Journal of Artificial Intelligence and Technology, 2024, 4, 208-215
https://doi.org/10.37965/jait.2024.0351

ISTP
RESEARCH ARTICLE

Vision-based Human Activity Recognition Using Local Phase Quantization

Madhuri Pandey, Richa Mishra, and Ashish Khare
Department of Electronics and Communication, University of Allahabad, Allahabad, India

(Received 29 May 2023; Revised 13 May 2024; Accepted 24 May 2024; Published online 09 June 2024)

Abstract: Human activity recognition (HAR) has been the most active and interesting area of research in recent years due to its wide range of applications in the field, such as healthcare, security and surveillance, robotics, gaming, and entertainment. However, recognizing vision-based human activity is still a challenging as input sequences may have cluttered background, illumination conditions, occlusions, degradation of video quality, blurring, etc. In the literature, several state-of-the-art methods have been trained and tested on different datasets but have yet to perform adequately to a certain extent. Moreover, extracting potential features and combining appropriate methods is one of the most challenging tasks in realistic video. This paper proposes an efficient frequency-based blur invariance local phase quantization feature extractor and multiclass SVM classifier that overcomes these challenges. The feature is invariant toward camera motion, misfocused optics, movements in the scene, and environmental conditions. The proposed feature vector is then fed to the classifier to recognize human activities. The experiment has conducted on two publicly available datasets, UCF101 and HMDB51, and has achieved 99.79% and 98.67% accuracies, respectively. The approach has also outperformed the existing state-of-the-art approaches in terms of computational cost without compromising the accuracy of HAR.

Keywords: LPQ; machine learning; SVM; texture-based; vision-based HAR

I. INTRODUCTION

Tremendous growth and development in the field of sensor-based devices and multi-view cameras constantly enhance the popularity of human activity recognition (HAR). It is widely used in several applications such as security and surveillance [1,2], video retrieval [3], human–robot interaction [4], human–computer interaction [5], abnormal activity detection [6], entertainment [7], etc. In the literature [8,9], HAR is broadly classified into sensor-based and vision-based activity recognition. Dang *et al.* [8] classified HAR into sensor-based and vision-based based on data collection, pre-processing methods, feature engineering, and the training process. Hussain *et al.* [10] divided sensor-based HAR into three types based on sensor deployment. They are wearable, object-tagged, and dense sensing. In sensor-based HAR, activity is recognized by sensors or machines like accelerometer, gyroscope, smart watches, etc. Sensor-based HAR has gained popularity due to its small size, low cost, and better accuracy. However, wearing a tag throughout the day is inconvenient, requires more power dependency, depends on environmental conditions, etc [10]. On the other hand, vision-based HAR is based on video or sequences of frame captured by multi-view cameras. Hence, there is no need to wear the tag to gain popularity. Despite its merits, the performance has faced challenges, such as illuminance change, moving camera, blurring, occlusion, jilter, inter-class similarity, intra-class variability, moving background, and noise. To overcome these limitations, conventional or handcrafted, deep learning (DL) and hybrid methods have been introduced in the literature [11–13]. These approaches give prominent results on various publicly available datasets. However, handcrafted methods require domain experts to perform feature extraction. Further, these features are fed into a suitable classifier for recognizing a human activity. On the other hand, DL methods extract features automatically from raw input data. It gives the concept of end-to-end learning and able to replace handcrafted approach [12]. However, handcrafted features are still prominently used due to some limitations in DL approach. Some of the best-performing DL-based methods still dependent on handcrafted features, and these methods have a higher computational cost [12].

This paper proposes a novel handcrafted feature extraction approach to recognize human activity. The proposed method uses a frequency-based feature extractor named local phase quantization (LPQ) to find an efficient feature vector to recognize human activity in an unconstrained and realistic environment. The feature vector is then fed to a one-to-one multiclass support vector machine (SVM) classifier to recognize human activity. In literature, several combinations of handcrafted feature descriptors have been used to achieve comparable accuracy, but no one can achieve similar accuracy using a single feature. This paper uses a single feature descriptor, which gives approximately the same or greater accuracy on complex datasets. It is simple and efficient and outperforms a combination of features (handcrafted, DL, and hybrid). The main contributions of the proposed approach are listed below:

1. The prominent role of our current research is to observe and find appropriate features that can increase the efficiency of HAR and overcome the degradation of image quality due to the presence of camera motion, misfocused optics, movements in the scene, and environmental condition, inter-class similarity, and intra-class variability. Such efficacy is more important when dealing with complex video in an unconstrained environment.

Corresponding author: Richa Mishra (email: richa_mishra@allduniv.ac.in).

2. A frequency-based LPQ feature descriptor has generated an efficient feature vector. The generated vector is fed as an input to the SVM classifier to recognize the human activity.
3. The approach has been trained and tested on publicly available datasets: University of Central Florida 101 (UCF101) [14] and Human Motion Database 51 (HMDB51) [15].
4. The approach has also outperformed the existing state-of-the-art approaches regarding computational efficiency.

The rest of the paper is divided into the following sections: Section II provides an overview of related work. Section III presents an overview of the proposed approach. Section IV discusses the experimental results and compared with the state-of-the-art techniques. The last section gives the conclusion and future scope.

II. LITERATURE REVIEW

The process of HAR in a realistic environment is still a challenging task. It suffers from different blurring, illuminance change, partial occlusion, complex background and foreground, complex structural arrangement of pixels, etc. Therefore, machine learning (ML) and DL-based approaches have been extensively used [11]. HAR mainly depends on how efficiently features are extracted from videos. Pareek and Thakkar [16] categorize features used in ML-based methods. They are trajectory-based, motion-based, texture-based, shape-based, gait-based, etc. Abdul-Azim and Hemayed [17] proposed trajectory-based HAR using the discriminative temporal relationship based on scale-invariant feature transform (SIFT) descriptor and SVM classifier. They obtained 95.36%, 97.77%, and 89.99% accuracies on Kungliga Tekniska högskolan (KTH), Weizmann, and UCF sports datasets, respectively. In motion-based techniques, optical flow and spatiotemporal filtering are used for capturing the motion of the target frame [16]. Bobick and Davis [18] proposed a motion-based HAR, invariant to linear changes in speed and run in real time. The shape-based feature provides human body dynamics and structure [16]. Vishwakarma *et al.* [19] present a shape-based feature with SVM classifier and found 100%, 85.80%, 95.5%, 93.25%, and 92.92% accuracies on Weizmann, INRIA Xmas Motion Acquisition Sequences (IXMAS), KTH, Ballet Movement, and Multi-view i3dPost datasets, respectively. Gupta *et al.* [20] proposed a gait-based feature for HAR on KTH and Weizmann datasets with 95.01% and 91.36% accuracies, respectively. Texture-based features provide a structural arrangement of intensities in an image. There are several texture-based features, and their combinations are available in literature, such as local binary pattern (LBP), local ternary pattern (LTP), LBP on three orthogonal planes (LBP-TOP), etc. [21–25]. Rahman *et al.* [26] proposed a HAR based on the combination of motion history image (MHI), LBP, and histogram of oriented gradients (HOG) feature with an SVM classifier. They obtained 86.67% and 94.3% accuracies on the KTH Action and the Pedestrian Action datasets, respectively. Wang *et al.* [27] have used the combination of HOG and LBP feature with a linear SVM classifier to recognize partial occluded human on National Institute for Research in Digital Science and Technology (INRIA) dataset with 97.9% accuracy. Carmona and Climent [28] used the template-based method to recognize human actions with Improved Dense Trajectories. The results were tested, and 89.3% and 65.3% accuracies were obtained on the UCF101 and HMDB51 datasets, respectively. Wang *et al.* [29] proposed a dense trajectories and motion boundary histogram (MBH)-based descriptors for HAR on nine publicly available datasets, KTH, YouTube, Hollywood2, UCF sports, IXMAS, University of Illinois Urbana Champaign (UIUC), Olympic Sports, UCF50, and HMDB51 with 94.2%, 84.1%, 58.2%, 88.0%, 93.5%, 98.4%, 74.1%, 84.5%, and 46.6% accuracies, respectively. Neggaz and AbdElminaam [30] used the combination of moment invariant (MI) mean block discrete cosine transform (MmDCT) and uniform LBP feature and neural network classifier for HAR. They have achieved 92.54%, 99.97%, and 99.9% accuracies on KTH, UCF11, and HMDB51 (for six activities) datasets, respectively. However, the approach fails to recognize the complex activities of UCF101 and HMDB51 datasets. Kushwaha *et al.* [31] proposed a linear feature fusion of optical flow and HOG feature for HAR with SVM classifier. The approach has achieved 99.3%, 97.96%, and 97.18% accuracies on UT Interaction, Institute of Automation, Chinese Academy of Sciences (CASIA), and HMDB51 datasets, respectively.

In a DL-based approach, Yu *et al.* [32] have proposed stratified pooling-based convolutional neural network (CNN) for HAR. The model has tested on UCF101 and HMDB51 datasets with 91.6% and 74.7% accuracies, respectively. Varol *et al.* [33] present a long-term temporal convolution (LTC)-based CNN model for activity recognition on the UCF101 and HMDB51 datasets and achieved 92.7% and 67.2% accuracies, respectively. Geng and Song [34] have proposed CNN-based HAR in which CNN is used for feature extraction, and SVM is used for pattern recognition. The result was tested on the KTH dataset and achieved 92.49% accuracy. Basak *et al.* [35] have recognized HAR using DL and swarm intelligence-based metaheuristic model. They have got 98.13%, 90.67%, and 89.98% accuracies on University of Texas at Dallas-Multimodal Human Action Dataset (UTD-MHAD), HMDB05, and Nanyang Technological University's Red Blue Green and Depth information (NTU RGB+D) 60 datasets, respectively. Tran *et al.* [36] showed accuracy improvement of 3D CNNs over 2D CNNs using the residual learning framework on Sports-1M, Kinetics, UCF101, and HMDB51 datasets. Xia *et al.* [37] have proposed the combination of long short-term memory (LSTM) and CNN-based model for activity recognition in which data is fed into the LSTM network followed by CNN. The result has tested on three datasets: University of California Irvine (UCI), Wireless Sensor Data Mining (WISDM), and OPPORTUNITY, with 95.78%, 95.85%, and 92.63% accuracies, respectively. Yin *et al.* [38] have proposed 1-D CNN-based bidirectional LSTM parallel model with an attention mechanism for activity recognition. The performance was tested on UCI and WISDM HAR datasets and found 96.71% and 95.86% accuracies, respectively. Huang *et al.* [39] have proposed Three-Stream Network model in which spatial, temporal, and sequential features are fused. Further, features are fed into a multilayer perception classifier to recognize human activity. The results are tested on UCF11, UCF50, and HMDB51 datasets with 99.17%, 97.40%, and 96.88% accuracies, respectively. Luo *et al.* [40] have used Dense Semantics-Assisted Networks for HAR. They have used a dense semantic segmentation mask to encode the semantics for network training and improve the accuracy of the proposed network. The network has been tested on UCF101, HMDB51, and Kinetics datasets and achieved 96.69%, 72.88%, and 76.52% accuracies, respectively. Majd and Safabakhsh [41] have proposed a correlational convolution LSTM network for capturing the motion and dependency between spatial and time of the input videos for activity recognition. They have achieved 93.6% and 66.2% accuracies on UCF101 and HMDB51 datasets, respectively. Wang *et al.* [42] have proposed CNN-based Semantic

Action-Aware Spatial-Temporal Features for action recognition with 71.2%, 45.6%, 95.9%, and 74.8% accuracies on Kinetics-400, Something-Something-V1, UCF101 and HMDB51 datasets, respectively. Xia and Wen [43] proposed a multi-stream based on key frame sampling for HAR. This framework consists three parts. First, it can use a self-attention mechanism to find the relationship between different regions; second, a key frame sampling mechanism is used to select a different video frame. Lastly, a deep spatiotemporal feature extraction mechanism is used to generate a fine grain feature, which is used for classification task. They have obtained 79.5%, 97.6%, 84.2%, and 71.6% accuracies on HMDB51, UCF101, Kinetics 400, and Something-Something Dataset, respectively. Saoudi *et al.* [44] used 3D CNN for feature extraction, followed by LSTM network with an attention mechanism for HAR. The network is tested on UCF101 and HMDB51 datasets and achieved 97.98% and 96.83% accuracies, respectively. Chen *et al.* [45] proposed improved residual CNNs with spatial attention modules. They have found 95.68% and 72.6% accuracies on UCF101 and HMDB51 datasets, respectively.

HAR based on the frequency domain are well known due to its robustness against blurring, geometric changes, and intensity changes [46]. In addition, it is computationally efficient for implementation [47,48]. Tran *et al.* [49] have used the frequency domain features to minimize the variability of effect and achieved 82.7% accuracy on the KTH dataset. Kumara *et al.* [50] have used foreground/background segmentation as preprocessing, discrete Fourier transform for feature extraction, and K-nearest neighbor (KNN) as a classifier for HAR. Lei *et al.* [51] have proposed a frequency-based descriptor for face recognition in low resolution. Ojansivu *et al.* [52] use local low-frequency Fourier phase information for rotation and blur-insensitive texture analysis. Ahonen *et al.* [53] use an LPQ-based feature to recognize blurred faces. Briassouli [54] describes motion statistics without requiring the estimation of optical flow, and it is useful where appearance features could be more informative. Foroosh *et al.* [55] have suggested estimating subpixel shifts using the phase correlation method. Feng *et al.* [56] have proposed 3D human skeleton-based HAR. They have partitioned the scene into several primitive actions based on the motion attention mechanism. They extract features from primitive motion and feed them into CNN architecture to recognize human action.

In this paper, we use the LPQ feature to overcome two main limitations of the HAR system: illumination variation and image blurring. Further, the result of the LPQ feature is fed into SVM classifier to recognize human activity. Moreover, the following section compares the results with those of the existing state-of-the-art approaches.

III. THE PROPOSED METHODOLOGY

This paper proposes a blur-invariant frequency-based feature descriptor for HAR. The descriptor can overcome illuminance changes and image degradation caused by camera motion, misfocused optics, movements in scene, and environmental condition. Work has been done to overcome the aforesaid limitations by considering a single or combination of different feature descriptors and feeding to ML classifiers [16]. However, there is a need of improvement in the performance of HAR in a blurring environment using single feature. It motivates the development of a robust blur-invariant LPQ feature for HAR on complex datasets. LPQ extracts texture information in a frequency domain. The extracted features are further fed into classifier to recognize human activity. Figure 1 shows an overview of the proposed approach. Frames are extracted in a video sequence, converted into grayscale image and resized to 128×128. Once data is preprocessed, texture-based LPQ features are extracted, and then the concatenation of the histogram of the features is fed to the SVM classifier to recognize activities.

A. FEATURE EXTRACTION

In this step, the proposed feature vector is the concatenation of the histogram of all frames in a video. LPQ feature is first proposed by Ojansivu and Heikkila [57]. It is a blur-invariant feature descriptor. Blurring is one of the issues frequently occurring in vision-based HAR by camera motion, misfocused optics, and movements in the scene. In literature, the LPQ feature is used in various applications like facial expression recognition, person authentication system, and so on [51–53,58].

It shows the blur invariance property of an image by separating the magnitude and phase part of the discrete Fourier transform of the blurred image $G(u)$. The blurred point spread function gives only two valued functions of the phase. In image processing, Fourier transform converts the signal from the spatial domain to the frequency domain. According to the Fourier transform property, a signal change in the spatial domain can reflect the change in the frequency domain. Let $f(x)$ be an image. The corresponding frequency domain representation of image $f(x)$ is obtained mathematically by the short-term Fourier transform (STFT) and is shown in equation (1) [57]:

$$F(u,x) = \sum_{y \in N_x} f(x-y) e^{-j2\pi u^T y} = w_u^T f_x \tag{1}$$

where $F(u,x)$ represents the transformed image at frequency u and pixel x, w_u represents basis vector of an image $f(x)$, f_x is a vector containing all samples from neighborhood 3×3 pixel (N_x), and w_u^T is the transpose of w_u.

In LPQ, the transformation matrix is computed by separating the real and imaginary parts of the four lowest frequencies using STFT and is shown in equation (2). Further, the covariance matrix of the transformed frequency F_x is obtained by assuming that the original image function $f(x)$ is a result of a first-order Markov process and obtained from the equation (3). The whitening transform G_x is calculated by using equation (4). Lastly, scalar quantization of G_x is done by using equation (5). The range of q_j is from 0 to 255 is shown by equation (6) [51]:

$$F_x = W f_x \tag{2}$$

$$D = WCW^T \tag{3}$$

$$G_x = V^T F_x \tag{4}$$

$$q_j = \begin{cases} 1, & if\ g_j \geq 0 \\ 0, & otherwise \end{cases} \tag{5}$$

$$b = \sum_{j=1}^{8} q_j 2^{j-1} \tag{6}$$

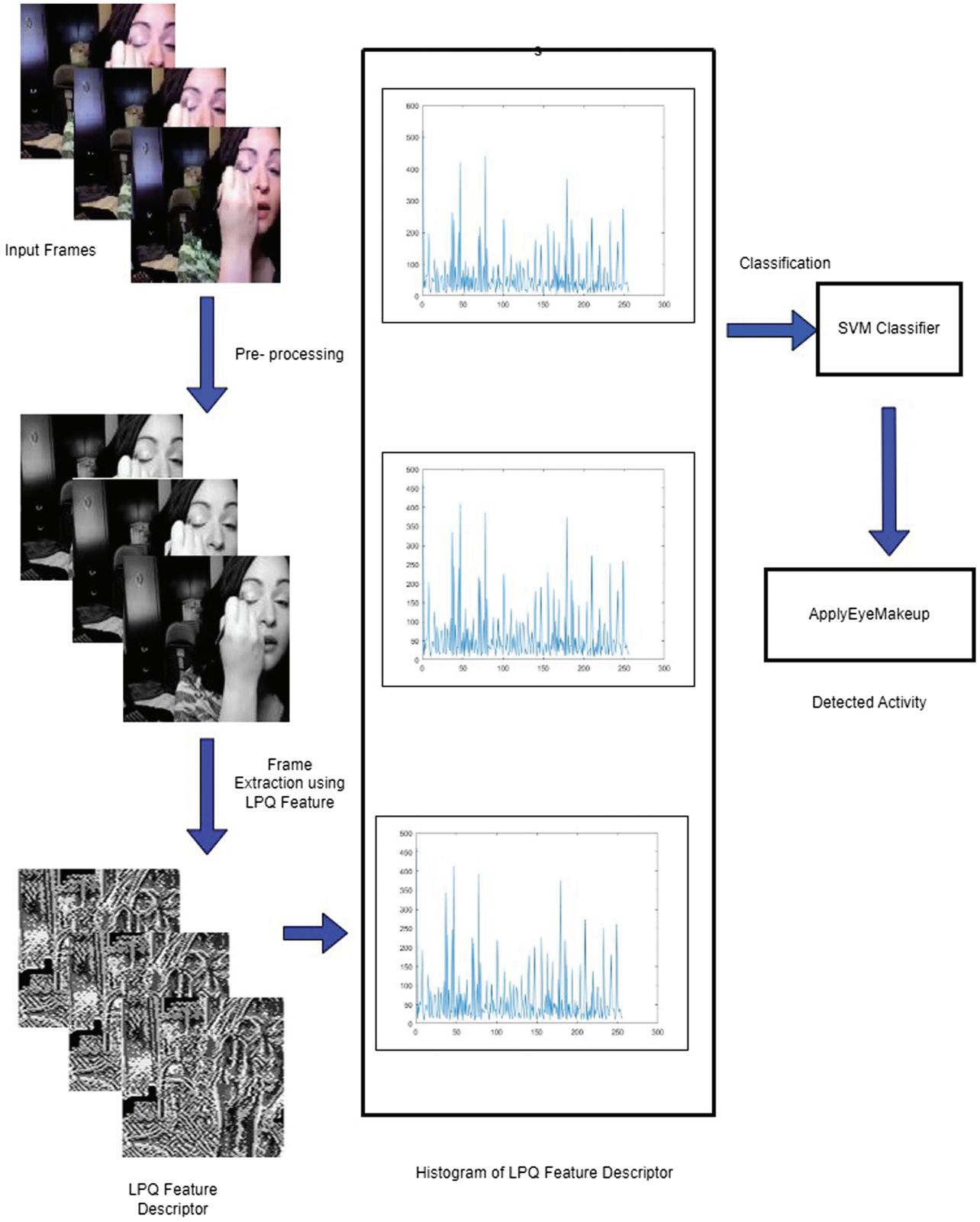

Fig. 1. An overview of the proposed approach.

B. LPQ ENCODING

LPQ coding and formation of histogram is similar to LBP. However, LBP feature extraction is done in spatial domain and LPQ work in frequency domain. LPQ is a local texture feature which outperforms LBP and LTP. The process of LPQ encoding is shown in Fig. 2.

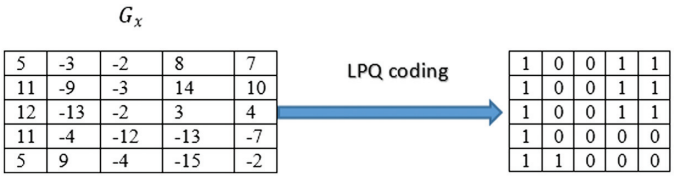

Fig. 2. Process of the LPQ encoding.

C. CLASSIFICATION

SVM [59–61] is one of the most commonly used supervised ML approaches for classification. The goal of SVM is to find optimal separating plane/hyperplane that can divide *n*-dimensional space into different classes. Here, we can ensure that the distance of hyperplane from its nearest data point on every class is maximum which is called the optimal hyperplane. The maximum distance of the data point to the hyperplane is called the margin. In this experiment, a one-to-one SVM classifier has been used to detect human activity using the LPQ feature in the *n*-dimensional plane.

IV. EXPERIMENTAL RESULTS & DISCUSSION

A. DATASET DESCRIPTION

The performance of the proposed approach has been extensively trained and tested on UCF101 [14] and HMDB51 [15] datasets. The datasets contain videos of different resolutions. The video files in the HMDB51 [15] dataset have different resolutions like 352×240, 320×240, 416×240, 424×240, etc., whereas the UCF101 [14] dataset has the exact resolution 320×240 for all the videos. For the simplicity of the experiment, we resized the video frame into 128×128. The entire dataset has been divided into training and testing dataset and are in the ratio of 7:3.

1. UCF101 DATASET. UCF101 [14] dataset has 101 human activity classes, which are divided into five categories: Human–object interaction, body–motion, human–human interaction, playing musical instruments, and sports. It has a fixed frame rate of 25 FPS and fixed resolution of 320×240. UCF101 dataset video includes camera motion, different lighting conditions, partial occlusion, cluttered background, and low-quality frame. Sample of the video frames of the dataset is shown in Fig. 3.

2. HMDB51 DATASET. HMDB51 [15] has 51 human action categories collected from different sources like You Tube and Google videos. They are grouped into five classes: general facial actions: smile, laugh, crew, talk, etc.; facial action with object manipulation: smoke, eat, drink; general body movements: cartwheel, clap hands, climb stairs, etc.; body movement with object interaction: brush hair, catch, draw sword, golf, kick ball, ride horse, etc.; and body movement for human interaction: fencing, hug, kick some one, punch, shake hands. This dataset is most realistic videos including complex high-level activities having a complex background, camera motion, and varying luminance. Sample of the video frames of the dataset is shown in Fig. 4.

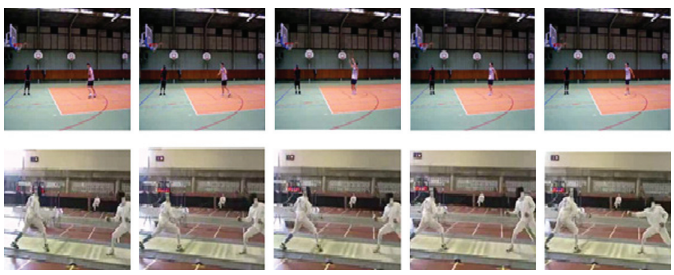

Fig. 3. Sample frame of UCF101 dataset.

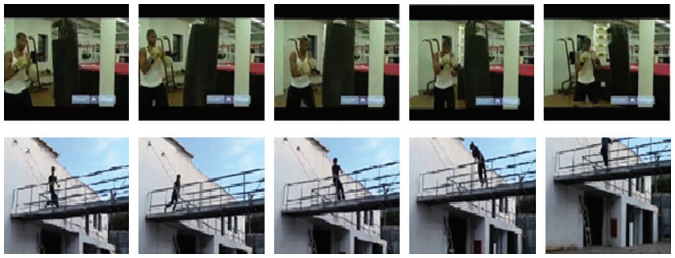

Fig. 4. Sample frame of HMDB51 dataset.

B. PERFORMANCE METRICS

To prove the authenticity and efficiency of the proposed approach, five performance metrics have been calculated named as accuracy, precision, sensitivity, specificity, and F-measure [62].

1. ACCURACY. Accuracy is the ratio between the correctly classified object to the total number of objects to be tested. Mathematically, accuracy can be written as:

$$\text{Accuracy} = (\text{correctly classified object}/\text{ total no. of object tested}) \times 100$$

2. PRECISION. Precision is calculated as the number of true positivity (T_P) divided by the sum of true positivity and false positivity (F_P). Mathematically, accuracy can be written as:

$$\text{Precision} = T_P/(T_P + F_P)$$

3. SENSITIVITY. Sensitivity is calculated as the number of true positivity (T_P) divided by sum of true positivity (T_P) and false negativity (F_N). Mathematically, accuracy can be written as:

$$\text{Sensitivity} = T_P/(T_P + F_N)$$

4. SPECIFICITY. Specificity is calculated as the number of true negativity (T_N) divided by sum of the false positivity (F_P) and true negativity (T_N).

$$\text{Specificity} = T_N/(T_N + F_P)$$

5. F1. It is the harmonic mean of precision and recall:

$$\text{F1} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

V. RESULTS AND DISCUSSION

In this section, the proposed approach has been analyzed on the mentioned datasets. Table I shows the performance of LPQ feature

Table I. The performance metrics of HAR on UCF101 and HMDB51 dataset

Performance metrics	UCF101 dataset	HMDB51 dataset
Accuracy (%)	99.78	98.67
Precision	.9978	.9867
Sensitivity	.9978	.9867
Specificity	.1000	.9997
F-measure	.9978	.9867

and SVM classifier-based HAR on five different metrics. The approach has achieved the considerable level of accuracy on the challenging dataset. It is found that the approach has obtained less accuracy on HMDB51 [15] dataset due to some inter-class variability. It is one of the common problems in which occurrences of intermediate actions of two activities from different class are same that results misidentification of accurate actions.

Table II compares the proposed approach's accuracy on both datasets with other state-of-the-art approaches. The table shows that the proposed approach has outperformed other handcrafted and DL approaches and is shown in the table as a boldface.

First, we compare the proposed approach with handcrafted-based method [28–31]. Neggaz and AbdElminaam [30] have achieved the approximately same accuracy as the proposed approach. They have performed the HAR using three features: MmDCT, Uniform Local Binary Patterns (ULBPs), and MIs. However, it has more mathematical computation to make an efficient feature vector. The dimension of the feature increases after the fusion of these three features. They have conducted the experiment on UCF101 [14] and HMDB51 [15] datasets by considering only six activities from both the datasets, respectively. However, the performance is limited to recognize complex activity of the stated dataset. Carmona and Climent [28] have used the template-based method for HAR which is combined with Improved Dense Trajectories. In this work, the Improved Dense Trajectories is obtained by the feature fusion of HOG, histogram of optical flow (HOF), and MBH feature. The resultant feature descriptor is more complex in terms of computation and size of feature vector than use of a single feature. Wang *et al.* [29] proposed a dense trajectories and MBH-based descriptors for HAR. It requires more time to combine dense trajectories and MBH features. Kushwaha *et al.* [31] used a linear feature fusion of optical flow and HOG feature for HAR with SVM classifier. The approach extracts large feature by HOG and optical flow combination. It results more computational time compared to single feature for extracting feature as well as classifying activity. On the other hand, the proposed approach is outperforming various DL methods [39–45] shown in Table II. The major challenge with DL methods is that it requires large sampled data to classify actions efficiently. It requires high computational cost, more memory, and large amount of input data. In addition, the proposed approach has also compared with Mohan *et al.* [63]. They have used the combination of HOG+LPQ with Fuz-SVM classifier for Object Face Liveness Detection on their own dataset. In this paper, the combination of HOG+LPQ feature with multiclass SVM has been tested on HMDB51 dataset and found the result shown in Table III. The single feature, that is, LPQ, has given comparable results in terms of accuracy. However, the number of features extracted in the combination is approximately 32 times higher than the number of features extracted by single feature. Moreover, the computational time required for feature extraction per video and classification also takes more time in the combination as compared to single feature.

Therefore, the proposed approach has outperformed the state-of-the-art approaches over HMDB51 [15] and UCF101 [14] datasets. HMDB51 [15] contains more blurred images compared to UCF101 [14]. Hence, the proposed approach is efficient and invariant toward blurring.

Table II. Comparison of proposed approach and the current state-of-art approaches on UCF101 and HMDB51 dataset

Author(s)	Handcrafted/deep learning-based feature	Feature(s)	Accuracy (%) for UCF101	Accuracy (%) for HMDB51
Carmona and Climent [28]	Handcrafted	Trajectory	89.30	65.30
Wang *et al.* [29]	Handcrafted	Trajectory and motion	89.10	48.30
Kushwaha *et al.* [31]	Handcrafted	Motion and shape	97.18	–
Neggaz and AbdElminaam [30]	Handcrafted	Texture and shape	99.97	99.92
Luo *et al.* [40]	Deep learning	Spatial and temporal	98.12	77.35
Majd and Safabakhsh [41]	Deep learning	Spatial and temporal	93.60	66.20
Wang *et al.* [42]	Deep learning	Spatial and temporal	95.90	74.80
Huang *et al.* [39]	Deep learning	Spatial, temporal, and sequential	97.40	96.88
Xia and Wen [43]	Deep learning	Temporal and sequential	97.6	79.5
Saoudi *et al.* [44]	Deep learning	Spatial and temporal	97.98	96.83
Chen *et al.* [45]	Deep learning	Spatial and temporal	95.68	72.60
Proposed method	**Handcrafted**	**Texture**	**99.78**	**98.67**

Table III. Comparison between HOG+LPQ and LPQ feature with SVM classifier on HMDB51 dataset

Feature combination	# Feature extracted	Execution time for feature extraction (per video) in sec.	Execution time for classification in sec.	Accuracy (%)
HOG+LPQ	8356	44.57	3654.47	99.48
LPQ	256	37.17	158.75	98.67

VI. CONCLUSION AND FUTURE DIRECTIONS

In this paper, the frequency-based technique has been used to recognize human activity. Here, LPQ feature descriptor has been considered to extract useful feature from input video and then fed into one-to-one SVM classifier to recognize human activity. The feature is more robust against the illuminance change and blurring caused by the camera motion, misfocused optics, movements in the scene, and environmental conditions. The proposed approach has achieved 99.78% and 98.67% accuracies on UCF101 and HMDB51 datasets, which is more complex and challenging datasets in HAR. The approach has also outperformed the existing state-of-the-art approaches in terms of computational cost without compromising the accuracy of HAR.

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: a survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.

[2] P. K. Roy and H. Om, "Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos," *Advances in Soft Computing and Machine Learning in Image Processing*. Cham: Springer, 2018, pp. 277–294.

[3] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 1168–1181, 2007.

[4] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," *2016 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2016, pp. 2702–2706, 2016.

[5] M. M. Islam, M. R. Islam, and M. S. Islam, "An efficient human computer interaction through hand gesture using deep convolutional neural network," *SN Comput. Sci.*, vol. 1, p. 211, 2020.

[6] B. Delgado, K. Tahboub, and E. J. Delp, "Automatic detection of abnormal human events on train platforms," *NAECON 2014-IEEE Natl. Aerosp. Electronics Conf.*, vol. 2014, pp. 169–173, 2014.

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *CVPR*, vol. 2011, pp. 1297–1304, 2011.

[8] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: a comprehensive survey," *Pattern Recognit.*, vol. 108, p. 107561, 2020.

[9] P. Girdhar, "Vision based human activity recognition: a comprehensive review of methods & techniques," *Turk. J. Comput. Math. Educ. (TURCOMAT)*, vol. 12, no. 10, pp. 7383–7394, 2021.

[10] Z. Hussain, Q. Z. Sheng, and W. Emma Zhang, "A review and categorization of techniques on device-free human activity recognition," *J. Netw. Comput. Appl.*, vol. 167, p. 102738, 2020.

[11] C.-Y. Zhang, Y.-Y. Xiao, J.-C. Lin, C. P. Chen, W. Liu, and Y.-H. Tong, "3-D deconvolutional networks for the unsupervised representation learning of human motions," *IEEE Trans. Cybern.*, vol. 52, no. 1, pp. 398–410, 2020.

[12] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Appl. Sci.*, vol. 7, no. 1, p. 110, 2017.

[13] N. Tasnim, M. K. Islam, and J.-H. Baek, "Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints," *Appl. Sci.*, vol. 11, no. 6, p. 2675, 2021.

[14] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Res Comput. Vision*, vol. 2, pp. 1–7, 2012.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," *2011 Int. Conf. Comput. Vision*, pp. 2556–2563, 2011.

[16] P. Pareek, and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artif. Intell. Rev.*, vol. 54, pp. 2259–2322, 2021.

[17] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egypt Inf. J.*, vol. 16, no. 2, pp. 187–198, 2015.

[18] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[19] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Rob. Auton. Syst.*, vol. 77, pp. 25–38, 2016.

[20] J. P. Gupta, N. Singh, P. Dixit, V. B. Semwal, and S. R. Dubey, "Human activity recognition using gait pattern," *Int. J. Comput. Vision Image Process. (IJCVIP)*, vol. 3, no. 3, pp. 31–53, 2013.

[21] M. Hazgui, H. Ghazouani, and W. Barhoumi, "Genetic programming-based fusion of HOG and LBP features for fully automated texture classification," *Vis. Comput.*, vol. 37, no. 1, pp. 1–20, 2022.

[22] Z. Li, Z. Zheng, F. Lin, H. Leung, and Q. Li, "Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN," *Multimedia Tools Appl.*, vol. 78, pp. 19587–19601, 2019.

[23] R. Mattivi and L. Shao, "Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor," *Comput. Analysis Images Patterns: 13th International Conference, CAIP 2009*, vol. 5702, pp. 740–747, 2009.

[24] K. B. Low and U. U. Sheikh, "Gait recognition using local ternary pattern (LTP)," *2013 IEEE Int. Conf. Signal Image Process. Appl.*, pp. 167–171, 2013.

[25] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," *2015 IEEE Winter Conf. Appl. Comput. Vision*, pp. 1092–1099, 2015.

[26] M. Rahman Ahad, M. N. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *J. Multimodal User Interfaces*, vol. 10, pp. 335–344, 2016.

[27] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," *2009 IEEE 12th Int. Conf. Comput. Vision*, pp. 32–39, 2009.

[28] J. M. Carmona and J. Climent, "Human action recognition by means of subtensor projections and dense trajectories," *Pattern Recognit.*, vol. 81, pp. 443–455, 2018.

[29] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, pp. 60–79, 2013.

[30] N. Neggaz and D. S. AbdElminaam, "Automatic sport video mining using a novel fusion of handcrafted descriptors," *2021 Int. Mobile, Intelligent, Ubiquitous Comput. Conf. (MIUCC)*, pp. 387–394, 2021.

[31] A. Kushwaha, A. Khare, and M. Khare, "Human activity recognition algorithm in video sequences based on integration of magnitude and orientation information of optical flow," *Int. J. Image Graph.*, vol. 22, p. 2250009, 2022.

[32] S. Yu, Y. Cheng, S. Su, G. Cai, and S. Li, "Stratified pooling based deep convolutional neural networks for human action recognition," *Multimedia Tools Appl.*, vol. 76, pp. 13367–13382, 2017.

[33] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, 2017.

[34] C. Geng and J. Song, "Human action recognition based on convolutional neural networks with a convolutional auto-encoder," *2015 5th Int. Conf. Comput. Sci. Autom. Engineering (ICCSAE 2015)*, pp. 933–938, 2016.

[35] H. Basak, R. Kundu, P. K. Singh, M. F. Ijaz, M. Woźniak, and R. Sarkar, "A union of deep learning and swarm-based optimization for 3D human action recognition," *Sci. Rep*, vol. 12, no. 1, p. 5494, 2022.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," *Proceedings IEEE conference Comput. Vision Pattern Recogn.*, pp. 6450–6459, 2018.

[37] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

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

[39] R. Huang, C. Chen, R. Cheng, Y. Zhang, and J. Zhu, "Human action recognition based on three-stream network with frame sequence features," *2022 7th Int. Conf. Image, Vision Comput. (ICIVC)*, vol. 1, pp. 37–44, 2022.

[40] H. Luo, G. Lin, Y. Yao, Z. Tang, Q. Wu, and X. Hua, "Dense semantics-assisted networks for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, pp. 3073–3084, 2021.

[41] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[42] F. Wang, G. Wang, Y. Huang, and H. Chu, "SAST: learning semantic action-aware spatial-temporal features for efficient action recognition," *IEEE Access*, vol. 7, pp. 164876–164886, 2019.

[43] L. Xia and X. Wen, "Multi-stream network with key frame sampling for human action recognition," *J. Supercomput.*, vol. 80, pp. 1–31, 2024.

[44] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, "Advancing human action recognition: a hybrid approach using attention-based LSTM and 3D CNN," *Sci. Afr.*, vol. 21, p. e01796, 2023.

[45] B. Chen, F. Meng, H. Tang, and G. Tong, "Two-level attention module based on spurious-3d residual networks for human action recognition," *Sensors*, vol. 23, p. 1707, 2023.

[46] S. M. Hejazi and C. Abhayaratne, "Handcrafted localized phase features for human action recognition," *Image Vis. Comput.*, vol. 123, p. 104465, 2022.

[47] S. Pal and C. Abhayaratne, "Phase feature-based activity level estimation for assisted living," *2nd IET Int. Conf. Technol. for Act. Assist. Living*, pp. 1–6, 2016.

[48] J. Ren, H. Zhao, J. Ren, and S. Cheng, "Sub-pixel motion estimation using phase correlation: comparisons and evaluations," *Int. J. Intell. Comput. Cybern.*, vol. 9, no. 4, pp. 394–405, 2016.

[49] A. Tran, J. Guan, T. Pilantanakitti, and P. Cohen, "Action recognition in the frequency domain," *ArXiv preprint arXiv:1409.0908*, 2014.

[50] S. Kumari and S. K. Mitra, "Human action recognition using DFT," *2011 Third Natl. Conf. Comput. Vis. Pattern Recognition, Image Process. Graph.*, pp. 239–242, 2011.

[51] Z. Lei, T. Ahonen, M. Pietikäinen, and S. Z. Li, "Local frequency descriptor for low-resolution face recognition," *2011 IEEE Int. Conf. Autom. Face Gesture Recogn. (FG)*, pp. 161–166, 2011.

[52] V. Ojansivu, E. Rahtu, and J. Heikkila, "Rotation invariant local phase quantization for blur insensitive texture analysis," *2008 19th Int. Conf. Pattern Recogn.*, pp. 1–4, 2008.

[53] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," *2008 19th Int. Conf. Pattern Recogn.*, pp. 1–4, 2008.

[54] A. Briassouli, "Unknown crowd event detection from phase-based statistics," *2018 15th IEEE Int. Conf. Adv Video Signal Based Surveill (AVSS)*, pp. 1–6, 2018.

[55] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 188–200, 2002.

[56] H. Feng, S. Wang, H. Xu, and S. S. Ge, "Object activity scene description, construction, and recognition," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 5082–5092, 2019.

[57] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," *Image Signal Process.: 3rd International Conference*, vol. 5099, pp. 236–243, 2008.

[58] S. Lakshmanan, P. Velliyan, A. Attia, and N. E. Chalabi, "Finger knuckle pattern person authentication system based on monogenic and LPQ features," *Pattern Analysis Appl.*, vol. 25, no. 2, pp. 395–407, 2022.

[59] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, 2002.

[60] Y. Ahuja, "Multiclass classification and support vector machine," *Glob J. Comput. Sci. Technol.*, vol. 12, no. G11, pp. 15–19, 2012.

[61] M. Gonen, A. G. Tanugur, and E. Alpaydin, "Multiclass posterior probability support vector machines," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 130–139, 2008.

[62] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mob. Netw. Appl.*, vol. 25, pp. 743–755, 2020.

[63] K. Mohan, P. Chandrasekhar, and S. A. K. Jilani, "A combined HOG-LPQ with Fuz-SVM classifier for object face liveness detection," *2017 Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud)(I-SMAC)*, pp. 531–537, 2017.