, presented in Table 1b and Fig. 9.

**Table 1:** Performance measures of (a) SUB-Interaction dataset and (b) HMDB51 dataset for HAR

| Classes | Recall | Precision | F1-score |
|---|---|---|---|
| **(a)** | | | |
| Punch | 1.00 | 0.50 | 0.67 |
| Approch | 1.00 | 1.00 | 1.00 |
| Depart | 1.00 | 1.00 | 1.00 |
| Handshake | 0.60 | 1.00 | 0.75 |
| Hug | 0.83 | 0.83 | 0.83 |
| Kick | 0.60 | 1.00 | 0.75 |
| Pass object | 1.00 | 0.71 | 0.83 |
| Punch | 1.00 | 1.00 | 1.00 |
| **(b)** | | | |
| Catch | 1.00 | 0.94 | 0.97 |
| Clap | 0.80 | 0.90 | 0.85 |
| Fencing | 0.91 | 0.71 | 0.80 |
| Kick | 0.87 | 0.87 | 0.87 |
| Hug | 0.71 | 0.83 | 0.77 |
| Punch | 0.88 | 0.74 | 0.81 |



(a) (b)

**Figure 9:** Confusion, matrix of a proposed HMM-based approach for recognising human interactions (a) shows SUB-Interaction outcomes and (b) illustrates HMDB51-Interaction outcomes

Fig. 9 confusion matrix reveals that our technique recognized most interaction kinds with few misclassifications. Our system for identifying human action in real-world surroundings works well and is resilient.

The comparison table of both datasets for detecting human interactions is shown in Table 2.

**Table 2:** An assessment of the accuracy of various action recognition approaches

| | HMDB51 dataset | SUB-Interaction | |
|---|---|---|---|
| Methods | Accuracy | Methods | Accuracy |
| Autoencoders (VideoMAE) [24] | 0.62% | Raw skeleton [25] | 0.79% |
| VGGNet + ConvNets [26] | 0.77% | Joint feature [27] | 0.80% |
| Temporal attention vectors (TAVs) [28] | 0.77% | Hierarchical RNN [29] | 0.80% |
| | | XGBoost [30] | 0.88% |
| Our | 0.86% | Our | 0.88% |

## 5 Conclusion

The novel system for HAR proposed in this paper achieves an 88% and 86% accuracy on the SUB-Interation and HMDB51 datasets, respectively. The proposed method consists of a few main steps: improvement of the frame and extract, the outline, the feature, the fusion and discrimination, and the classification by EKF. In other words, our method is computationally fast and low latency, which is ideal for edge device real-time applications. The research findings of the thesis have significance in computer vision and pattern recognition, with potential usage in biometrics, surveillance, and human-computer interaction. For the further enhancement of the system, following are the future possible enhancements can be associated with the deep learning method like CNN and RNN for the better feature extraction and classification. Incorporating the CNNs and the RNNs with the current method would enable the enhancement of feature optimization where CNN is used in the extraction of spatial features while RNN is used in the modeling of temporal features resulting in better accuracy and context. In addition, it is necessary to generalize by testing the methodology on larger data sets or in more complex situations so that its generality and effectiveness in practice can be confirmed.

**Acknowledgement:** Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R440), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Funding Statement:** The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Group Funding Program grant code (NU/RG/SERC/13/40). The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Kingdom of Saudi Arabia for funding this research work through the project number "NBUFFR-2024-231-08".

**Author Contributions:** Study conception and design: Tanvir Fatima Naik Bukht and Abdulwahab Alazeb; data collection: Naif Al Mudawi, Hui Liu, Ahmad Jalal, Bayan Alabdullah and Khaled Alnowaiser; analysis and interpretation of results: Tanvir Fatima Naik Bukht and Hui Liu; draft manuscript preparation: Tanvir Fatima Naik Bukht and Ahmad Jalal. 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] V. S. Nguyen, H. Kim, and D. Suh, "Attention mechanism-based bidirectional long short-term memory for cycling activity recognition using smartphones," *IEEE Access*, vol. 11, pp. 136206–136218, 2023. doi: 10.1109/ACCESS.2023.3338137.

[2] R. Saini, P. Kumar, P. P. Roy, and D. P. Dogra, "A novel framework of continuous human-activity recognition using Kinect," *Neurocomputing*, vol. 311, no. 2, pp. 99–111, 2018. doi: 10.1016/j.neucom.2018.05.042.

[3] T. F. N. Bukht, H. Rahman, and A. Jalal, "A novel framework for human action recognition based on features fusion and decision tree," in *2023 4th Int. Conf. Adv. Comput. Sci. (ICACS)*, Lahore, Pakistan, 2023, pp. 1–6. doi: 10.1109/ICACS55311.2023.10089752.

[4] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson and C. Fookes, "Continuous human action recognition for human-machine interaction: A review," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–38, 2023. doi: 10.1145/3587931.

[5] V. Nunavath *et al.*, "Deep learning for classifying physical activities from accelerometer data," *Sensors*, vol. 21, no. 16, 2021, Art. no. 5564. doi: 10.3390/s21165564.

[6] T. F. N. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujally and A. Jalal, "A review of video-based human activity recognition: Theory, methods and applications," *Multimed. Tools Appl.*, vol. 7, no. 2, pp. 1–47, 2024. doi: 10.1007/s11042-024-19711-w.

[7] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, 2014. doi: 10.1145/2499621.

[8] T. Zebin, P. J. Scully, and K. B. Ozanyan, "Human activity recognition with inertial sensors using a deep learning approach," in *2016 IEEE Sens.*, 2016, pp. 1–3.

[9] R. Kanjilal and I. Uysal, "Rich learning representations for human activity recognition: How to empower deep feature learning for biological time series," *J. Biomed Inform.*, vol. 134, no. 4, 2022, Art. no. 104180. doi: 10.1016/j.jbi.2022.104180.

[10] D. Hussein and G. Bhat, "CIM: A novel clustering-based energy-efficient data imputation method for human activity recognition," *ACM Trans. Embed. Comput. Syst.*, vol. 22, no. 5s, pp. 1–26, 2023. doi: 10.1145/3609111.

[11] R. Kanjilal, M. F. Kucuk, and I. Uysal, "Subtransfer learning in human activity recognition: Boosting the outlier user accuracy," *IEEE Sens. J.*, vol. 23, no. 20, pp. 25005–25015, 2023. doi: 10.1109/JSEN.2023.3312146.

[12] S. An, G. Bhat, S. Gumussoy, and U. Ogras, "Transfer learning for human activity recognition using representational analysis of neural networks," *ACM Trans. Comput. Healthc.*, vol. 4, no. 1, pp. 1–21, 2023. doi: 10.1145/3563948.

[13] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "A survey on artificial intelligence in posture recognition," *Comput. Model Eng. Sci.*, vol. 137, no. 1, pp. 35–82, 2023. doi: 10.32604/cmes.2023.027676.

[14] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning," *Sensors*, vol. 19, no. 7, 2019, Art. no. 1716. doi: 10.3390/s19071716.

[15] J. Zhao, S. Chong, L. Huang, X. Li, C. He, and J. Jia, "Action recognition based on CSI signal using improved deep residual network model," *Comput. Model. Eng. Sci.*, vol. 130, no. 3, pp. 1827–1851, 2022. doi: 10.32604/cmes.2022.017654.

[16] C. Xin, S. Kim, Y. Cho, and K. S. Park, “Enhancing human action recognition with 3D skeleton data: A comprehensive study of deep learning and data augmentation,” *Electronics*, vol. 13, no. 4, 2024, Art. no. 747. doi: 10.3390/electronics13040747.

[17] T. F. N. Bukht and A. Jalal, “Human action recognition based on embedded HMM,” in *2024 5th Int. Conf. Adv. Comput. Sci. (ICACS)*, Lahore, Pakistan, 2024, pp. 1–7.

[18] M. R. M. Hasan and N. H. S. Alani, “A comparative analysis using silhouette extraction methods for dynamic objects in monocular vision,” *Cloud Comput. Data Sci.*, pp. 1–12, 2022. doi: 10.37256/ccds.3220221201.

[19] D. S. Sathiya, “Texture classification with modified rotation invariant local binary pattern and gradient boosting,” *Int. J. Knowl. Based Intell. Eng. Syst.*, vol. 26, no. 2, pp. 125–136, 2022. doi: 10.3233/KES220012.

[20] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary robust invariant scalable keypoints,” in *2011 Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 2548–2555.

[21] W. Burger and M. J. Burge, “Scale-invariant feature transform (SIFT),” in *Digital Image Processing: An Algorithmic Introduction*. Vancouver, BC, Canada: Springer, 2022, pp. 709–763.

[22] D. C. Luvizon, H. Tabia, and D. Picard, “Learning features combination for human action recognition from skeleton sequences,” *Pattern Recognit. Lett.*, vol. 99, no. 9, pp. 13–20, 2017. doi: 10.1016/j.patrec.2017.02.001.

[23] S. Lo Feudo, J. -L. Dion, F. Renaud, G. Kerschen, and J. -P. Noël, “Video analysis of nonlinear systems with extended Kalman filtering for modal identification,” *Nonlinear Dyn.*, vol. 111, no. 14, pp. 13263–13277, 2023. doi: 10.1007/s11071-023-08560-1.

[24] Z. Tong, Y. Song, J. Wang, and L. Wang, “Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training,” *Adv. Neural Inf. Process Syst.*, vol. 35, pp. 10078–10093, 2022.

[25] Y. Ji, G. Ye, and H. Cheng, “Interactive body part contrast mining for human interaction recognition,” in *2014 IEEE Int. Conf. Multimed. Expo Workshops (ICMEW)*, Chengdu, China, 2014, pp. 1–6.

[26] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu and Q. Tian, “Pooling the convolutional layers in deep ConvNets for video action recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1839–1849, 2018. doi: 10.1109/TCSVT.2017.2682196.

[27] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, “Two-person interaction detection using body-pose features and multiple instance learning,” in *2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops*, Providence, RI, USA, 2012, pp. 28–35. doi: 10.1109/CVPRW.2012.6239234.

[28] Y. Bo, Y. Lu, and W. He, “Few-shot learning of video action recognition only based on video contents,” in *2020 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass, CO, USA, 2020, pp. 584–593. doi: 10.1109/WACV45572.2020.9093481.

[29] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Boston, MA, USA, 2015, pp. 1110–1118.

[30] T. F. N. Bukht and A. Jalal, “A robust model of human activity recognition using independent component analysis and XGBoost,” in *2024 5th Int. Conf. Adv. Comput. Sci. (ICACS)*, Lahore, Pakistan, 2024, pp. 1–7.