Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 16-33
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Optimal Features Selection for Human Activity Recognition (HAR) System Using Deep Learning Architectures

Subrata Kumer Paul[1,2], Rakhi Rani Paul[1,2], Md. Atikur Rahman[2], Md. Momenul Haque[2], Md. Ekramul Hamid[1]

[1]Department of Computer Science & Engineering (CSE), University of Rajshahi, Rajshahi, Bangladesh
[2]Department of Computer Science and Engineering (CSE), Bangladesh Army University of Engineering & Technology (BAUET), Dayarampur, Bangladesh
Email: sksubrata96@gmail.com, rakhipaul.cse@gmail.com, atik.cse.pust35@gmail.com, mominulhaquemim13@gmail.com, ekram_hamid@yahoo.com

How to cite this paper: Paul, S.K., Paul, R.R., Rahman, Md.A., Haque, Md.M. and Hamid, Md.E. (2024) Optimal Features Selection for Human Activity Recognition (HAR) System Using Deep Learning Architectures. *Journal of Computer and Communications*, **12**, 16-33.
https://doi.org/10.4236/jcc.2024.1212002

Received: June 15, 2024
Accepted: November 26, 2024
Published: December 10, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

One exciting area within computer vision is classifying human activities, which has diverse applications like medical informatics, human-computer interaction, surveillance, and task monitoring systems. In the healthcare field, understanding and classifying patients' activities is crucial for providing doctors with essential information for medication reactions and diagnosis. While some research methods already exist, utilizing machine learning and soft computational algorithms to recognize human activity from videos and images, there's ongoing exploration of more advanced computer vision techniques. This paper introduces a straightforward and effective automated approach that involves five key steps: preprocessing, feature extraction technique, feature selection, feature fusion, and finally classification. To evaluate the proposed approach, two commonly used benchmark datasets KTH and Weizmann are employed for training, validation, and testing of ML classifiers. The study's findings show that the first and second datasets had remarkable accuracy rates of 99.94% and 99.80%, respectively. When compared to existing methods, our approach stands out in terms of sensitivity, accuracy, precision, and specificity evaluation metrics. In essence, this paper demonstrates a practical method for automatically classifying human activities using an optimal feature fusion and deep learning approach, promising a great result that could benefit various fields, particularly in healthcare.

Keywords

Surveillance, Optimal Feature, SVM, Complex Tree, Human Activity Recognition, Feature Fusion

1. Introduction

In recent years, the e-vision community has directed its attention toward human activity recognition, which is crucial for numerous applications, including human-computer interaction [1], anti-terrorism [2], traffic surveillance [3], vehicle safety [4], pedestrian detection [5], video surveillance [6], real-time tracking [7], rescue operations [8], and human-robot interaction [9] are some examples of the various applications of digital surveillance. The efficient recognition of human activity in recorded videos is the focus of this study. Due to changes in appearance, color, and movement, developing a cost-effective algorithm to recognize a person from a video or image is difficult. Complicating factors include variations in the background and light. Numerous approaches, including feature extraction, segmentation strategies, and classifiers, have been developed to detect humans [10]. Unfortunately, several persons appearing in a scene or image is a challenge for present techniques, and they may not always produce the optimal results. Moreover, many methods are used to identify humans, including the Histogram of Gradients (HOG) [11], Haar-like features [12], adaptive contour features (ACF) [13], Hybrid Wind Farm (HWF) [14], Image Source Method (ISM) [15], edge detection [16], and movement characteristics [17]. When people are unclear or show notable positional variations, these extraction techniques might not be able to reliably capture the details. But, properly selecting pertinent features can greatly improve the ability to identify human activity. In this research, we propose a deep learning technique to overcome accuracy challenge of human activity recognition with selecting optimal features. This is accomplished by enhancing the quality of the frames that are extracted from videos and then classifying the areas based on specified feature vectors. There are five main phases in the proposed method including (a) data normalization, (b) image feature extraction, (c) best feature selection, (d) extract feature fusion, and (e) finally classify the targeted class. Several preprocessing steps, including background subtraction, noise reduction, and object extraction, are used during the normalization stage. We extract three types of features: HOG (Histogram of Oriented Gradients), Gabor, and color chromatic features. We use Principal Component Analysis (PCA) method on each set of features to get the best ones. Then, we combine these selected features. Finally, we use five different classifiers to find the most accurate results.

Major Contributions

Ineffective and lengthy preprocessing procedures decline the optimality as well as the accuracy of any algorithm. This work focuses on the efficient and accurate use of preprocessing and feature extraction steps. Thus, main contributions in this work include the following:

1) After removing the background, morphological operations are used to identify and define the specific area of interest precisely.

2) Independent scoring based on principal components for selecting feature subsets.

3) The optimal results are achieved by using a variety of classification methods.

2. Related Works

The following section provides a detailed examination of significant studies in the field of human activity recognition. In computer vision, one of the primary fields of study is action recognition, which is a subset of gesture recognition [18]. Researchers have employed a variety of methods to develop action-recognition systems, such as artificial intelligence (AI), hand-crafted features combined with traditional machine learning algorithms, and diverse deep learning approaches [19]. Traditional machine learning algorithms and a range of manual feature extraction methods were mostly used to construct the present human activity recognition (HAR) [20]. Conventional machine learning techniques for action recognition typically follow a three-step procedure. At the first phase, features are extracted using manually created descriptors. Then, a particular algorithm is used to encode these features. In the final stage, the encoded features are classified using a suitable machine-learning method [21]. Two distinct techniques are employed in different tasks: local feature-based approaches and global feature-based approaches. The main goal of local feature-based techniques is to characterize features as separate patches, interest points, and gesture information. The learned cues relevant to the current task are aligned with these features. Global features, on the other hand, encompass the entire region of interest.

Table 1. List of literature review and their performances.

SN	Authors List	Method(s)	Dataset	Accuracy (%)
1	Simonyan *et al.* [21]	Two-Stream CNN	JHMDB	Not specified
2	Wensel *et al.* [22]	ViT-ReT	JHMDB	85.20%
3	Feichtenhofer *et al.* [23]	Spatial-Motion	UCF101	Not specified
4	Tu *et al.* [24]	Multi-Stream CNN	JHMDB	71.17%
5	Gammulle *et al.* [25]	LSTM-based fused	UCF Sports	92.2%
6	Ijjina *et al.* [26]	Hybrid Technique	UCF11	69.0%
7	Meng *et al.* [27]	LSTM	Not specified	93.2%
8	Xu *et al.* [28]	Deep Learning	Not specified	Not specified
9	Najmul *et al.* [29]	BiLSTM	UCF11	85.30%
10	Riahi *et al.* [30]	Dilated CNN + BiLSTM + RB	UCF11	79.40%
11	Rama *et al.* [31]	Deep Learning Architecture	JHMBD	67.24%
12	Gammulle *et al.* [32]	Two-Stream Long Short-Term Memory (LSTM)	JHMBD	55.70%
13	Yang *et al.* [33]	Various DL Algorithms	JHMBD	65.00%

As noted in the references [22], background removal and tracking techniques are frequently used to accomplish this. Yasin *et al.* [23], who have presented a fundamental method for identifying actions in video sequences by utilizing keyframe selection and applying it to human activity recognition (HAR). Zhao *et al.* [24] presented the HAR system that leverages keyframes and employs conventional machine learning techniques for multi-feature fusion. Yala *et al.* [25] introduce a novel activity recognition system that relies on streaming data. The strategy shows a remarkable level of accuracy in recognizing important human activities. Nunes *et al.* [26] introduced a framework aimed at daily life human activity recognition. Many features are first extracted by the suggested method. Following this, two successive automatically recognized key positions are encircled around each human activity frame, from which the maximal static and dynamic features are retrieved. Kantorov and Laptev [27] identified the application of Fisher vectors for feature encoding and effectively used linear classifiers to obtain accurate action identification. In [28], Lan *et al.* presented a process for improving action recognition systems' operational strategies by switching from data-driven to data-independent approaches. While standard machine learning algorithms have made significant progress in the last 10 years, they still have limits because of human cognition. These limitations include labor intensiveness, time consumption, and the difficulty of feature engineering [29]. Deep learning is a move toward more automated and adaptive techniques that can get over the limitations of handcrafted features. Deep learning departs from the conventional three-step machine learning architecture by introducing a contemporary end-to-end framework. With this framework, classification tasks can be completed simultaneously with the learning and representation of highly discriminative visual features. This following presents the previous related publication for human activity recognition (HAR) using deep leaning techniques. **Table 1** presents the list of papers and its corresponding dataset with accuracy.

From this literature reviews, we can understand the deep learning model that uses feature extraction and classification techniques to improve the accuracy of human action recognition.

3. Proposed Method

The proposed method introduces a new technique for human activity recognition. This innovative approach involves five fundamental steps, which include: (a) identifying moving objects within the video sequence; (b) extracting the HOG, Gabor, and color features of the moving object; (c) selecting the most effective characteristics; (d) merging the selected features sequentially; and (e) classifying the moving object. **Figure 1** illustrates the entire process of the proposed technique.

3.1. Data Preprocessing Steps

During the preprocessing step, a region-wise sliding window approach is imple-

mented to account for variation in each consecutive frame. This approach helps in ignoring unnecessary regions, such as the background. The result of this process is a binary image is extracted through background subtraction, which is then subjected to a noise removal technique. To enhance the image, the binary image is initially converted to RGB color. Then, the RGB image is transformed into the Hue Saturation Intensity (HSI) format. The next stage is drawing a box around a person to identify them. This step aims to improve the quality of the video-extracted frames by enhancing the foreground features. The following are detailed steps involved in the preprocessing stage are presented in Figure 2.

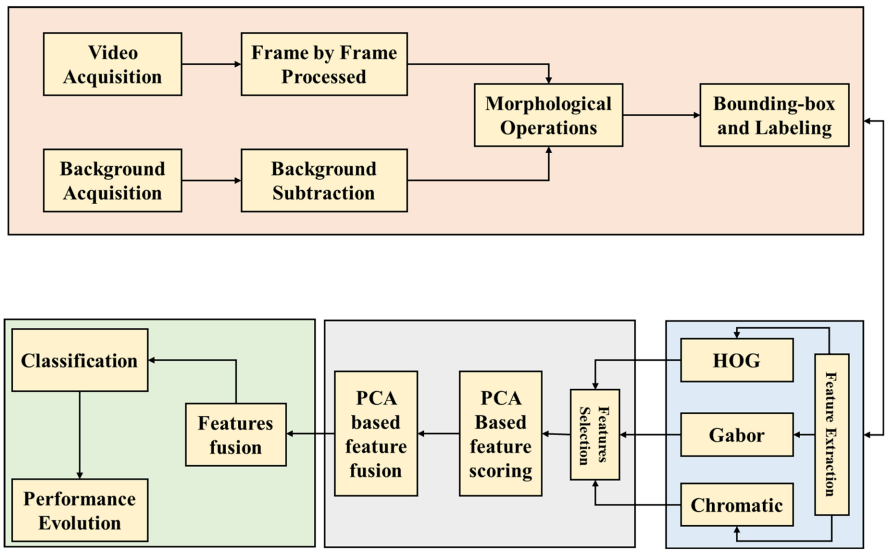


Figure 1. Proposed model block diagram.

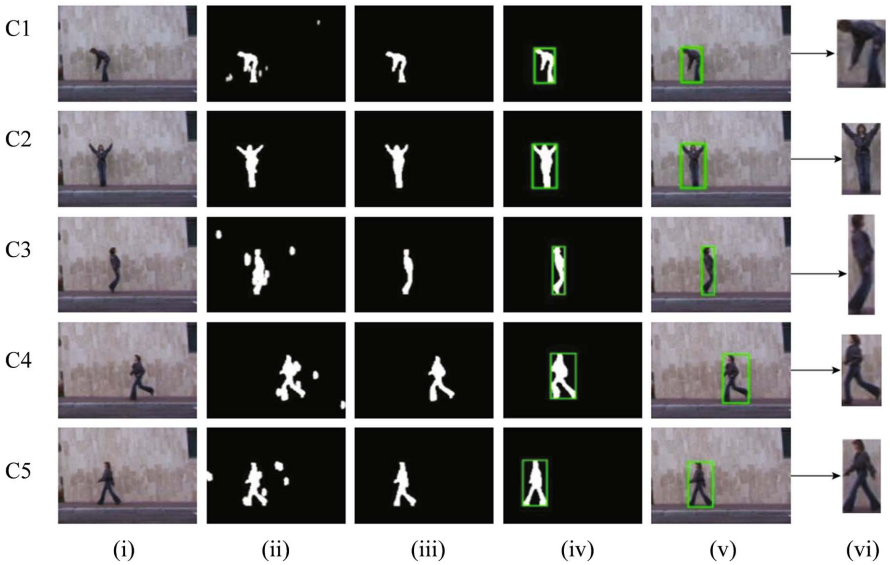


Figure 2. Initially, preprocessing stages include: (i) original images data; (ii) background subtraction images; (iii) image enhancement; (iv) object detection; (v) conversion: binary to RGB; (vi) image cropping.

3.2. Feature Extraction

In this stage, three types of feature extractors named Histogram of Oriented Gradients (HOG), Gabor, and chromatic features that are used to analyze each frame. These three types of feature extractors are considered in our study because the goal is to accurately identify and classify various activities, optimal feature selection becomes pivotal for creating models that are both accurate and practical for real-world deployment [34]. The resulting feature vectors have dimensions of 1 × 3780 for HOG, 1 × 60 for Gabor, and 1 × 9 for co-occurrence matrices and chromatic features. These features capture aspects like gradient distributions, textures, spatial relationships, and color characteristics in the frames, contributing to a comprehensive set for further analysis or classification.

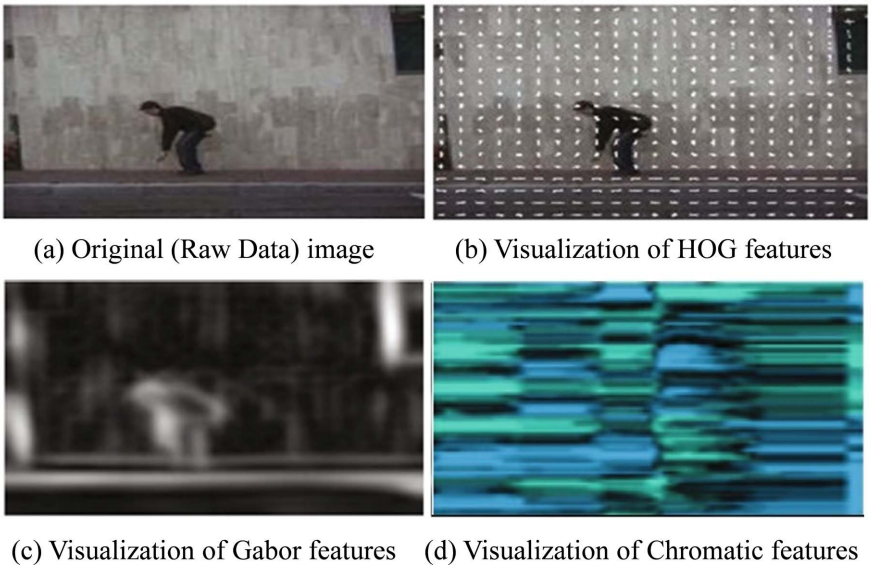
3.2.1. HOG Features

Feature extraction using HOG Method, the image is initially divided into smaller segments, which are then processed individually. Afterwards, these segments are combined back together. To compute directional gradients $\boldsymbol{G}_x$ and $\boldsymbol{G}_y$ the Sobel kernel function is utilized on the processed images according to the following mathematical equations [35].

$$F_{seg_{|G(i,j)|}} = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \tag{1}$$

$$F_{seg_{\varnothing G}}(i,j) = tan^{-1}\left(\frac{G_x(i,j)}{G_y(i,j)}\right) \tag{2}$$

where, $|G|$ denotes magnitude, $\varnothing_G$ supplies the gradient angle, i and j stand for rows and columns concurrently. Based on the gradient, the angle divides the cell votes into bins. Later, each block of the histogram is used to create the standardized vector.



(a) Original (Raw Data) image (b) Visualization of HOG features

(c) Visualization of Gabor features (d) Visualization of Chromatic features

Figure 3. The HOG features and its graphical representation.

The following equation represents the eight bin cells that are used to implement the HOG feature descriptor on the segmented image.

$$\boldsymbol{F}_{\boldsymbol{seg}^{\boldsymbol{V}_i^N}} = \frac{\boldsymbol{V}_i}{\sqrt{\boldsymbol{V}_2^2 + \in^2}} \tag{3}$$

where, the vector V is the non-normalized vector that contains all of the histograms in a block, and $\in$ is a minor constant that does not split by zero. A single block containing all of these vectors is the HOG feature vector. Each feature's range and mean variance are also measured. Figure 3 shows the HOG features and tis graphical representation.

3.2.2. Gabor Features

The following equation illustrates how a complex sinusoidal wave is used to use the "Gaussian Kernel" feature of a modified 2D Gabor filter in the geographic area.

$$F_{seg} = \frac{fs^2}{\pi Y \eta} \exp\left(-\frac{p' + Y^2 q'}{2\sigma^2}\right) \exp\left(2\pi\, fsx' + \varnothing\right) \tag{4}$$

here, Y represents the spatial characteristics where the elliptical support of the Gabor function is defined; p' and q' are detailed in the following equations. fs indicates sinusoidal frequency, $\varnothing$ represents band similarity direction of an activity described by Gabor, $\varnothing$ indicates the phase offset, and σ indicates the Standard Deviation (SD) of the Gaussian wrapper [36]. The gabor feature has five scales and six directions of implementation. The chosen measurement for the gabor feature is 1 × 30. By using the Gabor feature, one may measure the variance and mean. Figure 3(c) describes the graphical representation of HOG features.

3.2.3. Chromatic Features

Because of this method's extensive use, it has become a standard. In contrast, other studies depended on a limited set of functions, including entropy (H), correlation (COR), energy (E), and local homogeneity (LH). To calculate this Chromatic Features, we used these equations (5)-(9) [37].

$$E = \sum_i \sum_j \left[h\left(i, j \middle| d, \theta\right)\right]^2 \tag{5}$$

$$H = -\sum_i \sum_j \left[h(d,\theta) \log h(d,\theta)\right] \tag{6}$$

$$I = \sum_i \sum_j \left[(i-j)^2 h(d,\theta)\right] \tag{7}$$

$$LH = \sum_i \sum_j \frac{\left(i, j \middle| d, \theta\right)}{1 + (i+j)^2} \tag{8}$$

$$COR = \sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y) h\left(i, j \middle| d, \theta\right)}{\sigma_x \sigma_y} \tag{9}$$

where, σ_x and σ_y are the vertical statistics, μ_x is the horizontal mean, and σ_x is the variance. Quantifying color information in images in demonstrate in Figure 3(d).

3.3. Feature Selection

In the presented technique, Principal Component Analysis (PCA) is employed for the purpose of feature selection. This allows the identification and selection of the most significant features from the outcomes of various methods, namely Histogram of Oriented Gradients (HOG), Gabor filters, and chromatic feature vectors. In general, the PCA method transforms a set of n vectors from a d-dimensional space to another space with d' dimensions [39]. This is represented by the equation, which gives the resulting in vectors as ($x_1^{'}$, $x_2^{'}, \cdots$, $x_i^{'}, \cdots$, $x_n^{'}$).

$$x_r^{'} = \sum_{n=1}^{d'} a_{n,r}\, e_n,\ d' \le d \tag{10}$$

where, e_n displays the greatest eigenvalues of the distribution and the eigenvectors corresponding to the d'-dimensional space. Conversely, an, r represents predictions of the vectors x_r across the eigenvectors.

3.4. Feature Fusion

The action recognition method will be efficient and effective because of feature fusion. Additionally, in complex settings, this improves the human action classification rate. Compared to the original Gabor and HOG features, feature fusion in this method yields significantly better results in both the high brightness environment and the dark background. The feature vectors have sizes of 1 × 60, 1 × 3780, and 1 × 9 for HOG, Gabor, and chromatic features, respectively. Let, C1, C2, C3, ⋯, Cn be the human activity classes that require classification in order to perform feature fusion. Consider, $\Delta = \{\varnothing \text{v} \varnothing \in RN\}$ presents the total number of model training samples $\{'Y_{\mathbf{HOG}}, 'Y_{\mathbf{Gab}}, 'Y_{\mathbf{Chrom}}\} \in \boldsymbol{R}^{N}_{\mathbf{HOG+Gab+Chrom}}$ are the three feature vectors that have been extracted. The size is defined as:

$$FV_1 = \{j_1, \cdots, j_k\}, FV_2 = \{y_1, \cdots, y_k\}, FV_3 = \{d_1, \cdots, d_k\} \tag{11}$$

The sizes of the feature vectors are denoted by FV_1, FV_2, and FV_3, representing HOG, Gabor, and cooccurrence matrices with chromatic features, respectively. These feature vector sizes can be further described using the set k, where $k \in \{60, 3780, 9\}$. The sizes of extracted feature sets are: ('Y_HOG→1 × 3780,'Y_Gab→1 × 60,'Y_Chrom→1 × 9). The final extracted vector is indicated as:

$$\boldsymbol{F}(\boldsymbol{\emptyset}) = \sum_{j}^{FV_1} 'Y_{\boldsymbol{HOG}} + \sum_{t}^{FV_2} 'Y_{\boldsymbol{Gab}} + \sum_{d}^{FV_3} 'Y_{\boldsymbol{Chrom}} \tag{12}$$

$$F(\emptyset) = \{(1\times 3780) + (1\times 60) + (1\times 9)\} \tag{13}$$

$$Final(\emptyset) = \{(1\times 3849)\} \tag{14}$$

3.5. Classification

We compared five different ways of making predictions: linear SVM (LS), cubic-SVM (CS), Complex tree (CT), fine-KNN (FK), and subspace KNN (SK). **Figure 4** shows how we selected and combined features to get the best results.

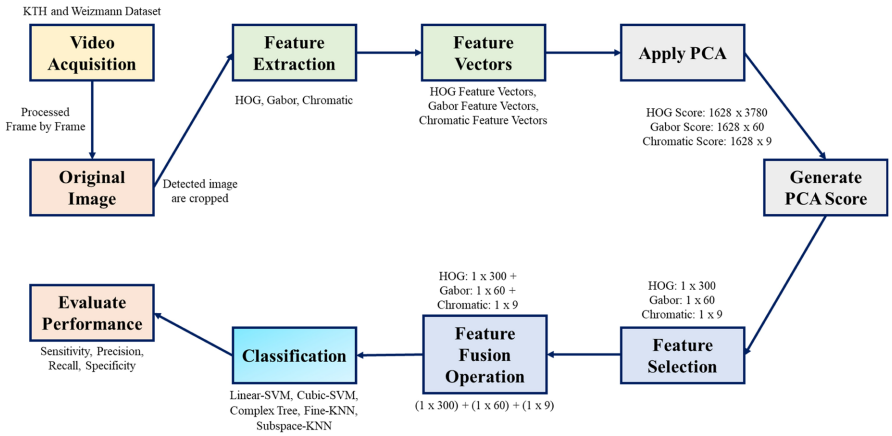


Figure 4. Summary of selecting feature vectors, combining them, and classifying it.

Subspace-KNN did the best on the KTH dataset, and cubic-SVM did better than the other classifiers model on the Weizmann dataset.

4. Results and Analysis of Experiment

In this section, we talk about the datasets we used for our experiments and the results we got based on different performance measures.

4.1. Datasets

The total success of the machine learning model, which includes the suggested network, is heavily influenced by the caliber and applicability of the dataset. In this study, we consider Weizmann Dataset and KTH dataset.

4.1.1. Weizmann Public Dataset

Weizmann dataset comprises 2513 images depicting various human activities, performed by nine different actors and covering five types of human behavior. After selecting and combining features, classification techniques are functional to assess the results. **Figure 5** provides a sample of images from the Weizmann dataset. This dataset consists of five classes: Hand Waving, Running, Jumping, Walking, and Bending [39].

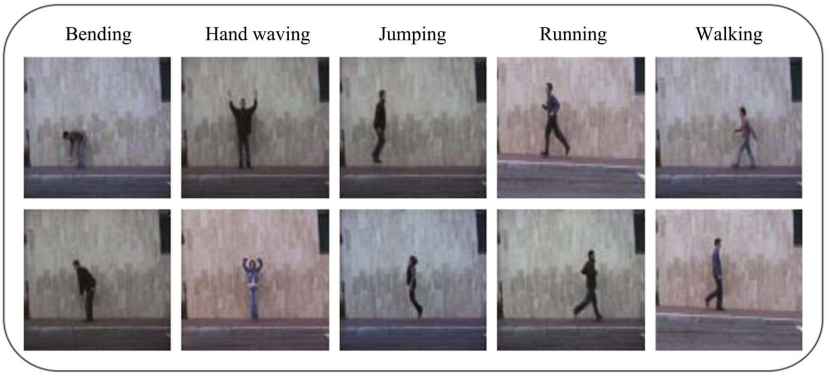


Figure 5. Images from the Weizmann dataset.

4.1.2. KTH Public Datasets

The KTH dataset comprises 1628 images showcasing six distinct types of human activities. Figure 6 displays sample images from the Weizmann datasets, which encompass activities such as boxing, clapping, hand waving, running, and walking [40]. Table 2 presents the combined summary of both datasets (Cn for Class label, n = 1, 2, 3…).

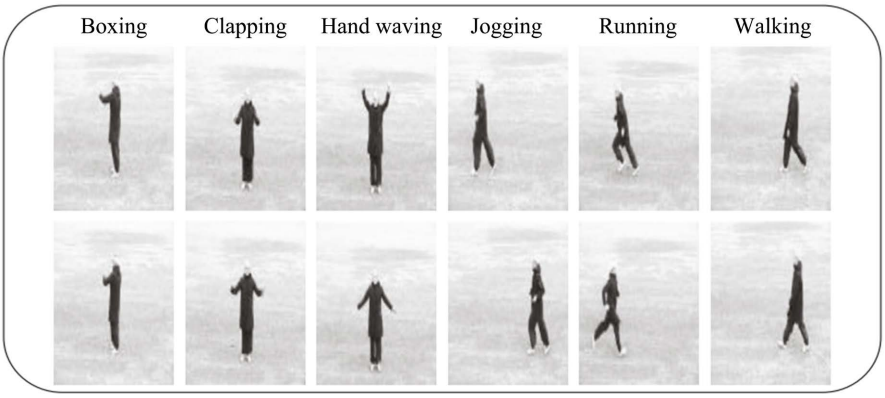


Figure 6. Image from the KTH dataset.

Table 2. Datasets summary.

KTH Dataset			Weizmann Dataset		
Classes	**Activity**	**Images**	**Classes**	**Activity**	**images**
C1	Clapping	312	C1	Hand Waving	624
C2	Jogging	191	C2	Running	206
C3	Hand Waving	581	C3	Jumping	421
C4	Running	109	C4	Walking	271
C5	Walking	27	C5	Bending	375
C6	Boxing	408			
Total images		**1628**	**Total images**		**2513**

4.2. Performance Measures

In Section 4.2 of the research, they checked how well their new algorithm performed. They used a few different measurements to do this:

- **Specificity (SPE):** This checks how good the algorithm is at identifying things that are not what it's looking for.
- **Area Under the Curve (AUC):** This measures how well the algorithm can tell the difference between what it's looking for and what it's not.
- **Precision (PRE):** This looks at how often the algorithm correctly identifies the things it's looking for out of all the things it identifies.
- **Sensitivity (SEN):** This checks how good the algorithm is at finding the

things it's looking for.

- **Accuracy (ACU):** This is a general measure of how often the algorithm is correct in its predictions overall.

These measurements help us to calculate overall the classification performance.

4.3. Experimental Result and Discussion

To quantify the results, three distinct experiments are conducted, each involving a different number of features. Table 3 provides a detailed description of all experiments, specifying the number of classes, folds, and features. When assessing the performance of a machine learning model, a technique called "k-fold cross-validation" is commonly used. This involves dividing the dataset into subsets (folds), training the model on some folds, and evaluating it on others. The process is repeated multiple times. Using different values for k helps in obtaining a more reliable estimate of how well the model performs. After all the runs, the results are averaged to get a comprehensive assessment of the model's effectiveness. This approach provides a more robust evaluation compared to a single train-test split. Each experiment uses five classification methods and calculates sensitivity, specificity, precision, and AUC (Area Under Curve) for the Weizmann and KTH datasets. This helps us see which classification method is the best fit for these specific datasets.

In **Experiment 1**, we observed that cubic-SVM achieved the highest specificity of 99.80% for the Weizmann dataset among all the algorithms. Meanwhile, Fine-KNN attained an accuracy of 99.93%, specifically in terms of precision, for the KTH dataset, as indicated in Table 4.

In **Experiment 2**, we observed that cubic-SVM achieved the highest sensitivity of 99.85% for the Weizmann dataset among all the algorithms. Meanwhile, Subspace-KNN attained an accuracy of 99.94%, specifically in terms of specificity, for the KTH dataset, as indicated in Table 4.

In **Experiment 3**, we observed that cubic-SVM achieved the highest sensitivity of 99.84% for the Weizmann dataset among all the algorithms. Meanwhile, Fine-KNN attained an accuracy of 99.93%, specifically in terms of specificity, for the KTH dataset, as indicated in Table 4.

Table 3. Images, its features and validation for KTH and Weizmann datasets.

Experiment no.	No. of classes				Image Features			Validation
	No. of Images		No. of Images					
	KTH Class	Total Images	Weizmann Class	Total Images	Shape	Texture	Color	Folds
1	6	1628	5	2513	100	60	9	5
2	6		5		300	60	9	10
3	6		5		800	58	9	5

Table 4. Classification results for KTH and Weizmann datasets.

Experiment no.	Methods	Weizmann dataset				KTH dataset			
		Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
1	LS	98.18	99.17	98.55	98.18	99.83	99.92	99.04	99.81
	CS	98.83	99.80	98.96	99.31	99.81	99.90	99.19	99.71
	CT	85.96	97.35	86.16	89.01	99.68	98.28	97.75	98.41
	FK	98.98	99.79	99.33	99.03	99.79	99.77	99.93	99.61
	SK	90.38	98.36	91.75	93.38	99.88	99.87	99.75	99.83
2	LS	98.83	99.76	98.53	98.87	99.84	99.92	99.23	99.82
	CS	98.85	99.84	98.97	99.23	99.82	99.90	99.20	99.73
	CT	85.93	97.36	86.07	89.03	98.46	99.69	97.78	98.42
	FK	98.97	99.79	99.25	99.02	99.76	99.94	99.77	99.62
	SK	90.33	98.36	91.74	93.37	99.85	99.93	99.76	99.82
3	LS	98.87	99.73	98.56	98.86	99.83	99.92	99.23	99.82
	CS	98.86	99.84	98.99	99.36	99.82	99.89	99.24	99.72
	CT	85.91	97.43	86.16	89.54	98.01	99.47	97.67	98.44
	FK	99.67	99.79	99.20	99.05	99.76	99.93	99.77	99.64
	SK	90.34	98.32	91.76	93.38	99.65	99.91	99.73	99.82

LS = Linear-SVM, CS = Cubic-SVM, CT = Complex Tree, FK = Fine-KNN, SK = Subspace-KNN.

4.4. Result Comparison

Table 5 compares the previously implemented algorithms and the proposed algorithm. The table provides a clearer understanding of the performance metrics and highlights the proposed algorithm's superiority. The basis for this conclusion is derived from the discussion that follows.

The proposed algorithm distinguishes itself by incorporating three distinct feature extractors. This strategic combination yields a notable improvement in accuracy compared to the algorithms that were previously implemented. The enhanced accuracy is a result of the synergistic effect created by the integration of these feature extractors, which collectively contribute to a more robust and effective algorithm. In summary, the proposed algorithm surpasses its predecessors in terms of accuracy, making it a promising advancement in the field. The utilization of multiple feature extractors enhances the algorithm's ability to capture and leverage diverse information, leading to improved performance in comparison to existing methods. This reinforces the significance of the proposed approach and its potential for applications requiring high-precision and reliable results.

Table 5. Comparison of action recognition results.

Dataset Name	Reference Paper	Publication Year (Sort by Year)	Accuracy (%)
Weizmann	Li *et al.* [41]	2013	95.43
	JPaul *et al.* [42]	2014	95.54
	Candès *et al.* [43]	2016	88.16
	Imran *et al.* [44]	2016	90.43
	Ahemed Sharif *et al.* [38]	2017	95.88
	S. Aly *et al.* [45]	2019	99.02
	D. K. Vishwakarma *et al.* [46]	2020	96.06
	Our Proposed Method	**2023**	**99.80**
KTH	Le. Shao *et al.* [47]	2014	95.09
	Jain *et al.* [48]	2015	95.23
	J. Yang *et al.* [49]	2015	96.55
	H. Liu *et al.* [50]	2016	97.17
	Ribeiro *et al.* [51]	2017	94.93
	M. Sharif *et al.* [45]	2017	96.36
	Kong *et al.* [52]	2020	94.82
	Ibrahim *et al.* [53]	2019	91.63
	Our Proposed Method	**2024**	**99.94**

5. Conclusion

This study suggests a novel method for identifying and detecting human activity in multimedia frames and films. Preprocessing, feature extraction, feature selection, serial feature fusion, and classification are the five main steps of the algorithm. Through a series of experiments using the KTH and Weizmann datasets, the algorithm demonstrates superior performance, particularly excelling in activities. The study emphasizes the importance of shape features for accurate classification and identifies texture and color features as crucial for detecting various human activities. Additionally, the integration of feature selection and fusion significantly enhances the system's accuracy and sensitivity. The proposed algorithm is much more accurate than existing methods, with a 99.94% accuracy rate on the KTH dataset and 99.80% on the Weizmann dataset. This shows how effective it is at recognizing and classifying human activities. Overall, the research contributes a robust approach to activity detection and classification in multimedia, outperforming current methods.

Author Contributions

- Conceptualization: Subrata Kumer Paul, Md. Atikur Rahman, Md. Ekramul Hamid, Rakhi Rani Paul.

- Methodology: Subrata Kumer Paul, Md. Momenul Haque.
- Data collection and preprocessing: Rakhi Rani Paul, Md. Atikur Rahman.
- Used Software: Subrata Kumer Paul, Md. Ekramul Hamid.
- Writing Original Draft and final copy: Subrata Kumer Paul, Md. Atikur Rahman.
- Overall Supervision: Throughout the project, Md. Ekramul Hamid provided overall supervision, ensuring coherence and adherence to project goals.

Data Availability Statement

The data are available in a publicly accessible repository. The data presented in this study are openly available.

- Weizmann Dataset: https://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
- KTH Dataset: https://www.nada.kth.se/cvap/actions/
- Source code available in this GitHub repository: https://github.com/Subrata11/HAR_Deep_Learning_Architecture

Acknowledgements

I extend my sincere thanks to the Information and Communication Technology Division of the Ministry of Posts, Telecommunication, and Information Technology of Bangladesh, People's Republic of Bangladesh, for their invaluable support and funding of my *ICT fellowship program in the MPhil program*. Additionally, I would like to express my gratitude to my supervisor and co-authors for their guidance and contributions to this research.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Thombre, D.V., Nirmal, J.H. and Lekha, D. (2009) Human Detection and Tracking Using Image Segmentation and Kalman Filter. 2009 *International Conference on Intelligent Agent & Multi-Agent Systems*, Chennai, 22-24 July 2009, 1-5. https://doi.org/10.1109/iama.2009.5228040

[2] Goodrich, M.A. and Schultz, A.C. (2007) Human-robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction*, **1**, 203-275. https://doi.org/10.1561/1100000005

[3] Srinivasan, S., Latchman, H., Shea, J., Wong, T. and McNair, J. (2004) Airborne Traffic Surveillance Systems. *Proceedings of the ACM 2nd International Workshop on Video Surveillance & Sensor Networks*, New York, 15 October 2004, 131-135. https://doi.org/10.1145/1026799.1026821

[4] Tahboub, K., Guera, D., Reibman, A.R. and Delp, E.J. (2017) Quality-Adaptive Deep Learning for Pedestrian Detection. 2017 *IEEE International Conference on Image Processing* (*ICIP*), Beijing, 17-20 September 2017, 4187-4191. https://doi.org/10.1109/icip.2017.8297071

[5] Wang, Q., Tao, Z., Ning, J., Jiang, Z., Guo, L., Luo, H., Wang, H., Men, A., Cheng, X. and Zhang, Z. (2024) Pedestrian Navigation Activity Recognition Based on Segmentation Transformer. *IEEE Internet of Things Journal*, **11**, 26020-26032. https://doi.org/10.1109/jiot.2024.3394050

[6] Ye, Q., Han, Z., Jiao, J. and Liu, J. (2013) Human Detection in Images via Piecewise Linear Support Vector Machines. *IEEE Transactions on Image Processing*, **22**, 778-789. https://doi.org/10.1109/tip.2012.2222901

[7] Conde, C., Moctezuma, D., Martín De Diego, I. and Cabello, E. (2013) Hogg: Gabor and Hog-Based Human Detection for Surveillance in Non-Controlled Environments. *Neurocomputing*, **100**, 19-30. https://doi.org/10.1016/j.neucom.2011.12.037

[8] Satpathy, A., Jiang, X. and Eng, H. (2014) Human Detection by Quadratic Classification on Subspace of Extended Histogram of Gradients. *IEEE Transactions on Image Processing*, **23**, 287-297. https://doi.org/10.1109/tip.2013.2264677

[9] Obaigbena, A., Lottu, O.A., Ugwuanyi, E.D., Jacks, B.S., Sodiya, E.O., Daraojimba, O.D. and Lottu, O.A. (2024) AI and Human-Robot Interaction: A Review of Recent Advances and Challenges. In: *GSC Advanced Research and Reviews* (Vol. 18, Issue 2, pp. 321-330). GSC Online Press. https://doi.org/10.30574/gscarr.2024.18.2.0070

[10] Aboussaleh, I., Riffi, J., Mahraz, A.M. and Tairi, H. (2021) Brain Tumor Segmentation Based on Deep Learning's Feature Representation. *Journal of Imaging*, **7**, Article 269. https://doi.org/10.3390/jimaging7120269

[11] Akbar, S., Sharif, M., Akram, M.U., Saba, T., Mahmood, T. and Kolivand, M. (2019) Automated Techniques for Blood Vessels Segmentation through Fundus Retinal Images: A Review. *Microscopy Research and Technique*, **82**, 153-170. https://doi.org/10.1002/jemt.23172

[12] Fletcher, N.D. and Evans, A.N. (n.d.) Texture Segmentation Using Area Morphology Local Granulometries. In: Ronse, C., Najman, L. and Decencière, E., Eds., *Mathematical Morphology*: 40 *Years On*, Springer-Verlag, 367-376. https://doi.org/10.1007/1-4020-3443-1_33

[13] Paul, S.K., Paul, R.R., Nishimura, M. and Hamid, M.E. (2021) Throat Microphone Speech Enhancement Using Machine Learning Technique. In: Favorskaya, M.N., Peng, SL., Simic, M., Alhadidi, B. and Pal, S., Eds., *Intelligent Computing Paradigm and Cutting-Edge Technologies ICICCT* 2020, Springer International Publishing, 1-11. https://doi.org/10.1007/978-3-030-65407-8_1

[14] Chen, Z., Jiang, C., Xiang, S., Ding, J., Wu, M. and Li, X. (2020) Smartphone Sensor-Based Human Activity Recognition Using Feature Fusion and Maximum Full *a Posteriori*. *IEEE Transactions on Instrumentation and Measurement*, **69**, 3992-4001. https://doi.org/10.1109/tim.2019.2945467

[15] Ding, R., Sun, Q., Liu, M. and Liu, H. (2017) A Compact Representation of Human Actions by Sliding Coordinate Coding. *International Journal of Advanced Robotic Systems*, **14**. https://doi.org/10.1177/1729881417746114

[16] Alavigharahbagh, A., Hajihashemi, V., Machado, J.J.M. and Tavares, J.M.R.S. (2023) Deep Learning Approach for Human Action Recognition Using a Time Saliency Map Based on Motion Features Considering Camera Movement and Shot in Video Image Sequences. *Information*, **14**, Article 616. https://doi.org/10.3390/info14110616

[17] Abdellaoui, M. and Douik, A. (2020) Human Action Recognition in Video Sequences Using Deep Belief Networks. *Traitement du Signal*, **37**, 37-44. https://doi.org/10.18280/ts.370105

[18] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 *IEEE Conference on Computer Vision and Pattern*

Recognition (*CVPR*), Honolulu, 21-26 July 2017, 4724-4733. https://doi.org/10.1109/cvpr.2017.502

[19] Haque, M.M., Paul, S.K., Paul, R.R., Islam, N., Rashidul Hasan, M.A.F.M. and Hamid, M.E. (2023) Improving Performance of a Brain Tumor Detection on MRI Images Using DCGAN-Based Data Augmentation and Vision Transformer (ViT) Approach. In: Solanki, A. and Naved, M., Eds., *GANs for Data Augmentation in Healthcare*, Springer International Publishing, 157-186. https://doi.org/10.1007/978-3-031-43205-7_10

[20] Monteiro, J., Granada, R., Aires, J.P. and Barros, R.C. (2018) Evaluating the Feasibility of Deep Learning for Action Recognition in Small Datasets. 2018 *International Joint Conference on Neural Networks* (*IJCNN*), Rio de Janeiro, 8-13 July 2018, 1-8. https://doi.org/10.1109/ijcnn.2018.8489297

[21] Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T. (2014) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv: 1411.4389. https://doi.org/10.48550/ARXIV.1411.4389

[22] Gupta, T.K. and Raza, K. (2020) Optimizing Deep Feedforward Neural Network Architecture: A Tabu Search Based Approach. *Neural Processing Letters*, **51**, 2855-2870. https://doi.org/10.1007/s11063-020-10234-7

[23] Hara, K., Kataoka, H. and Satoh, Y. (2018) Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet? 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6546-6555. https://doi.org/10.1109/cvpr.2018.00685

[24] Lee, J., Abu-El-Haija, S., Varadarajan, B. and Natsev, A. (2018) Collaborative Deep Metric Learning for Video Understanding. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 481-490. https://doi.org/10.1145/3219819.3219856

[25] Wang, Z., Zheng, Y., Liu, Z. and Li, Y. (2022) A Survey of Video Human Behaviour Recognition Methodologies in the Perspective of Spatial-Temporal. 2022 2*nd International Conference on Intelligent Technology and Embedded Systems* (*ICITES*), Chengdu, 23-26 September 2022, 138-147. https://doi.org/10.1109/icites56274.2022.9943587

[26] Chen, A.T., Biglari-Abhari, M. and Wang, K.I. (2019) Investigating Fast Re-Identification for Multi-Camera Indoor Person Tracking. *Computers & Electrical Engineering*, **77**, 273-288. https://doi.org/10.1016/j.compeleceng.2019.06.009

[27] Kumar, D. and Kukreja, V. (2022) Early Recognition of Wheat Powdery Mildew Disease Based on Mask RCNN. 2022 *International Conference on Data Analytics for Business and Industry* (*ICDABI*), Sakhir, 25-26 October 2022, 542-546. https://doi.org/10.1109/icdabi56818.2022.10041613

[28] Plizzari, C., Cannici, M. and Matteucci, M. (2021) Skeleton-Based Action Recognition via Spatial and Temporal Transformer Networks. *Computer Vision and Image Understanding*, **208**, Article 103219. https://doi.org/10.1016/j.cviu.2021.103219

[29] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. https://doi.org/10.1023/b:visi.0000029664.99615.94

[30] Haindavi, P., Sharif, S., Lakshman, A., Aerranagula, V., Reddy, P.C.S. and Kumar, A. (2023) Human Action Recognition by Learning Spatio-Temporal Features with Deep Neural Networks. *E3S Web of Conferences*, **430**, Article 01154. https://doi.org/10.1051/e3sconf/202343001154

[31] Pang, Y., Jin, A.T.B. and Ling, D.N.C. (2005) A Robust Face Recognition System. *AI* 2005: *Advances in Artificial Intelligence*, Sydney, 5-9 December 2005, 1217-1220. https://doi.org/10.1007/11589990_173

[32] Kumer Paul, S., Ala Walid, M.A., Rani Paul, R., Uddin, M.J., Rana, M.S., Kumar Devnath, M., *et al.* (2024) An Adam Based CNN and LSTM Approach for Sign Language Recognition in Real Time for Deaf People. *Bulletin of Electrical Engineering and Informatics*, **13**, 499-509. https://doi.org/10.11591/eei.v13i1.6059

[33] Mahasseni, B. and Todorovic, S. (2013) Latent Multitask Learning for View-Invariant Action Recognition. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 3128-3135. https://doi.org/10.1109/iccv.2013.388

[34] Paul, S.K., Zisa, A.A., Ala Walid, M.A., Zeem, Y., Paul, R.R., Haque, M.M., *et al.* (2023) Human Fall Detection System Using Long-Term Recurrent Convolutional Networks for Next-Generation Healthcare: A Study of Human Motion Recognition. 2023 14*th International Conference on Computing Communication and Networking Technologies* (*ICCCNT*), Delhi, 6-8 July 2023, 1-7. https://doi.org/10.1109/icccnt56998.2023.10308247

[35] Bobick, A.F. and Davis, J.W. (2001) The Recognition of Human Movement Using Temporal Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 257-267. https://doi.org/10.1109/34.910878

[36] Patel, C.I., Garg, S., Zaveri, T., Banerjee, A. and Patel, R. (2018) Human Action Recognition Using Fusion of Features for Unconstrained Video Sequences. *Computers & Electrical Engineering*, **70**, 284-301. https://doi.org/10.1016/j.compeleceng.2016.06.004

[37] Gutoski, M., Lazzaretti, A.E. and Lopes, H.S. (2020) Deep Metric Learning for Open-Set Human Action Recognition in Videos. *Neural Computing and Applications*, **33**, 1207-1220. https://doi.org/10.1007/s00521-020-05009-z

[38] Malhi, A. and Gao, R.X. (2004) PCA-Based Feature Selection Scheme for Machine Defect Classification. *IEEE Transactions on Instrumentation and Measurement*, **53**, 1517-1525. https://doi.org/10.1109/tim.2004.834070

[39] Das, A., Mitra, A., Bhagat, S.N. and Paul, S. (2020) Issues and Concepts of Graph Database and a Comparative Analysis on List of Graph Database Tools. 2020 *International Conference on Computer Communication and Informatics* (*ICCCI*), Coimbatore, 22-24 January 2020, 1-6. https://doi.org/10.1109/iccci48352.2020.9104202

[40] Schuldt, C., Laptev, I. and Caputo, B. (2004) Recognizing Human Actions: A Local SVM Approach. *Proceedings of the* 17*th International Conference on Pattern Recognition*, 2004. *ICPR* 2004, Cambridge, 26 August 2004, 32-36. https://doi.org/10.1109/icpr.2004.1334462

[41] Li, R., Yun, L., Zhang, M., Yang, Y. and Cheng, F. (2023) Cross-View Gait Recognition Method Based on Multi-Teacher Joint Knowledge Distillation. *Sensors*, **23**, Article 9289. https://doi.org/10.3390/s23229289

[42] Paul, R.R., Paul, S.K. and Hamid, M.E. (2022) A 2D Convolution Neural Network Based Method for Human Emotion Classification from Speech Signal. 2022 25*th International Conference on Computer and Information Technology* (*ICCIT*), Cox's Bazar, 17-19 December 2022, 72-77. https://doi.org/10.1109/iccit57492.2022.10054811

[43] Candès, E.J. (2008) The Restricted Isometry Property and Its Implications for Compressed Sensing. *Comptes Rendus. Mathématique*, **346**, 589-592. https://doi.org/10.1016/j.crma.2008.03.014

[44] Imran, J. and Raman, B. (2019) Deep Motion Templates and Extreme Learning Ma-

chine for Sign Language Recognition. *The Visual Computer*, **36**, 1233-1246. https://doi.org/10.1007/s00371-019-01725-3

[45] Sharif, M., Khan, M.A., Akram, T., Javed, M.Y., Saba, T. and Rehman, A. (2017) A Framework of Human Detection and Action Recognition Based on Uniform Segmentation and Combination of Euclidean Distance and Joint Entropy-Based Features Selection. *EURASIP Journal on Image and Video Processing*, **2017**, Article No. 89. https://doi.org/10.1186/s13640-017-0236-8

[46] Aly, S. and Sayed, A. (2019) An Effective Human Action Recognition System Based on Zernike Moment Features. 2019 *International Conference on Innovative Trends in Computer Engineering* (*ITCE*), Aswan, 2-4 February 2019, 52-57. https://doi.org/10.1109/itce.2019.8646504

[47] Le, V., Tran-Trung, K. and Hoang, V.T. (2022) A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition. *Computational Intelligence and Neuroscience*, **2022**, Article 8323962. https://doi.org/10.1155/2022/8323962

[48] Jain, S.B. and Sreeraj, M. (2015) Multi-posture Human Detection Based on Hybrid HOG-BO Feature. 2015 *Fifth International Conference on Advances in Computing and Communications* (*ICACC*), Kochi, 2-4 September 2015, 37-40. https://doi.org/10.1109/icacc.2015.99

[49] Yang, J., Ma, Z. and Xie, M. (2015) Action Recognition Based on Multi-Scale Oriented Neighborhood Features. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, **8**, 241-254. https://doi.org/10.14257/ijsip.2015.8.1.21

[50] Liu, H., Ju, Z., Ji, X., Chan, C.S. and Khoury, M. (2017) Study of Human Action Recognition Based on Improved Spatio-Temporal Features. In: Liu, H., Ju, Z., Ji, X., Chan, C.S. and Khoury, M., Eds., *Human Motion Sensing and Recognition*, Springer, 233-250. https://doi.org/10.1007/978-3-662-53692-6_11

[51] Ribeiro, M., Lazzaretti, A.E. and Lopes, H.S. (2018) A Study of Deep Convolutional Auto-Encoders for Anomaly Detection in Videos. *Pattern Recognition Letters*, **105**, 13-22. https://doi.org/10.1016/j.patrec.2017.07.016

[52] Kong, Y. and Fu, Y. (2022) Human Action Recognition and Prediction: A Survey. *International Journal of Computer Vision*, **130**, 1366-1401. https://doi.org/10.1007/s11263-022-01594-9

[53] Ibrahim, M.J., Kainat, J., AlSalman, H., Ullah, S.S., Al-Hadhrami, S. and Hussain, S. (2022) An Effective Approach for Human Activity Classification Using Feature Fusion and Machine Learning Methods. *Applied Bionics and Biomechanics*, **2022**, Article 7931729. https://doi.org/10.1155/2022/7931729