International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Assessing the Performance of Handcrafted Features for Human action Recognition

Aditi Jahagirdar*[1], Dhanashri Wategaonkar[2], Rashmi Phalnikar[3]
[1, 2, 3], Department of Computer Engineering and Technology,
Dr. Vishwanath Karad MIT World Peace University,
Pune, India
[1] *aditi.jahagirdar@mitwpu.edu.in*
[2] *dhanashri.wategaonkar@mitwpu.edu.in*
[3] *rashmi.phalnikar@mitwpu.edu.in*

Abstract— Recognition of Human action such as running, punching, bending, kicking etc. plays an vital role in futuristic applications like intelligent video surveillance, health care monitoring, robotics, smart automation system, computer gaming etc. This field relies on various approaches based on hand crafted features like PCA, HOG, LBPH, DWT, STIP, SWF, SWFHOG and deep learning techniques like CNN, RNN and their variants. Though many approaches have been proposed and implemented by researchers, the literature survey suggests that a detailed understanding of the approaches and a comparison of advantages and limitations is required to develop more accurate action recognition method. This paper focuses on this issue and gives detailed analysis of results obtained by implementing algorithms on standardize open source datasets of varying complexity namely Weizmann, KTH, UT Interaction and UCF sports. The results are compared based on the classification accuracy as it is one of the performance measure for checking reliability of the method. The comparison shows that, SHFHOG feature gives the best classification accuracy as compared to other handcrafted features and also outperforms the simple CNN.

Keywords- Human Action Recognition, Hand crafted features, PCA, LBPH, HOG, SWFHOG, CNN, Deep Learning

I. Introduction

In recent years, human action recognition (HAR) has received attention from the researchers. Role of video action recognition is vital in applications like intelligent video surveillance, smart automation system, robotics, healthcare monitoring etc. The development in machine learning and computer vision domains has motivated the researchers to develop various techniques for automatic HAR. Since there are many machine learning techniques, it is important to identify the most appropriate method for best results in given scenario [1]. One of the challenges is that one action can be performed differently by people, owing to it high intra-class variations is present in the human actions. Further, similar actions like jogging, walking and running produce high inter class similarity. These two parameters make human action recognition a challenging task. Moreover, occlusion, camera view angle changes, changes in the scale, illumination changes etc. add to the difficulty level. The approach to be used for human action recognition depends on the complexity of action to be recognized and the application in which it is used.

The human actions are categorized into gestures, simple actions, interactions and group activities. Gestures are small movements of body parts done with some specific purpose, e.g. palm movements in sign language, nodding of head. Simple actions are intended body movements of random complexity that can be sequence of gestures performed periodically. Interaction can be between two humans or between human and object. e.g., a person carrying a bag, two persons shaking hands. A group activity is defined as several humans involved in common objective e.g. group conference, group dance. Complexity of understanding the gestures is lowest and group activity is the highest.

The approaches to this problem of HAR use handcrafted features, deep learning methods or combination of both. General block schematic of the HAR technique using handcrafted features is shown in Fig. 1.

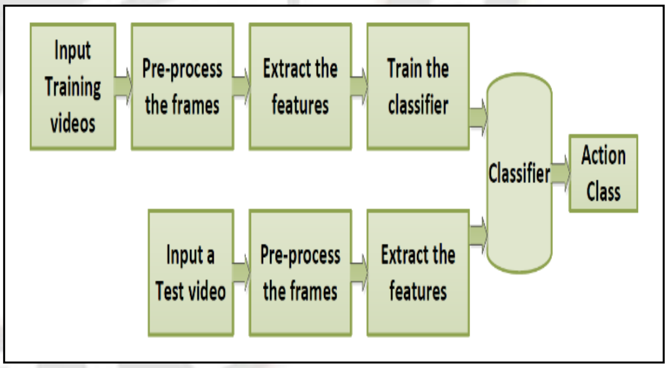

Figure 1. Block schematic of HAR using handcrafted features.

Use of computer vision has made HAR an important step in many applications [2]. On the basis of the literature survey, the techniques for HAR, differ from each other mainly based on type of feature used for representing an action. Local as well as global handcrafted features have been explored by the researchers and both approaches are seen to have advantages and disadvantages.

Local features look at the image as group of patches and extracts information from them to represent the action. This makes it robust against background changes and occlusion. The main disadvantage of local features is, it sometimes fails to look at the sequence as a whole and does not capture adequate spatial and temporal information. Global features on the other hand, look at the video as a whole and represent the action using whole

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

human body information. Global features are easy to capture but might not give good results in case of cluttered background and occlusion. In most of the applications, Global features are used for object detection, while the local features are used for object recognition. A comprehensive survey of human activity recognition techniques is given by L. Minh Dang et al [3]. In most cases, the performance of a HAR system is improved by blending a global and local feature. Invariant Moments like Hu, Zernike, Shape Matrices, Histogram Oriented Gradients (HOG) and Co-HOG are some examples of global features. There are several local features that are used for determining the action such as the LBP, SIFT, SURF, FREAK, MSER, and BRISK.

The most popular spatio temporal detectors are based on Harris 3D detector. An interest point detection method, based on geometric algebra [4], which can capture the appearance and motion information from the video uniformly is proposed. Nazir et al.[5] proposed a modification of popular bag of visual words called Dynamic Spatio-Temporal Bag of Expressions (D-STBoE). 3D Harris detector is implemented to extract the local features. The bag of visual words is then implemented using K-means clustering algorithm. [6] have proposed a spatio temporal local feature based on RGB as well as depth channel, called RGBD cuboids. Spatio temporal interest points are first extracted from depth data and then mapped to RGB data. 3D spatio temporal cuboids are formed around depth interest points. Cuboid size inversely proportional to depth is selected to tackle the problem of scale of the object. Nazir et al. [7] have proposed a method based on spatio temporal interest point detection and scale invariant feature transform representation. For extracting sparse STIPs, 3D Harris detector is implemented. The extracted STIPs are then represented using 3D SIFT.

Use of ensemble of multiple descriptors is explored by [8] [9]. LBP, HOG, Haar wavelets, SIFT, velocity and displacement are the features extracted after applying the background subtraction. Sequential minimal optimization technique is implemented for getting better results.

Al-Berry et al. [10] have used a combination of local and global features to construct a feature descriptor to take advantage of both the techniques. Discrete wavelet transforms have been used successfully by other researchers to detect motion. Few researchers [11][12][13] have proposed use of detail coefficients obtained by applying DWT as features. The method takes advantage of multi resolution property of DWT by finding the features from level 1 to level 7. The method gives promising results in terms of recognition accuracy but fails to give high values of precision and recall. Yang Li et al [14] proposed sensor based human activity recognition. Based on the data obtained from the sensors a matrix is formed to predict the activity. It out performs sensor based methods in terms of daily behavior recognition accuracy and time consumption. C. In [15], high level features are extracted and used to classify the actions using HMM. The paper focuses on non-machine learning approach for human action recognition and emphasizes selection of appropriate feature to represent an action.

The area of video-based HAR has undergone a revolution owing to development of deep learning techniques. Deep learning techniques have shown promising results with high accuracy and robustness, however they have several drawbacks. These algorithms have limited capacity in handling complex temporal information and also it requires huge number of data samples for training the model. As number of actions are many getting the large number of samples for each action becomes difficult. Owing to this, using handcrafted features becomes good option for video based HAR. This study focuses on comparison of performance of various simple handcrafted features for publically available well known datasets.

Various deep learning based approaches have been explored for HAR in the recent years. In. [16], authors have proposed use of RNN for recognizing the actions from depth videos. It uses stateful and stateless convolutional LSTM networks for recognizing the action. Around 80% accuracy is achieved with these methods on RGB + Depth dataset. Use of LSTM with a light weight feature is proposed in [17]. This method makes it possible to detect the people and recognize the action in real time. High accuracy is obtained with this method for datasets used. A detail review of various deep learning based methods is given in [18] by Hieu H. Pham et.al.

An inclusive survey of CNN based methods is given by Yang Jiaxin [19]. Zhang [20] proposed integration of the discrete wavelet transform (DWT) technique with the DT model for gaining more descriptive human action features. It further combines features obtained from the pre trained CNN-RNN model with handcrafted features to form a Fisher vectors. The system has shown promising results for benchmark datasets. A two stream CNN is proposed in [21]. To increase recognition accuracy, two-stream CNNs incorporate spatial and temporal data from RGB frames and optical flow, respectively. A 3D Convolutional Network is implemented in [22]. It uses 3D motion cuboids for representing the actions. The results show potential of the 3D CNN technique.

After studying the literature it is seen that deep learning methods face challenges like managing large variations in appearance and motion, handling the occlusions, less number of samples in a dataset and class imbalance in action datasets. All these issues can be tackled by using handcrafted features or their combinations. Use of various handcrafted features can help in improving the performance of HAR system at low computational complexity.

This work focuses on emphasizing use and importance of various simple handcrafted features by implementing them. Evaluation of these features is done on publically available datasets. The results prove the fact that handcrafted features play an important role in HAR system. The detail result analysis of each feature tries to give in-site of why each feature is giving results in a certain way. This will help the researchers to select the features based on type of data and application.

II. Methodology

For assessing the performance of hand crafted features, various simple hand crafted features are extracted from the action videos. The features and their combinations are implemented for finding the classification accuracy. CNN is also implemented for comparing the accuracy with deep learning method. This sections gives brief description of all the feature techniques implemented in this work.

A. Principal Component Analysis

Principal Component analysis is a second order statistical measure to represent data with maximum possible variance values and lower dimensions. Principal Components (PCs) give global information. Principal component analysis (PCA) is a technique from multivariate analysis and linear algebra. It uses statistical approach for representing a large and highly correlated data in smaller number of samples. The first and most common

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

application of PCA is in dimensionality reduction when dealing with large datasets. Other applications where PCA is used are data compression and de-noising of data. In general, PCA transforms the data in the vector space with less number of variables, called principal components having maximum variance. This helps in preserving most of the information present in the data and makes it possible to identify the patterns in the data. This property of PCA we have used here to describe the action. Eigenvectors and Eigenvalues are computed for the covariance matrix of zero mean data. Eigenvectors having higher eigenvalues are selected for representing the data. A feature vector is formed by converting diagonal of eigenvalues to row matrix. A matrix of eigenvalues computed for each frame, represent a video. Choice of number of frames used to represent one video is not defined and is a tradeoff between computational cost and accuracy of recognition. Steps to find the Eigenvalues using PCA are given as:

1. Input data

Data = matrix of dimension [m n]

Each frame of the video is considered as a matrix of m rows and n columns and passed as an input to PCA algorithm.

2. Find mean of each column

Arithmetic mean of each column is found using formula in Eq. (1)

$$\bar{x} = \frac{\sum_{k=1}^{m} x_k}{m} \qquad (1)$$

3. Subtract the mean from each column

Arithmetic mean computed is subtracted from each column to get zero centered data as shown in Eq. (2).

$$x_{zk} = x_k - \bar{x} \qquad (2)$$

4. Find Co-variance matrix

Covariance between two variables gives dependence of one variable on other. Covariance between two variables is computed using Eq. (3)

$$Cov_{xy} = \frac{\sum_{k=1}^{m}(x_k-\bar{x})(y_k-\bar{y})}{m-1} \qquad (3)$$

As in a frame there are multiple columns, covariance is computed between every two rows, generating a covariance matrix. The matrix obtained is shown in Eq. (4).

$$Cov_{matrix} = \left\{ \begin{matrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{matrix} \right\} \qquad (4)$$

5. Find Eigen vectors and Eigen values of the covariance matrix

Eigen decomposition method is applied to the covariance matrix to identify the dimension with the largest variance. Then it identifies dimension having second largest variance from remaining variances and so on. These variances computed are called as Eigenvalues. PCA replaces the original variables with the new transformed variables called principal components. The first principal component is the component with highest Eigen value or variance.

6. Forming a feature vector

The matrix of Eigen values is a diagonal matrix as the elements of the diagonal represent variance of the parameters while remaining components are covariance between various parameters. The diagonal elements represent the eigenvalues. The diagonal elements are arranged in descending order and then extracted as a feature.

7. Selection of features

As the extracted features are in descending order, first few values having significant variance are selected as features. This preserves most of the information of each frame and at the same time reduces the size of the feature vector.

B. Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) was first introduced by Dalal and Triggs. The HOG feature was initially introduced for detecting the human figure from a still image. HOG is the histogram of edge orientations of the gradients found in localized regions of an image. It gives the rough shape information of the object present in an image. In case of edge feature, only the pixels on the edges are identified. In HOG computation, magnitude as well as direction of the edge pixels is identified as gradient and orientation. The image is first divided in number of small regions called cells. The magnitude and directions are then computed for all the pixels within these cells. Histograms of gradient directions, based on the magnitude, are then computed for all these cells separately. Concatenation of all these histograms is done to generate the HOG feature descriptor. Block normalization is applied over the bigger regions called blocks for increasing the accuracy. For applying block normalization, intensity of pixels is measured across the blocks and using this value all the histograms of cells within that block are normalized. This method results in contrast normalizing and thus better representation of shape of the object at the same time making the HOG robust to variations in illumination. As HOG is computed over small local regions, it is robust to geometric variations which become crucial for larger image regions.

In this work, HOG technique is applied to each frame of the input video and features are extracted. The features computed for all frames are then stored in a matrix form and used to represent the video. The process for computing the HOG features is given in detail in following steps.

1. Gradient computation

Gradient Fx in horizontal direction and Fy in vertical direction is computed using the derivative masks. D_x is the derivative mask applied in horizontal direction and D_y is the derivative mask applied in vertical direction as given in Eq. (5).

$$D_x = [-1, 0, 1], \quad D_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \qquad (5)$$

The magnitude and orientation of the gradient is computed as given in Eq. 6).

$$|G| = \sqrt{{F_x}^2 + {F_y}^2} \text{ and } \qquad \theta = arctan\frac{F_y}{F_x} \qquad (6)$$

2. Orientation binning (creating the cell histograms)

In this work, unsigned gradients are used which range from 0^0 to 180^0. Nine gradient bins, each of 20^0, are formed for computing the histograms. Gradients having high magnitude have high influence on the histogram. In this work as cells of 8X8 size is selected, 64 bit gradient vector is generated per cell distributed in 9 bins.

3. Histogram Normalization

To make the HOG feature descriptor robust to the variations in the illumination, normalization is applied. The normalization process stretches the signal values over the possible range by using maximum range. The HOG descriptor is then computed by concatenating normalized cell histograms.

4. *Descriptor blocks and Block normalization*

Block normalization is applied to cells for stretching the regions of low contrast present in the image. Multiple spatially connected blocks are formed by grouping cells together. In this work, L2 normalization is used for block normalization. The equation used for block normalization is given in (7)

$$L2_{norm} = \frac{v}{\sqrt{||v||^2+e^2}} \qquad (7)$$

C. *Local Binary Pattern Histogram*

Texture is one of the important feature used to discriminate between two images. Local Binary Pattern (LBP) is a simple but very efficient texture operator used in object recognition tasks.

The LBP considers a small neighborhood of a pixel and then compares the current pixel under consideration with each neighboring pixel. If the value of center pixel is greater than neighboring pixel then it writes one otherwise zero. Thus each pixel is represented by sequence of 1s and 0s. For a neighborhood of 3 x 3 pixels, sequence of 8 bits will be generated for each pixel. Thus all the pixels of the image are represented in binary pattern. The image thus generated is divided in multiple grids of same size. Histogram is then calculated over each grid. As gray scale values are from 0 to 255, each grid generates 256 bins. The feature descriptor is computed by concatenating the histograms. In this work, circular neighborhood with 8 neighbors is implemented. So each pixel is represented as a binary number of 8 bits. The transformed image is then divided into multiple grids of same size. Histogram is then computed for each grid depending on the value of the pixel in transformed image.

In this work, circular neighbourhood with 8 neighbours is implemented. So each pixel is represented as a binary number of 8 bits. The equation to generate this N bit binary number and then represent it in its decimal equivalent can be given in generalized form as in Eq. (8)

$$LBP_n = \sum_{n=1}^{N} f(x) * 2^n$$

$$f(x) = f(g_n - g_c) = \begin{cases} 1, & if\ x \geq 0 \\ 0, otherwise \end{cases} \qquad (8)$$

D. *Spatio Temporal Interest Points*

Local feature based methods are the most favored for human action recognition. The main advantage of the local feature based methods is that they do not require localizing the human figure or modeling the human body. The STIPs are extracted directly from the video without foreground extraction. This removes the hard task of background modeling. The most explored interest point detectors are Harris detector, Laplacian – DoG, Harris - /Hessian - Laplace, Harris - /Hessian - Affine detector [23]. As compared to global features, STIPs are robust to variations in illumination and geometric transformation. The method proposed by Harris and Stephen is used here for extracting the interest points. It is widely explored technique because of its strong invariance to scale, rotation and illumination. The Harris corner detector uses autocorrelation function which calculates the local changes in small windows shifted by small amount. The implementation steps for computing Harris corners are given here

1. Calculate x and y derivatives of image as given
   $dI_x = G_{\sigma_d,x} * I$, $dI_y = G_{\sigma_d,y} * I$
2. At every pixel, calculate the product of derivatives as given
   $dI_x^2 = dI_x * dI_x$, $dI_y^2 = dI_y * dI_y$, $dI_{xy} = dI_x * dI_y$
3. Compute the weighted sums of products of derivatives
   $S_x^2 = G_{\sigma_i} * dI_x^2$, $S_y^2 = G_{\sigma_i} * dI_y^2$, $S_{xy} = G_{\sigma_i} * dI_{xy}$
4. Compute the autocorrelation at each pixel as
   $M(x,y) = \begin{bmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{bmatrix}$
5. Compute the response at each pixel as
   $R_H = \det(M) - k.(trace(M))^2$
6. Set a threshold value empirically and apply non maximal suppression to extract required interest points.

E. *Discrete Wavelet Transform*

Holistic motion patterns can be extracted using frequency domain approaches. Complex images consist of varying levels of details. To capture these details multi-resolution analysis plays an important role. 2 D wavelet transform is most common yet very efficient tool for moving object detection technique. Detail coefficients obtained by wavelet decomposition represent high-frequency components because of which they are able to capture significant edge information. In this work, detail coefficients are used for representing the action. While there are many types of wavelets, Daubechies wavelets (db) are most commonly used, since they provide slightly longer support. The db1 wavelet or Harr wavelet is used in this work. The sub-bands generated by wavelet decomposition are used as local features. Steps to find wavelet coefficients are given as:

1. Obtain low pass and high pass decomposition filter coefficients.
2. Convolve input image matrix row-wise with low pass decomposition filter coefficients obtained in step 1.
3. Down-sample to keep only even indexed elements to get intermediate matrix z.
4. Convolve matrix z obtained in step 3 column-wise with low pass and high pass decomposition filter coefficients separately to obtain the average and detail horizontal coefficients.
5. Convolve input image row-wise with high pass decomposition filter coefficients obtained in step 1.
6. Down-sample to keep only even indexed elements to get intermediate matrix z.
7. Convolve matrix z obtained in step 3 column-wise with low pass and high pass decomposition filter coefficients separately to obtain detail vertical and detail diagonal coefficients.
8. Concatenate the detail coefficients to obtain the feature vector.

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

F. Salient Wavelet Features-Histogram of Oriented Gradients

The Salient Wavelet Feature (SWF) and its extension, the Salient Wavelet Feature Histogram of Oriented Gradients (SWFHOG), are local features that are investigated. In the SWF technique, local spatio - temporal interest point detection is done for capturing the movement in small local areas of the frame. Salient regions are chosen from prominent local areas depending on the information they contain. DWT is applied to 3D volume of salient regions for detecting motion patterns. Steps to extract SWF feature are given in figure 2 [24].

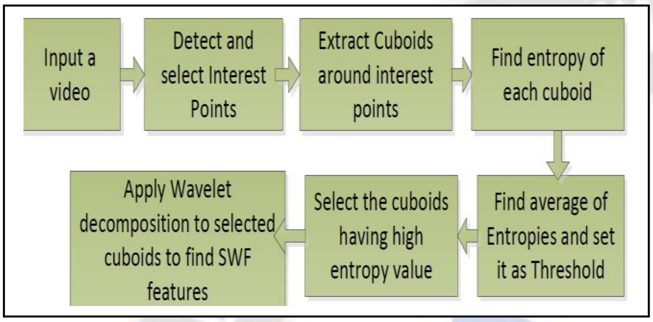
Input a video
Detect and select Interest Points
Extract Cuboids around interest points
Find entropy of each cuboid
Find average of Entropies and set it as Threshold
Select the cuboids having high entropy value
Apply Wavelet decomposition to selected cuboids to find SWF features

Figure 2. Block schematic of SWF feature technique.

SWHHOG is an extension of SWF technique. The Histogram of Oriented Gradient (HOG) and the SWF feature are merged to enhance the SWF feature's performance, creating the SWFHOG feature descriptor. SWF gives local and HOG gives global information about the action. The SWF and HOG features are extracted separately for the frames and then concatenated to form final feature descriptor. The steps implemented to extract the SWFHOG are given as

1. Detection and selection of Interest Points
2. Formation of Cuboids
3. Selection of Cuboids
4. Apply Wavelet decomposition to selected cuboids to find SWF
5. Apply HOG technique to obtain global feature
6. Select significant SWF and HOG features by applying PCA to both separately.
7. Concatenate selected SWF and HOG features to form SWFHOG feature descriptor

G. Convolutional neural network

To compare the performance of handcrafted features with deep learning techniques, Convolutional neural network is implemented. A variety of deep learning models, including CNN, 3D CNN, LSTM, recurrent neural network (RNN), region-based convolutional neural network (RCNN), faster RCNN, etc., are available for feature extraction and action recognition [25][26]. CNN is the most explored algorithm for image classification. Researchers have suggested a variety of CNN modifications for image categorization. The CNN can be defined as an extension of a neural network, but it operates on volume instead of a vector. Dense CNNs are preferred for image classification applications, since detailed features are required for this task. Multiple convolution and pooling layers are used in dense CNNs. As the number of layers go on increasing, computational complexity of the algorithm increases.

III. EXPERIMENTATION

The standardized open source datasets used for performance evaluation have various challenges like cluttered background, illumination change, occlusion, change in camera view angle, scale variation, imperfect actions and varied objects present along with the human. A detailed analysis of the datasets used is given here.

The Weizmann dataset has 90 video sequences with nine different actors performing ten simple actions. The video is captured at a frame rate of 25 frames per second. A static camera is used to record the videos in a controlled environment. Running, strolling, jumping forward on two legs, galloping sideways, skipping, bending, jumping jack, jumping in one spot, waving with one hand, and waving with both hands are among the actions included in the dataset.

In comparison to the Weizmann dataset, the KTH dataset is more complex. The six simple behaviors in the dataset are "walking," "running," "jogging," "boxing," "waving a hand," and "clapping." There are 25 different actors who participated in the actions. The videos are recorded with 25fps. The actions are recorded in four different scenarios, shown as S1, S2, S3 and S4.

In comparison to the Weizmann and KTH datasets, the UT Interaction dataset is complex. It is recorded in realistic environment. It is having six different human-human interactions, namely, “handshaking”, “punching”, “pushing”’ “kicking”, “hugging” and “pointing a finger”. The videos are recorded with speed of 30fps. Both the sets contain 10 videos of six actions but with different challenges. For the objective of identifying human action, the Weizmann, KTH, and UT interaction datasets were specifically created.

The UCF sports dataset is the collection of various sport videos which are normally telecasted on sport channels like ESPN or BBC. The dataset has total 150 video sequences spanning over 10 action. "Diving," "golf swing," "kicking a ball," "weight lifting," "horseback riding," "running," "skateboarding," "swing bench," and "walking" are a few of the available activities.

The RGBD Microsoft Kinect Sensor is used to build the SBU Kinect Two Person Interaction dataset. It has color images as well as depth maps with ground truths. It comprises of seven actors performing in 21 pairs eight different events, including approaching, leaving, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. The identical lab setting serves as the backdrop for all videos. The dataset has total of 300 videos of interactions. Specifications of the datasets used are given in table 1.

For evaluating the techniques mentioned in section 3, feed forward neural network is implemented. For the fair evaluation, ratio of samples used for the training, testing and validation is kept 80%, 10% and 10% respectively for all the experiments. Stratified random sampling is used for avoiding influence of any one class. Performance of the implemented algorithms is analyzed based on the classification accuracy achieved. Experimentation is done by using single feature as well as combination of features for understanding effect of each feature. Six-layered CNN architecture having one input layer, two convolution layers, two pooling layers and one fully connected output layer is utilized for evaluating CNN technique. Three filters of size 5 x 5 are used in both convolution layers. The CNN is trained for 50 epochs.

TABLE I. SPECIFICATIONS OF DATASETS

Categories	**Datasets**					
	Weizmann	***KTH***	***UT Interaction 1***	***UT Interaction 2***	***SBU Kinect***	***UCF Sports***
Type of action	Simple	Simple	Interaction	Interaction	Interaction	Complex
No. of Classes	10	6	6	6	8	10
No. of videos	90	600	60	60	300	150
No. of actors	9 actors	25 actors	15 pairs	15 pairs	21 pairs	Real sports videos
Environment	Controlled	Controlled	Realistic background	Realistic, windy background	Controlled	Realistic background
Frame Rate	25 fps	25 fps	30 fps	30 fps	10 fps	10 fps
Resolution	180 x 144	160 × 120	720 x 480	720 x 480	720 x 480	720 x 480
Complexity	Simple	Marginally complex	Complex	Complexity more than in set 1	Complex	Most Complex
Other remarks	Background is almost constant	4 scenarios: Indoor, Outdoor, Scale change, Different clothes	Some camera jitter, Clothes of various colours	More camera jitter, cluttered background, more than two people present in the frame	Background is almost constant	High intra class variation. Scale, illumination variations cluttered background

IV. RESULTS AND DISCUSSION

The techniques mentioned in section 3 are vigorously tested on action recognition datasets. Classification accuracy is computed and plotted for each dataset. Fig. 3 to Fig. 7 show graphs of classification accuracy obtained on Weizmann, KTH, UCF Sports, UT Interaction 1, and UT Interaction 2 dataset respectively. The complexity level of the actions goes on increasing as we go from Weizmann dataset to UT Interaction dataset. Challenges like, low resolution, occlusion, cluttered background, interclass similarity are present in the datasets used for testing. It is seen that, for all these datasets, highest classification accuracy is obtained with SWFHOG technique. For Weizmann and KTH datasets, higher values of classification accuracies are obtained for PCA, HOG and PCA+HOG as background subtraction technique is implemented before extracting these features. Foe Weizmann and KTH datasets, constant background is available which makes it possible to extract foreground satisfactorily. For UT Interaction and UCF sports datasets, as constant background is not available, extracting exact foreground is not possible. This results in less classification accuracy using PCA, HOG and PCA+HOG features.

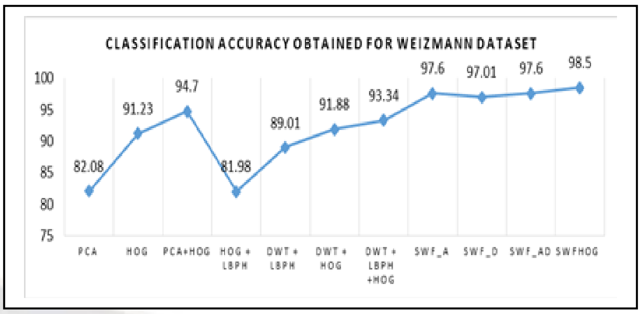

Figure 3. Classification accuracy - Weizmann Dataset.

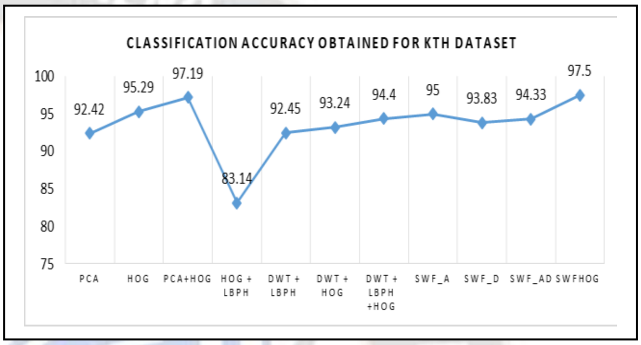

Figure 4. Classification accuracy - KTH Dataset.

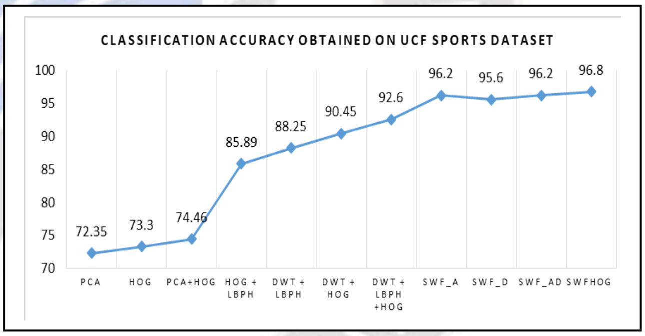

Figure 5. Classification accuracy – UCF Sports dataset

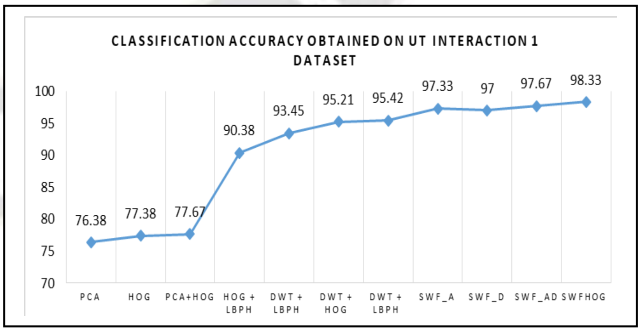

Figure 6. Classification accuracy – UT Interaction 1 dataset

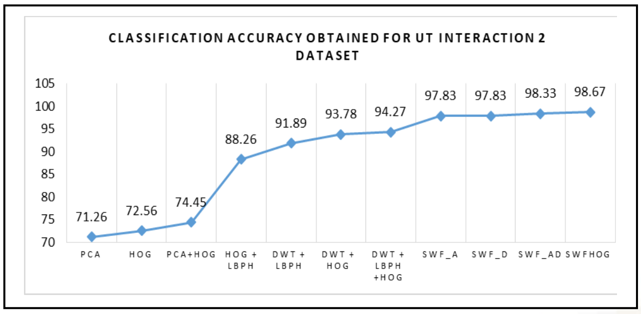

Figure 7. Classification accuracy – UT 2 Interaction dataset

Table 2 gives classification accuracy obtained with all the handcrafted features implemented.

TABLE II. COMPARISON OF RESULTS

	% Recognition Accuracy				
Techniques	***Weizmann***	***KTH***	***UCF Sports***	***UT1***	***UT2***
PCA	82.08	92.42	72.35	76.38	71.26
HOG	91.23	95.29	73.3	77.38	72.56
PCA+HOG	94.7	97.19	74.46	77.67	74.45
HOG + LBPH	81.98	83.14	85.89	90.38	88.26
DWT + LBPH	89.01	92.45	88.25	93.45	91.89
DWT + HOG	91.88	93.24	90.45	95.21	93.78
DWT + LBPH	93.34	94.4	92.6	95.42	94.27
SWF_A	97.6	95	96.2	97.33	97.83
SWF_D	97.01	93.83	95.6	97	97.83
SWF_AD	97.6	94.33	96.2	97.67	98.33
SWFHOG	98.5	97.5	96.8	98.33	98.67

From the results obtained, it is seen that, significant improvement is obtained in recognition accuracy when SWFHOG technique is used. In the experimentation where PCA, HOG and PCA+HOG features are used to describe the actions, foreground extraction is implemented.

In PCA technique, Eigen values are used as a feature. It is seen that, when PCA is used as a feature less recognition accuracy is achieved as compared to that when HOG and PCA+HOG are used as features. As PCA considers only the statistical characteristics of the frames, the feature values are similar for frames having very less change. This happens especially when movement is less. For HOG features, the accuracy increases as compared to that with PCA as it captures the direction and magnitude of the movement. HOG is able to capture the shape of the object as movement is more significant at the edges of the object. PCA+HOG feature improves the accuracy further.

It is seen that, satisfactory performance is achieved for the Weizmann and KTH datasets as compared that of UCF Sports and UT Interaction dataset. As Weizmann and KTH datasets are recorded in controlled environment with constant background, foreground object is detected perfectly. In UCF Sports and UT Interaction datasets, where background is not constant and multiple objects are present in the frame, foreground objects are not detected satisfactorily, hampering the accuracy results. From this experimentation it is concluded that, PCA is not able to represent an action satisfactorily.

HOG, LBPH and DWT methods are further explored by applying them to frame as a whole. It is seen that, performance of HOG+LBPH is less as compared to performance of HOG+DWT and DWT+LBPH. Detail coefficients obtained by applying discrete wavelet transform are used as features to represent the action. LBPH technique gives the texture information whereas DWT gives edge information of the objects. It is seen that, maximum recognition accuracy is obtained when HOG, LBPH and DWT features are used together.

Highest accuracy of 95.42% is achieved for UT 1 Interaction dataset and lowest accuracy of 92.6% is achieved for UCF Sports dataset. It is seen that LBPH feature or texture information is not much useful in representing the action. In action videos, frames at the start of the action and at the end of the action have very less movement and tend to produce similar LBPH features. DWT and HOG features give promising results and are explored further.

Taking into consideration performance of all the implemented techniques, a new local feature called SWF is introduced. Salient region extraction using spatio temporal interest point detection, entropy based cuboid selection and extraction of wavelet coefficients from the cuboids are key points of the technique. The SWF technique is further enhanced by using it along with HOG technique.

The results show that, highest recognition accuracy of 98.67% is achieved for UT 2 Interaction dataset whereas 96.8% is achieved for UCF Sports dataset. The selection of cuboids helps in picking the parts of the videos that have maximum information whereas discrete wavelet transform describes the local movements of the human body satisfactorily. As STIPs are localized in nature, they make the SWFHOG features robust to occlusion, geometric changes and illumination changes to large extent. Use of Gabor filter in the process of interest point detection makes the feature rotation invariant. Use of HOG descriptor helps in giving shape and gradient of movement. All these properties of SWFHOG technique are reflected in the result where high recognition accuracy is achieved for datasets having various challenges.

As mentioned in section 3, CNN technique was implemented as a deep learning method and classification accuracy is computed. Number of epochs used in CNN play a crucial role in training of the model. It is also directly proportional to the amount of time required to train the model. For evaluating the performance of CNN profoundly, recognition accuracy is computed at different numbers of epochs. Results obtained with various number of epochs are given in the table 3. It is seen that, for most of the datasets, recognition accuracy increases when the number of epochs are increased from 5 to 100 but it remains almost constant when number of epochs are increased from 100 to 200. It happens due to overfitting taking place as number of epochs are increased beyond 100.

TABLE III. CLASSIFICATION ACCURACY FOR DIFFERENT NUMBER OF EPOCHS

	Datasets				
No. of Epochs	***Weizmann***	***KTH***	***UCF***	***UT 1***	***UT 2***
5	84	86	78	83	82
10	84	82	82	89	84
20	86	78	86	90	82
50	88	81	89	91	83
100	88	81	93	91	83
200	87	81	91	92	83

Table 4 shows comparison of results obtained with SWFHOG technique and CNN technique.

TABLE IV. COMPARISON OF ACCURACY OBTAINED WITH CNN AND SWFHOG FEATURE

	% Recognition Accuracy				
Techniques	***Weizmann***	***KTH***	***UCF Sports***	***UT1***	***UT2***
CNN	88.06	86.25	90.31	91.5	83.75
SWFHOG	98.5	97.5	96.8	98.33	98.67

It is seen that higher recognition accuracy is obtained when SWFHOG feature is used with simple feed forward neural network as compared to CNN. Higher recognition accuracy can be obtained with the CNN by increasing number of layers. Number of samples available also affect the performance of CNN.

V. CONCLUSION AND FUTURE SCOPE

In this work, various handcrafted features and their combinations are studied and implemented for human action recognition to emphasize their usefulness in the field. Work is carried out on varied datasets having different complexity levels and challenges. Simple well known features like PCA, LBPH, DWT coefficients, STIPs and their combinations are implemented. A feed forward neural network is used for classification of the action classes. Highest classification accuracy is obtained when SWFHOG feature is used for representing the action. It emphasizes the fact that a local feature along with a global feature can improve the action classification performance. As compared to deep learning technique, CNN, higher accuracy is obtained by the SWFHOG technique with less computational complexity. Better performance can be obtained by using various variants of CNN and RNN. Research in this domain can be taken ahead by exploring more hand crafted features and combining them with deep learning techniques. Recognizing different actions performed by various people simultaneously is a challenging task and can be explored further.

REFERENCES

[1] M. F. Aslan, A. Durdu, and K. Sabanci, “Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization,” Neural Comput. Appl., vol. 9, 2019, doi: 10.1007/s00521-019-04365-9.

[2] M. G. Morshed, T. Sultana, A. Alam, and Y. K. Lee, “Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities,” Sensors, vol. 23, no. 4, pp. 1–40, 2023, doi: 10.3390/s23042182.

[3] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, “Sensor-based and vision-based human activity recognition: A comprehensive survey,” Pattern Recognit., vol. 108, 2020, doi: 10.1016/j.patcog.2020.107561.

[4] Y. L. Q. L. Q. H. R. X. X. Li, “Spatiotemporal interest point detector exploiting appearance and motion-variation information,” J. Electron. Imaging, vol. 28, no. 3, pp. 1–14, 2019.

[5] S. Nazir, M. H. Yousaf, and S. A. Velastin, “Dynamic Spatio-Temporal Bag of Expressions (D-STBoE) Model for Human Action Recognition,” Sensors, vol. 19, no. 2790, pp. 1–21, 2019, doi: 10.3390/s19122790.

[6] H. A. Sial, M. H. Yousaf, and F. Hussain, “Spatio - Temporal RGBD Cuboids Feature for Human Activity Recognition,” Nucl., vol. 3, no. 3, pp. 139–149, 2018.

[7] S. Nazir, M. H. Yousaf, and S. A. Velastin, “Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition,” Comput. Electr. Eng., vol. 0, pp. 660–669, 2018, doi: 10.1016/j.compeleceng.2018.01.037.

[8] H. Naveed, G. Khan, A. U. Khan, A. Siddiqi, and M. U. G. Khan, “Human activity recognition using mixture of heterogeneous features and sequential minimal optimization,” Int. J. Mach. Learn. Cybern., vol. 10, no. 9, pp. 2329–2340, 2019, doi: 10.1007/s13042-018-0870-1.

[9] A. S. Jahagirdar and M. S. Nagmode, Silhouette-based human action recognition by embedding HOG and PCA features, vol. 673. 2018.

[10] M. Al-Berry, A.-M. Mohammed, H. Ebied, A. Hussein, and M. Tolba, “Weighted Directional 3D Stationary Wavelet-based Action Classification,” Egypt. Comput. Sci. J., vol. 39, no. 2, pp. 83–97, 2015.

[11] S. Yousefi, M. T. Manzuri Shalmani, J. Lin, and M. Staring, “A Novel Motion Detection Method Using 3D Discrete Wavelet Transform,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 12, pp. 3487–3500, 2019, doi: 10.1109/TCSVT.2018.2885211.

[12] S. Yousefi and J. Lin, “A Novel Motion Detection Method Resistant to Severe Illumination Changes,” arXiv Comput. Vis. Pattern Recognit., 2016, doi: 10.1109/TCSVT.2018.2885211.

[13] M. Khare and M. Jeon, “Towards discrete wavelet transform-based human activity recognition,” Second Int. Work. Pattern Recognit., vol. 10443, p. 1044308, 2017, doi: 10.1117/12.2280346.

[14] Y. Li, G. Yang, Z. Su, S. Li, and Y. Wang, “Human activity recognition based on multienvironment sensor data,” Inf. Fusion, vol. 91, no. February, pp. 47–63, 2023, doi: 10.1016/j.inffus.2022.10.015.

[15] Y. Hartmann, H. Liu, S. Lahrberg, and T. Schultz, “Interpretable High-level Features for Human Activity Recognition,” no. February, pp. 40–49, 2022, doi: 10.5220/0010840500003123.

[16] Sánchez-Caballero, Adrián, David Fuentes-Jiménez, and Cristina Losada-Gutiérrez. "Real-time human action recognition using raw depth video-based recurrent neural networks." Multimedia Tools and Applications 82, no. 11 (2023): 16213-16235.

[17] Cob-Parro, Antonio Carlos, Cristina Losada-Gutiérrez, Marta Marrón-Romera, Alfredo Gardel-Vicente, and Ignacio Bravo-Muñoz. "A new framework for deep learning video based Human Action Recognition on the edge." Expert Systems with Applications 238 (2024): 122220.

[18] Pham, Hieu H., Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin. "Video-based human action recognition using deep learning: a review." arXiv preprint arXiv:2208.03775 (2022).

[19] Y. Jiaxin, W. Fang, and Y. Jieru, "A review of action recognition based on Convolutional Neural Network," J. Phys. Conf. Ser., vol. 1827, no. 1, 2021, doi: 10.1088/1742-6596/1827/1/012138.

[20] C. Zhang, Y. Xu, Z. Xu, J. Huang, and J. Lu, "Hybrid handcrafted and learned feature framework for human action recognition," Appl. Intell., vol. 52, no. 11, pp. 12771–12787, 2022, doi: 10.1007/s10489-021-03068-w.

[21] Y. Zhao, K. L. Man, J. Smith, K. Siddique, and S. U. Guan, "Improved two-stream model for human action recognition," Eurasip J. Image Video Process., vol. 2020, no. 1, 2020, doi: 10.1186/s13640-020-00501-x.

[22] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos," Procedia Comput. Sci., vol. 133, pp. 471–477, 2018, doi: 10.1016/j.procs.2018.07.059.

[23] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li, "Survey of Spatio-Temporal Interest Point Detection Algorithms in Video," IEEE Access, vol. 5, pp. 10323–10331, 2017, doi: 10.1109/ACCESS.2017.2712789.

[24] A. S. Jahagirdar and M. S. Nagmode, "A Novel Human Action Recognition and Behaviour Analysis Technique using SWFHOG," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, pp. 571–580, 2020, doi: 10.14569/IJACSA.2020.0110475.

[25] R. S. Sheikh, S. M. Patil, and M. R. Dhanvijay, "Framework for deep learning based model for human activity recognition (HAR) using adapted PSRA6 dataset," Int. J. Adv. Technol. Eng. Explor., vol. 10, no. 98, pp. 37–66, 2023, doi: 10.19101/IJATEE.2021.876325.

[26] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects," Comput. Biol. Med., vol. 149, 2022, doi: 10.1016/j.compbiomed.2022.106060.