Composite Feature Vector Assisted Human Action Recognition through Supervised Learning

K. Ruben Raju, Yogesh Kumar Sharma, Birru Devender

Abstract: ***Human Action Recognition is a key research direction and also a trending topic in several fields like machine learning, computer vision and other fields. The main objective of this research is to recognize the human action in image of video. However, the existing approaches have many limitations like low recognition accuracy and non-robustness. Hence, this paper focused to develop a novel and robust Human Action Recognition framework. In this framework, we proposed a new feature extraction technique based on the Gabor Transform and Dual Tree Complex Wavelet Transform. These two feature extraction techniques helps in the extraction of perfect discriminative features by which the actions present in the image or video are correctly recognized. Later, the proposed framework accomplished the Support Vector Machine algorithm as a classifier. Simulation experiments are conducted over two standard datasets such as KTH and Weizmann. Experimental results reveal that the proposed framework achieves better performance compared to state-of-art recognition methods.***

Keywords: Action Recognition, Gabor, Wavelet, KTH, Weizmann, Accuracy.

I. INTRODUCTION

In recent years, analyzing and understanding the human actions have become one of the major challenges for future technical systems focused over the analysis of human behavior through visual sensors [1]. Acquiring the knowledge about the person's action is important and it is very crucial at several circumstances [2]. For instance, in automated visual surveillance systems, action analysis is very important which allows helping in the detection of potential threats emanating from a single person or a group of persons. Furthermore, in Human-Computer Interactions (HCI), the computer can analyze the actions to know the intentions and objectives of a user thereby provide an appropriate support, protect the user or guide the user.Action Recognition is a more crucial and important step in the human actions analysis which has vast variety applications in various fields like sports video analysis, intelligent surveillance [3], video retrieval, medical health care, smart home management [4] and 3D video games.

In earlier, a significant amount of research has been carried out in the field of Human Action Recognition (HAR) based on two types of sensors such as Wearable sensors [8] and video sensors. In the case of wearable sensor based HAR, the sensors are attached to the human body parts.

Revised Manuscript Received on February 15, 2020.

*Corresponding Author

K. Ruben Raju, Research Scholar, Dept. of Computer Science Engineering, JJT University, Rajasthan India.

Dr. Yogesh Kumar Sharma, Head & Associate Professor, Dept. of Computer Science Engineering, JJT University, Rajasthan India..

Dr. Birru Devender, Associate Professor, Dept. of Computer Science Engineering, Holy Mary Institute of Technology & Science, Hyderabad, India.

Various types of sensors or measuring equipment like gyroscopes, accelerometers and magnetometers are connected to the human body parts to measure the motion features which give prior information regarding the human action or movements [5-7]. However, the main disadvantages are that the wearable sensors can't be connected to human body for a long time, consequences to a restricted movement due to the wired connections, and also results in more complexity at the device settings.

In order to overcome these problems, video sensor based HAR technology has come into picture in which the human actions are monitored through cameras. With an advancement of imaging technology and the availability of cost-effective cameras, a new paradigm has been started in the HAR based on various types of image sequences acquired through these cameras. Compared to the wearable sensor technology, the images/videos acquired through the cameras reveal more information about the body parts and their movements. In the video sensor technology, the human actions are initially recorded by cameras and then they are fed to an automatic HAR system which analyzes the actions based on computer vision methods.Broadly the automatic HAR system is implemented in two phases, feature extraction and classification. In the feature extraction phase, the input video or image sequence is subjected to extract the features through which the action is represented. In the classification phase, the extracted features are processed for classification through machine learning algorithms.

Several feature extraction techniques are developed in earlier to extract an efficient feature set from an action image. Broadly they are classified as spatial domain techniques and transform domain techniques. In the spatial category, the pixel intensities and their derivatives such as Histogram of Gradients (HOGs) [11], space-time interest points (STIPs) [9, 10] are considered as features. In the transform domain, initially the action image is transformed through some transformation techniques like Fourier Transform, Wavelet Transform etc. Both the methods have their individual advantages and disadvantages. To overcome such disadvantages, this paper proposed a hybrid feature extraction technique by combining the Wavelet Filter with Gabor Filter.

For a given input action, this approach extracts both the Gabor feature maps and Wavelet feature maps and fuses them to form a single feature vector called as Fused Feature Vector (FFV), which provides a perfect discrimination between different actions Further, a novel super learning technique, machine Support Vector Machine (SVM) is considered as classifier. Simulation experiments are conducted over two standard benchmark datasets such as KTH dataset and Weizmann dataset. The performance of developed HAR system is analyzed through Recognition Accuracy.

Rest of the paper is organized as follows; Section II discusses the literature survey details. Section III discusses the complete details of proposed HAR framework. Section IV discusses the simulation experiments and performance analysis details and finally section V concludes the paper.

II. RELATED WORK

Recently, several methods have been developed for HAR based on the Spatio-temporal features of a video sequence and also have been demonstrated to achieve better results in recognizing human actions. Examples of such methods are 3-D speed-up robust features (3-D SURF) [12], 3-D HOGs [13], 3-D scale invariant feature transform (3-D SIFT) [14], and optical flow methods [15]. Further based on the applicability over the action frames, they are categorized as global and local feature extraction techniques.

Global feature extraction methods can capture the motion information of the complete human body, provides rich and expressive motion information for recognizing human actions. Generally, the human motion in a video creates a space-time shape in 3-D volume and this shape can capture both the dynamic information of human body and spatial information of human pose at different times. In the case of global representation, the entire human body is treated as singe region and hence it can capture the dynamic variations. Motion Energy Image (MEI) and Motion History Image (MHI) [16] are two popular methods through which the human action can be found where and how it is occurring. But they are not robust to view point changes and [17] introduced Motion history Volume (MHV) to remove the view-point dependency. In this approach, Fourier transform is used to create features invariant to rotations and locations. However, the common drawback of Global representation methods is that they are sensitive to noise variations in the action image. These methods capture the motion information in a particular rectangle area, and hence may introduce noise from cluttered background and can also introduce some irrelevant information.

In contrast to the global features, the local features represent the human motion through local Spatio-time regions. Since the information persists in these local regions is more salient and informative than the surrounding regions, they are only focused to detect in local feature representation. After detecting the informative regions, they are represented through various types of features. STIPs [18, 10] and Motion Trajectories [19-23]and are the two local feature representation techniques which had shown a better performance in the action recognition under varying translation and appearances. Laptev [24] applied Harries Corner detector to detect the space-time points and it applied Spatio-temporal separable Gaussian Kernel over a video to find the out the larger changes in Spatio-temporal motions.

Next, one more method is proposed by P. Dollar et al. [18] by considering the 2-D Gaussian Kernel along the spatial dimension and 1-D Gabor filter along the temporal dimension. For every interest point, gradients, optical flow vectors and raw pixels are extracted and formulated as a single feature vector. Further PCA is applied for dimensionality reduction and K-means clustering is applied for codebook construction. Further, Bregonzio e al. [10] accomplished Gabor filter and G. Willems [25] accomplished Hessian Matrix to detect the STIPs. However, the STIPs only acquire the information only for short-time but not capture long-time duration. Motion Trajectory is a better way to track these long-time STIPs [22].

H. Wang and C. Schmid [19] proposed an improved trajectory based action representation by considering the camera motion and to estimate the camera motion, a feature matching is accomplished between the frames using SURF descriptors and dense optical flow vectors. To further improve the recognition performance, Wang et al. [20] accomplished the feature matches to predict the homography. Further, this approach considered the Histogram of Optical Flows (HOF) and Motion Boundary Histograms (MBHs) to layout the spatial information. Furthermore, the concatenation of HOG with HOF and MBH features will give more efficient and continuous trajectories [21], [23]. Though the local feature representations methods has gained a great recognition performance, the computational complexity is more and also might be delivering the more information than the HAR system required.

Recently a simple HAR system has been developed by considering the basic filters and signal processing algorithms as feature extraction methods and machine learning algorithms as classifiers.S. Kanagamalliga and S. Vauski [26] used Gabor and Optical Flow features based contour model for motion estimation and to perform object tracking. This approach applied background subtraction over the obtained optical field through Expectation Maximization based Effective Gaussian Mixture Model and applied Adaboost algorithm for classification.Jin Jiang et al. [27] considered the Wavelet Packet Transform (WPT) as a feature extraction and Support Vector Machine algorithm as a classifier in the proposed Human activity recognition. Moreover, the SVM algorithm is optimized by determining optimal values for two kernel parameters through Improved Adaptive Genetic Algorithm (IAGA). Further, considering the advantages of Gabor filter and Ridgelet Transform, D. K Vishwakarma et al. [28] proposed a novel action recognition framework. The rotation and scale invariant property of Gabor Wavelet Transform (GWT) and the orientation dependent property of Ridgelet Transform are helped in the determination more discriminant properties between human actions.

M. H Siddiqi et al. [29] proposed to deploy a new dimensionality reduction technique, named as the Stepwise Linear Discriminant Analysis (SLDA) in the Human action recognition system after extracting the features through Wavelet Transform.

At wavelet based feature extraction, the ‘Symlet’ called wavelet filter is accomplished to extract the sub bands of action frames and to select most prominent features, SWLDA is applied. Finally, a well-known sequential classifier called hidden Markov model (HMM) [30] is applied to give the appropriate labels to the actions. Two standard datasets such as Weizmann and KTH are used to validate the developed system. Some more variants of Wavelet family are also accomplished as feature extraction techniques in the HAR system.

Manish Khare et al. [31] Applied the Discrete Wavelet Transform (DWT) to extract the features of an action frames at multiple resolutions. Next, E. Mohammadi et al. [32] applied 3D-DWT as a preprocessing step in the feature extraction phase for the HAR system proposed under two SVM kernels such as Sigmoid and Polynomial kernels. Further, DTCWT [33] has better edge representation and approximate shift-invariant properties compared to real-valued wavelet transforms. Considering these facts, Manish Khare et al. [34] proposed a method for human action recognition based on dual tree complex wavelet transform (DTCWT). KTH and MSR datasets are used for simulation validation. Recently, considering the correlation between wavelet coefficients, H. A. Moghaddam and Amin Zare [35] proposed Spatio-temporal wavelet correlogram (SWTC) as a new feature for human action recognition in videos. SWTC utilizes the multi-resolution and multi-scale property of wavelet transform and considered the correlation of wavelet coefficients.

III. PROPOSED HARFRAMEWORK

A. Overview of framework

This section discusses the complete details of proposed Human Action Recognition framework. The total framework is accomplished in two phases, one is training and another is testing. In the training phase, the HAR system is trained with larger number of videos with different actions. In the next phase, the trained HAR system is subjected to testing through various action videos giving as input one-by-one. In the both phases, the raw videos are not processed. Initially, the video is subjected to feature extraction to extract a set of features through which the action in the video can be represented. After feature extraction, in the training phase, the features of all videos of respective actions are trained and in the testing phase, they are fed to classifier. Here the classifier is the final stage in HAR system which has two inputs and output. One input is from trained dataset and another is from testing. The classifier performs a systematic classification by comparing the features of test action video with the features of trained action videos and produces one class label which denotes the name of action present in the test video.

In this paper, a new feature extraction technique is proposed by combining two different feature extraction techniques, they are Gabor Features and Wavelet Features. For a given action, the Gabor filter to extracts the scale and rotation invariant features and the wavelet filter extracts multi-resolution features. After these two features extraction, a composite feature vector is formulated by integrating them. Once the feature extraction phase is completed for a given action, then it is fed to classifier. This approach used Support Vector Machine as a classifier. The novelty of this approach is the formulation of a composite feature vector which is more efficient and provides a robust classification with scale and rotation invariant features. Unlike the conventional approaches, which accomplished either Gabor filter or Wavelet Filter, this approach combined both of them to achieve more accurate results. Furthermore, this approach also applied Principal Component Analysis to solve the dimensionality problem. The overall architecture of proposed framework is shown in figure.1.

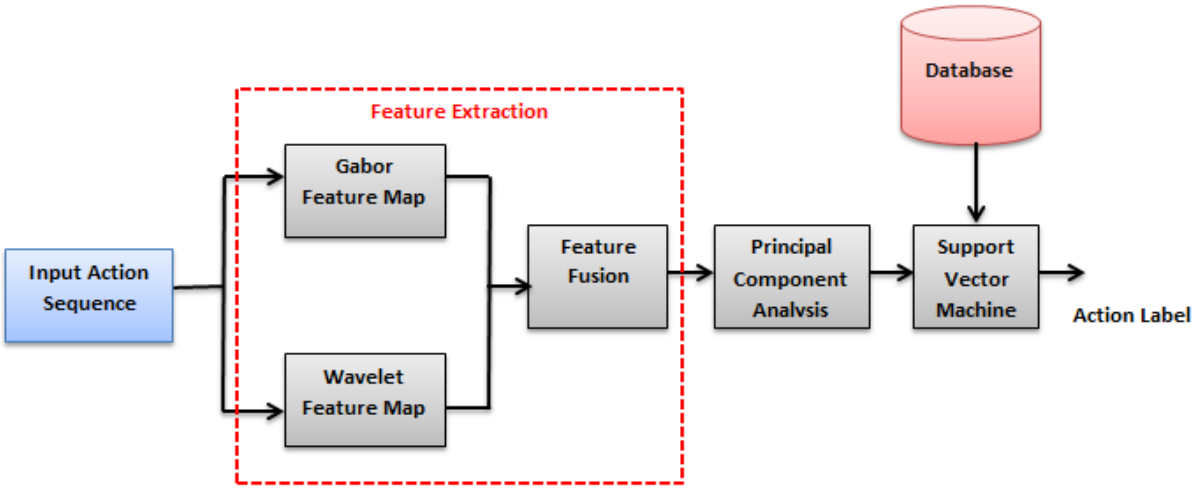


Figure.1 Block diagram of proposed HAR framework

B. Feature Extraction

Feature extraction is most important step in HAR system. Under this feature extraction phase, the input videos are processed to extract a sufficient set of features through which the HAR system can acquire required knowledge about the dynamics of Human action. Further, due to the variations of semantics in the Human actions, the feature set extracted for one action is different from the features extracted from another action. This difference ensures a sufficient discrimination between actions and helps in achieving maximum recognition accuracy. In this paper, the proposed feature extraction technique is a hybrid technique, which is composed of two different feature extraction techniques such as Gabor filter based feature extraction and wavelet filter based feature extraction. The main intention of Gabor filter is to extract the features at multiple orientations and the wavelet filter is for multi-resolution features. Finally, the multi-orientation features and multi-resolution features are formulated into a single feature vector, called as Fused Feature Vector (FFV). The detail about individual feature extraction techniques is illustrated in the following subsections.

a. Gabor Features Gabor filter is a most popular filter which is generally used in various image oriented applications. The main advantage of Gabor filter is it can represent an image in multiple orientations, i.e., we can analyze an image and the possible dominant features at various angles. Though the Gabor we can study the image variants at multiple angles. In this paper, Gabor filter is applied with eight orientations: 0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, and 315^0 and with six different scales: 5×5, 7×7, 9×9, 11×11, 13×13 and 15×15. Hence the total feature maps obtained after the accomplishment of Gabor Filter is $8 \times 6 = 48$. The mathematical response of Gabor filter is defined as;

$$G(x,y) = exp\left(\frac{X^2+\gamma Y^2}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}x\right) \quad (1)$$

Where

$$X = xcos\theta - ysin\theta, \;\; Y = xsin\theta + ycos\theta \quad (2)$$

and (x, y) is position relative to the center of filter. According to the mathematical expressions (1) and (2), the θ value varies from 0^0 to 315^0 with an angular deviation of 45^0. For instance, let's consider the scale 7×7, initially the action frame is scaled and then the Gabor filter is applied over it for eight orientations. Similarly, the action frame is processed for remaining scales also and hence we obtain totally 48 feature maps. The obtained 48 feature maps are belongs to only one action frame and a video consists of N number of frames. For example, if a video consist of 100 frames, then the total feature maps obtained for a video sequence is $100 \times 48 = 4800$, which is a very larger count. This creates a heavy computational overhead at the classification level and also needs more memory to store at the database. Hence, to solve these problems, a max pooling is applied at every scale to extract only key feature maps. Let's $S = \{5 \times 5, 7 \times 7, 9 \times 9, 11 \times 11, 13 \times 13, 15 \times 15\}$ and $\theta = \{0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, 315^0\}$, the max-pooling at different scales is formulated as;

$$F_{max,i} = \max_{\substack{i=1\ to\ length(\theta)\\ j=1\ to\ length\ (S)}}(x,y)\{f_{S_i}(x,y,\theta_j)\} \quad (3)$$

Where f_{S_i} represents the feature map at i_{th} orientation and θ_j represents the j_{th} scale. For $i = 1$, the orientation $\theta = 0^0$ is picked up and the feature maps obtained at four scales are chosen and the expression (3) picks up the final feature map with all maximum values. For a given co-ordinate (x, y), the expression (3) searches for maximum value in the total four feature maps obtained at i_{th} orientation. Hence, finally we obtain totally eight feature maps which cover almost all scale and rotation invariant features for a given action frame. The max-pooling based feature selection has the following advantages. 1). It extracts a key feature map for action recognition and provides a more robust response for an action under cluttered background and occlusions. 2). It reduces the unnecessary computational burden by reducing the feature maps at different orientations. 3). Results in a scale and rotation invariant feature map which is more helpful in the recognition of an actions captured at different angles and scales. The Gabor filtered outputs for a scaled action image is shown in fgure.2.

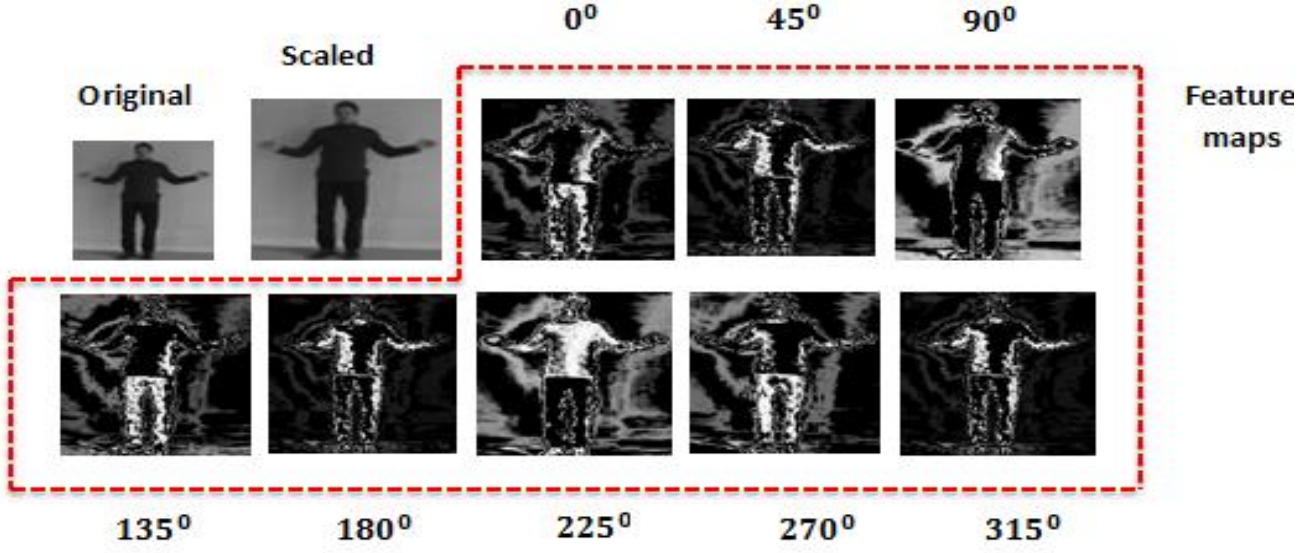


Figure.2 Gabor filtered outputs for a scaled action image

b. Wavelet Filter

Wavelet transform is an efficient transform technique which provides a simultaneous representation of an action in both spatial and frequency domain. In wavelet transform, this effective representation is possible due to the decomposition of an image over dilated (scale) and translator (time) version of a wavelet. Hence, most of the signal, image oriented applications prefers wavelet transform for feature analysis based on the accomplishment of shifting and scaling through wavelet filters. Wavelet family has number of variants like Discrete Wavelet Transform (DWT), Complex Wavelet Transform (CWT) and Wavelet Packet Transform (WPT) etc. Among those variants, DWT and DTCW have gained much importance due to their hierarchical decomposition structure.

1. DWT

DWT is the most popular wavelet transform which represents the image as a linear combination of its basis functions.

In DWT, the image is decomposed into several frequency bands called as Approximations (CA), Horizontals (CH), Verticals (CV) and Diagonal Details (CD). To obtain these four bands, the action image need to be subjected to decomposition through 2-D DWT where there exists one 2-D scaling function $\varphi(x,y)$ and three 2-D wavelet functions such as $\psi^H(x,y), \psi^V(x,y)$ and $\psi^D(x,y)$. The scaling function is used to approximate the image at different level of approximations. In DWT, the scaling functions and the wavelet functions are formulated as the product of one scaling function φ and the respective wavelet function ψ which are modeled as;

$$\varphi(x,y) = \varphi(x)\varphi(y) \quad (4)$$

$$\psi^H(x,y) = \psi(x)\varphi(y) \quad (5)$$

$$\psi^V(x,y) = \psi(y)\varphi(x) \quad (6)$$

$$\psi^D(x,y) = \psi(x)\psi(y) \quad (7)$$

Where $\psi^H(x,y)$ evaluates the horizontal variations, $\psi^V(x,y)$ evaluates the vertical variations and the $\psi^D(x,y)$ evaluates the variations along the diagonal. The wavelet functions are mainly used to detect the edges in the image. The block diagram of DWT is shown in figure.3 and the obtained output bands after its accomplishment over an action frame is shown in figure.4.

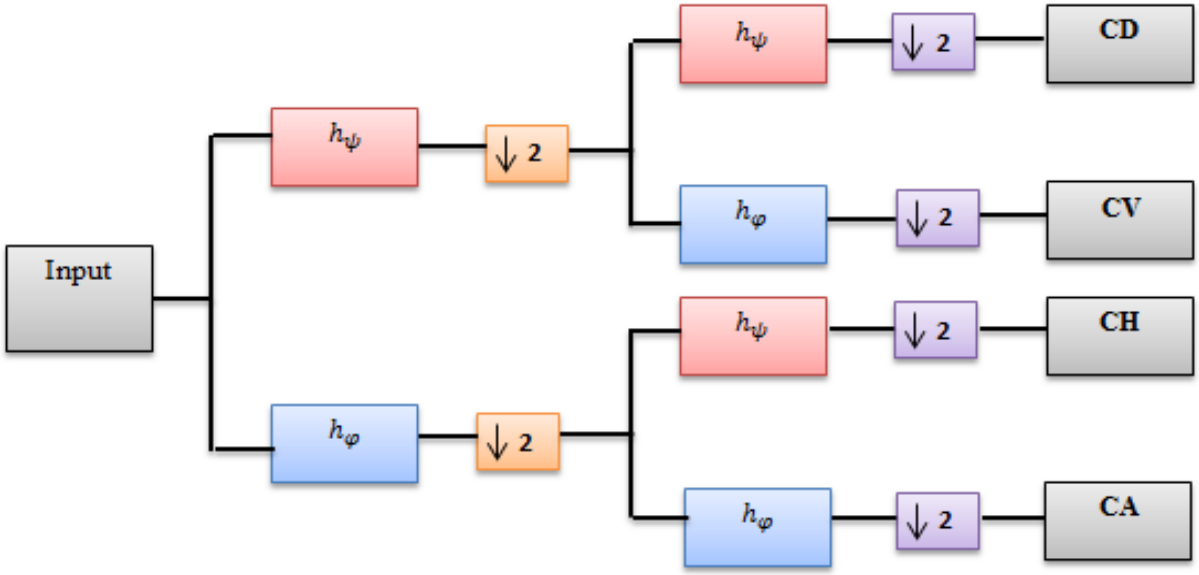


Figure.3 Block diagram of DWT

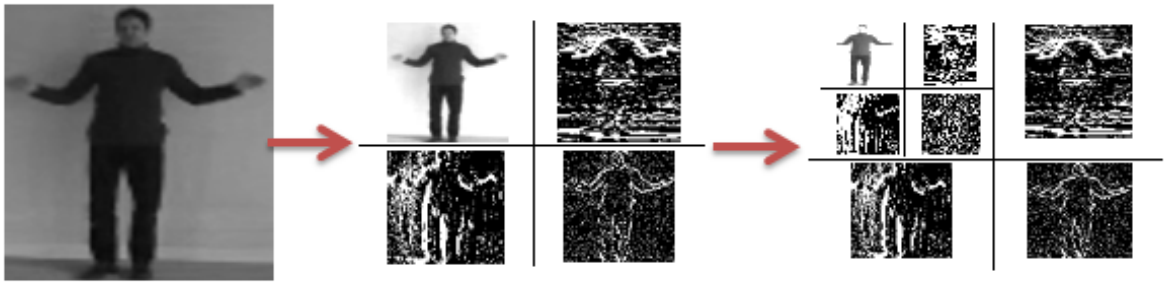


Figure.4 Two level decomposition of 2-D DWT

2. DTCWT

DTCWT is one of the most important and effective transform in the wavelet family. This was developed to overcome the major problem of DWT, i.e., lack of shift invariance. Though the traditional DWT has gained effective results in the image decomposition, it suffers from several problems and the lack of shift invariance is the most distinct problem due to which the reconstructed signal will have distortions. The main reason behind this problem is the presence of a down sampling module at every stage of DWT implementation. Due to this reason, the shifts in the input signal can't be analyzed through DWT coefficients. The best solution to overcome this problem is an accomplishment of an un-decimated DWT. But, this solution consequences to a higher computational cost followed by very high redundancy in the obtained sub bands.

At this instant, DTCWT has come into picture which can overcome the DWT problem without any removal of down-sampler. By finding complex wavelet coefficients which are 90^0 out of phase with each other, this problem can be solved and here the DTCWT follows the same procedure which is inspired from the Fourier transform. Because, the Fourier coefficients obtained are of complex sinusoid form and they constitute a Hilbert transform Pair.

Further the Fourier transform has no problem of shift invariance, the Fourier coefficients are perfectly shift variant. Considering these facts, the DTCWT employs a complex valued wavelet and scaling functions and they are defined as follows;

$$\varphi_S(x,y) = \left(\varphi_{rs}(x) + j\,\varphi_{is}(x)\right)\left(\varphi_{rs}(y) + j\,\varphi_{is}(y)\right) \quad (8)$$

$$\psi_w^H(x,y) = \left(\varphi_{rs}(x) + j\,\varphi_{is}(x)\right)\left(\psi_{rw}^H(y) + j\,\psi_{iw}^H(y)\right) \quad (9)$$

$$\psi_w^V(x,y) = \left(\psi_{rw}^V(x) + j\,\psi_{iw}^V(x)\right)\left(\varphi_{rs}(y) + j\,\varphi_{is}(y)\right) \quad (10)$$

$$\psi_w^D(x,y) = \left(\psi_{rw}^D(x) + j\,\psi_{iw}^D(x)\right)\left(\psi_{rw}^D(y) + j\,\psi_{iw}^D(y)\right) \quad (11)$$

Where $\psi_w(x,y)$ is a wavelet coefficient obtained from the real wavelet coefficient $\psi_{rw}(x,y)$, and imaginary wavelet coefficient, $\psi_{iw}(x,y)$. Similarly, the term $\varphi_s(x,y)$ is a scaled coefficient and it is obtained from the real scaled coefficient $\varphi_{rs}(x,y)$, and imaginary scaled coefficient, $\varphi_{is}(x,y)$. The DTCWT applies two wavelet filters, one extracts the real part and other extracts the imaginary part. These two filters combination is called as analytical filter. The structure of three level DTCWT is shown in figure.5, where the upper DWT is called as real part of Complex wavelet transform and the lower tree of DWT is called as imaginary part of complex wavelet transform.

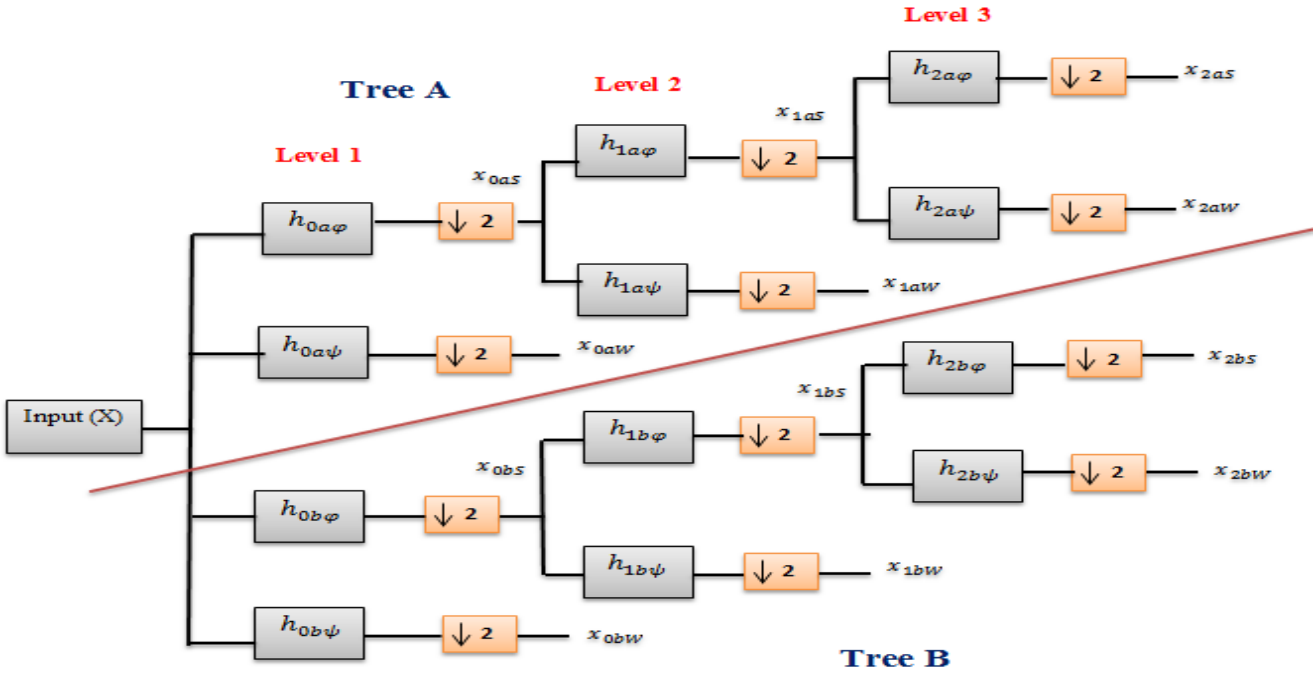


Figure.5 Three level DTCWT architecture

In this paper, the DTCWT is accomplished for five levels of decomposition over an action frame from a given action sequence and at every level, only the approximation is considered for further decomposition to construct a wavelet pyramid. Finally, to obtain a wavelet pyramid feature map, the adjacent levels are subtracted, i.e., $F_w = W_i - W_{i-1}$, where W_i and W_{i-1} are the adjacent levels from a wavelet pyramid and F_w is the final wavelet feature map.

After the extraction of Gabor feature maps and wavelet feature maps, they are concatenated and formulated into a single feature vector called as FFV. Since the FFV is composed of larger number of features, to reduce the dimensionality problem, Principal Component Analysis is accomplished and 90% of principal components are considered as required features.

IV. SIMULATION RESULTS

To evaluate the performance of developed HAR system, we used two different and standard benchmark datasets that include KTH dataset and Weizmann Dataset. To simulate the proposed model, this paper used MATLAB software. The details of the two datasets and the results obtained after the deployment of proposed HAR over them are discussed in the following subsections. Furthermore, the comparative analysis between evaluated between proposed and conventional approaches is also described here.

A. Datasets

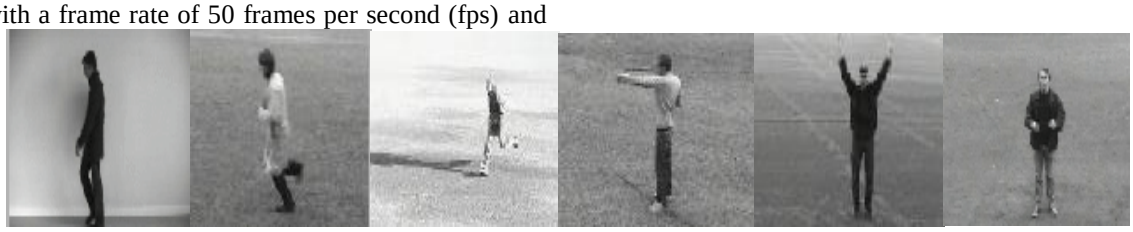
a. KTH Dataset

This dataset consists of totally six different actions such as *Handclapping*, *Hand waving*, *Boxing*, *Running*, *Jogging* and *Walking* [36]. All these actions are performed under four different environments such as Outdoors, Indoors, Outdoors with several scales and with several clothes. The same actions are performed several times by totally 25 subjects and hence the total number of actions videos present in this dataset are $25 \times 6 \times 4 = 600$. All the videos are in unique format, i.e., .AVI format and total sequences present in this dataset are 2391. The entire action video set is captured through a static camera with a frame rate of 25 frames per second (fps) with a homogeneous background. the resolution of each frame is 160×120. Some examples frames of this dataset are shown in figure.6.

b. Weizmann Dataset

This dataset totally consists of totally ten different actions like *walking*, *running*, *skipping*, *jumping jack*, *jump forwrad on two legs*, *jump in place on two legs*, *gallopsideways*, *handwaving*, *wave one hand*, *and bend* and totally consist of 90 videos [37].

All the actions are captured with the help of a static camera. The entire action video set is captured through a static camera with a frame rate of 50 frames per second (fps) and the resolution of each frame is 180 × 144. Some examples frames of this dataset are shown in figure.7.



(a) (b) (c) (d) (e) (f)

Figure.6 KTH samples, (a) Walking, (b) Jogging, (c) Running, (d) Boxing, (e) Handwaving, (f) Handclapping

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

Figure.7 Weizmann samples, (a) Bend, (b) Jump, (c) Run, (d) Boxing, (e) Wave two hands, (f) Walk

B. Performance Metrics and Results

To measure the performance of developed HAR system, this paper accomplished several performance metrics like Detection Rate, Precision, F-Measure and Accuracy. These metrics are implemented accoridng to the mathematical expression shown below;

$$Detection\ Rate\ (Recall) = \frac{TP}{TP+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

$$F - Measure = \frac{2\times Recall\times Precision}{Recall+Precision} \quad (14)$$

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN+TP} \quad (15)$$

$$False\ Discovery\ Rate\ (FDR) = \frac{FP}{FP+TP} \quad (16)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

All these metrics are obtained based on the secondary metrics derived from confusion matrix. In the above expressions, the term TP stands for True Positive, TN stands for True Negative, FP stands for False Positive and FN stands for False Negative. The basic confusion matrix from which these secondary metrics are derived is shown in table.1. After testing the actions from both datasets, these performance metrics are measured and they are shown in table.2 and table.3.

Table.1. Sample Confusion Matrix

		Predicted Action	
		Action Class 1	**Action Class 2**
Actual Action	**Action Class 1**	TP	FN
	Action Class 2	FP	TN

Table.2 Performance Measures of different actions in KTH dataset

Action/Metric	Detection Rate (%)	Precision (%)	F-Measure (%)	FNR (%)	FDR (%)
Walking	95.3647	98.4751	96.8949	4.6353	1.5249
Jogging	88.4178	94.7814	91.4891	11.5822	5.2186
Running	83.7941	90.4412	86.9909	16.2059	9.5588
Boxing	100.00	99.8512	99.9255	0.0000	0.1488
Hand Clapping	85.4175	91.7896	88.4890	14.5825	8.2104
Hand Waving	92.4147	96.4578	94.3930	7.5853	3.5422

Table.3 Performance Measures different actions in Weizmann dataset

Action/Metric	Detection Rate (%)	Precision (%)	F-Measure (%)	FNR (%)	FDR (%)
Bend	98.1123	97.2975	97.7032	1.8877	2.7025
Jumping Jack	98.5674	97.7526	98.1583	1.4326	2.2474
Jump forward	97.1457	96.3309	96.7366	2.8543	3.6691
Jump in space	100.00	99.1852	99.5909	0.0000	0.8148
Running	88.7419	87.9271	88.3326	11.2581	12.0729

Gallop side ways	92.8127	91.9979	92.4035	7.1873	8.0021
Skipping	85.6631	84.8483	85.2538	14.3369	15.1517
Walking	94.9987	94.1839	94.5895	5.0013	5.8161
Wave one hand	100.00	99.5610	99.7800	0.0000	0.4390
Wave two hands	95.7843	94.9695	95.3752	4.2157	5.0305

Table.2 and Table.3 reveals the details of performance measures of different actions of KTH and Weizmann datasets respectively. The first measure, detection rate (recall) measures the total number of true positives for a given total number of inputs. For example, let's consider the walking action, the recall is measured as the ratio of total number of walking action frames detected as walking to the total number of walking actions frames given as input for testing process. In this case, the TP is the total number of walking action frames classified correctly and FN is the total number of walking actions frames classified incorrectly. This process is applied for all the remaining actions also and the respective Detection rates are measured.
Next, the precision is measured as the ratio of TPs to the sum of TP and FP. In the above example of walking action as input, the TP is the total number of walking action frames classified correctly and FP is the total number of action frames classified as walking when the input is not a walking action. In this case, the input is not required action but the output is required action.
Next, the F-measure a simple harmonic mean of recall and precision. The FNR and FDR are having inverse relations with Recall and Precision respectively. From table.2 it can be noticed that the maximum recall (100%) and precision (99.8512) is observed for *boxing* action and minimum recall (83.7941%) and precision (90.4412) for *running* action. Simultaneously, the maximum FNR (16.2059%) and FPR (9.5588%) are for running action and minimum FNR (0%) and FPR (0.1488%) is for boxing action. The boxing actions have constant movements and also we can observe minor movements in the body, gives a detailed structure of action for the HAR system which results in a more detection rate. In contrast to the boxing action, in the running action we can observe more variations in the movement of human parts which results a more confusion to HAR system, results in less detection rate and precision.
Next, from the Table.3, we can observe that the *Wave one hand* action and *jumping in space* action has gained a maximum performance with respect to recall and precision. Compared to the remaining actions, these two actions have less variation in the movement of body parts and hence the HAR system is efficiently detected them. The maximum recall rate observed from table.3 is 100% and minimum recall rate is 85.6631% for Wave *one hand* action and *jumping in space* and *skipping* actions respectively. Similarly the maximum precision (99.5610%) is observed for *Wave one hand* action and minimum (84.8483%) is observed for *skipping* action.

C. Comparitive Analysis

To alleviate the performance enhancement, a comparitive analysis is performed between the results obtained through propsoed approach and conventional approaches such as DWT [31], DTCWT [34] and Gabor transform with Ridgelet Transform [28]. Manish Khare and MoonguJeon [31] used DWT as a feature extraction technique to recognize the human activities. DWT is an effective technique in the extrcation of contour features of a human body but due to the non-shift invariant property, the features at differnet shifts cannot be extracted properly. This is due to the presnece of a downsampler at every decomposition level. This problem is solved by M. Khare et al. [34] by using DTCWT. In DTCWT, the action image is simultaneously undergone through the two phase decomposition such as Real part and imaginary part decompositions. Even though the DTCWT has downsamplers, due to the two phases, it will acquire shift invarinat property. Though the wavelet family has achieved good recogniton results they are not roust for scale and rotation variations in the action image. This drawback is solved by Gabor filter which provides a scale and rotation invarinat features. By considering this advantage, D. K. Vishwakarma et al. [28] combined Gabor Transform with Ridgelet Transform to perform human activity recognition. However, compared to RT, DTCWT has more efficiency in the feature representation for shift invariants and hence the proposed approach combined the Gabor transform with DTCWT for action reocgnition. In the comparitive analysis, the propsoed approach and conventional approaches are compared through Recall, Precision, F_meausres, FNR, FDR and Accuracy for both KTH and weizmann datasets.
Fiugre.8 shows the comparitive analysis between the proposed and conventional approaches through Recall. As it can be observed from the above figure, the Recall of proposed approach (Gabor with DTCWT) is high compared to the conventional approaches.
Since the proposed approach can find both scale and rotation invariant features of an action, the actions can be discriminated more effectively and hence results in higher recall and it is approximately 93.7618% whereas it is approximately 80.5303%, 90.8203% and 91.8103% for DWT, DTCWT and Gabor with RT respectively. A higher number of TPs results in higher recall and this is achieved when the system has classified almost all the input actions correctly.

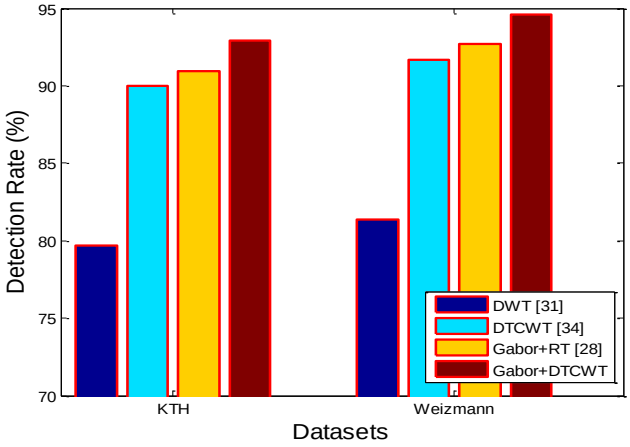


Figure.8 Detection rate comprison for various datasets

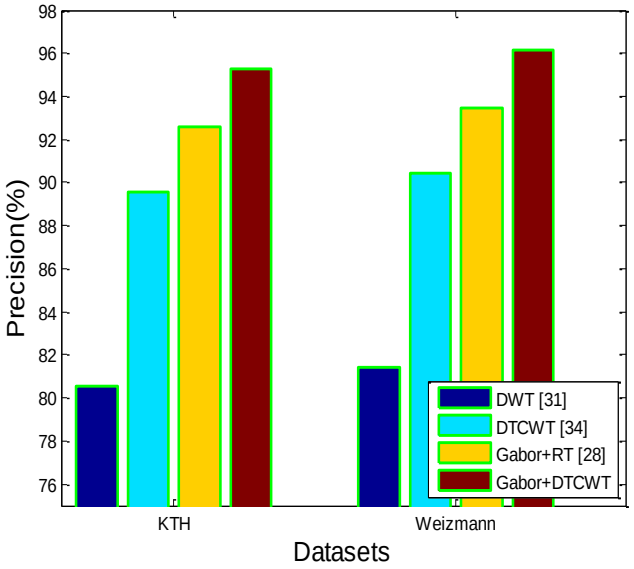


Figure.9 Precision comprison for various datasets

Fiugre.9 shows the comparitive analysis between the proposed and conventional approaches through Precision. As it can be observed from the above figure, the Precision of proposed approach (Gabor with DTCWT) is high compared to the conventional approaches. On an average, the precision of proposed approach is observed as 96.1503% whereas it if of 81.4003%, 90.4103% and 93.4203% for conventional approaches DWT, DTCWT and Gabor with RT respectively. A higher number of TPs and lower number of FPs leads to a higher precision. For a given number of action images, the total number of actions classified correctly increases the precicion and the proposed approach has achieved a higher precision compared to the conventional appraches.

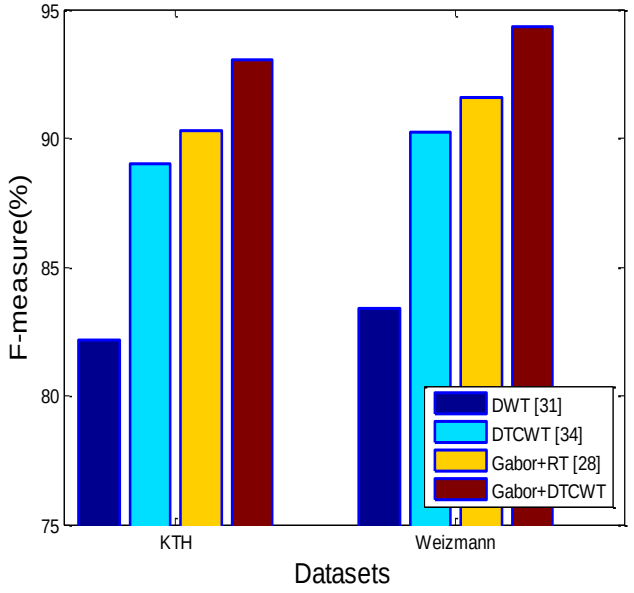


Figure.10 F-measure comprison for various datasets

F-Measure is the harmonic mean of precision and recall. As the precision and recall are high, the F-Measure is also high. From the above figure, it can be observed tha the proposed approach has an average F-meausre of 94.9410% and for conventional aproaches, it is of 80.9630%, 90.6148% and 92.6083% for DWT, DTCWT and Gabor with RT respectively.

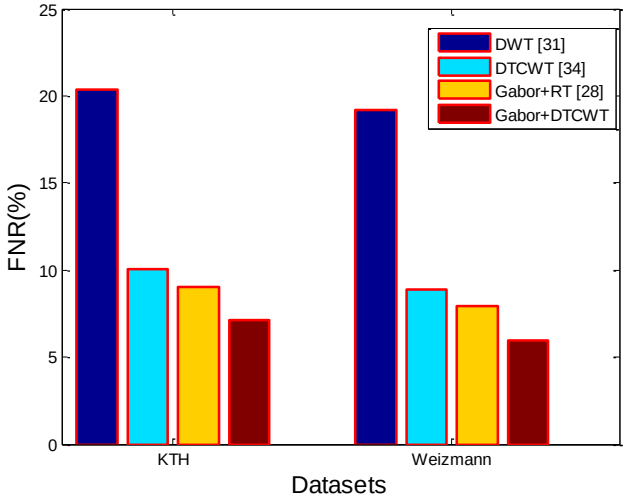


Figure.11 FNR comprison for various datasets

FNR measures the total number of false negatives from the total outputs. For a given N number of action images, the FNR is meausred as tht tpotal number of actions those are wrongly clasisified, i.e, the input action is one and out label is another.For example, for an input frame having *walking action*, if the system had shown it as some *other action*, then it is considered as FN. FNR is measured by suumming up such types of outputs. A system with less FNR is considered as more efficient. According to figure.11, the FNR of proposed approach is less compared to the conventioal approaches. Moreover, FNR obtained by subtracting the Recall from 100 and hence there exists an opposite relation between recall and FNR. From the above figure, on an average, the proposed approach has obtained 6.2382% FNR and for conventional approaches, it is observed as 19.4697%, 9.1797% and 8.1897% for DWT, DTCWT and Gabor with RT respectively.

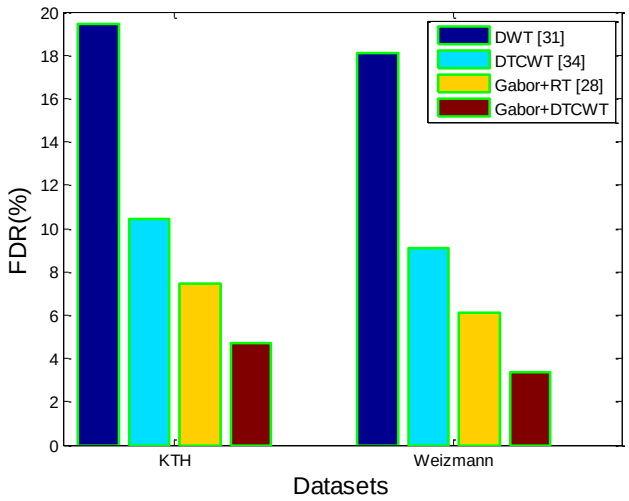


Figure.12 FDR comprison for various datasets

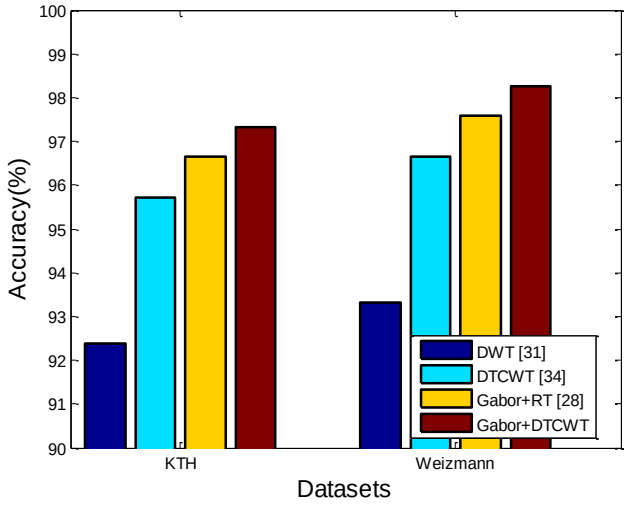


Figure.13 Accuracy comparison for various datasets

FDR measures the total number of false discovers from the total outputs. For a given N number of action images, the FDR is measured as total number of falsely discovered outputs. For an action, the FDR is meausred as total number of instances at which it is obtained as output when the input is other action. A system with less FDR is considered as more efficient. According to figure.12, the FDR of proposed approach is less compared to the conventioal approaches. Moreover, FDR obtained by subtracting the Precision from 100 and hence there exists an opposite relation between precision and FDR. From the above figure, on an average, the proposed approach has obtained 3.8497% FNR and for conventional approaches, it is observed as 18.5997%, 9.5897% and 6.5797% for DWT, DTCWT and Gabor with RT respectively.

Next, Figure.13 shows the accuracy comparisons between proposed and conventional approaches over KTH and Weizmann datasets.

From this figure, the average accuracy of proposed approach is measured as 97.6949% and for conventional approaches, it is observed as 92.7549%, 96.0849% and 96.2249% for DWT, DTCWT and Gabor with RT respectively. Since the proposed approach combined Gabor features (scale and rotation invarinat) and Shift invariant features, the accuracy of proposed approach is high compared to simple DWT [31], DTCWT [34]. Furtermore, the proposed approach also gained an improved accuracy than the method proposed by D. K. Vishwakarma et al. [28] in which the Gabor features are combined wuth Ridgelet features. The RT has similar characteristics with DWT and DTCWT is more efficient than DWT, the proposed approach achieved a higher recognition accuracy. The main reason is that the proposed features are more effective in the provision of a perfect discrimination between different actions. Moreover, as the feature count increases, a classifier will get more clairty and classifies the given input actions more accurately.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we presented an end-to-end framework for human action recognition. This framework is based on the study of a more distinctive and discriminative features through which the action can be analyzed more perfectly. This method considered the Gabor filter and DTCWT as feature extraction techniques and SVM algorithm as a classifier. These two feature extraction techniques are more effective in the extraction of scale, rotation and shift invariant features of an action image. Based on these features, the classifier got more clarity on the internal motion dynamics of human body and thus resulted in a more accurate recognition results. For simulation purpose, we accomplished two datasets such as KTH and Weizmann and the performance is measured through Recall, Precision and Accuracy. Compared to the existing state-of-art methods, our proposed recognition framework achieved improved recognition Accuracy and less Misclassification rate.

Focusing over the further improvisation of recognition accuracy, the future work of this paper can be directed in the reduction of computational time. Due to the more number of features extracted at feature extraction phase, the computational time required for classification will increase and hence the future work is focused over the computational time reduction by detecting the self-similarities between the frames of a single action video.

REFERENCES

1. Ronald Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding (*CVIU*) 108 (1-2) (2007) 4–18.
2. A.Veeraraghavan, A. Roy Chowdhury, R.Chellappa, Matching shape sequences in video with applications in human movement analysis,*IEEETrans. PatternAnal.Mach.Intell*.27(12)(2005)1896–1909.
3. J.Kim, D.Yeom, Y.Joo, Fast and robust algorithm of tracking multiple moving objects for intelligent video surveillance systems, *IEEE Trans. Consum. Electron*.57 (2011) 1165–1170.
4. A.Jalal,N.Sharif,J.Kim,T.Kim,Humanactivityrecognitionviarecognize d body partsofhumandepthsilhouettesforresidentsmonitoringservicesat smart homes,IndoorBuiltEnviron.22(2013)271–279.
5. J. Ward, P. Lukowicz, G. Troster, T. Starner, Activity recognition of assembly tasks using body-worn microphones and accelerometers, *IEEE Trans. Pattern Anal. Mach. Intell*. 28 (2006) 1553–1567.
6. Kwapisz, J.; Weiss, G.; Moore, S. Activity recognition using cell phone accelerometers. *ACM SigKDDExplor. Newsl*. 2011, 12, 74–82.
7. Siirtola, P., Röning, J., Recognizing Human Activities User-Independently on Smartphones Based on Accelerometer Data. *Int. J. Int. Multimed. Artif. Intell*. 2012, 1, 38–45.
8. Zhang, M.; Sawchuk, A.A. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. *In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh*, PA, USA, 5–8 September 2012; ACM: New York, NY, USA; pp. 1036–1043.
9. I. Laptev and T. Lindeberg, Space-time interest points, *in ICCV*, 2003, pp. 432–439.
10. M. Bregonzio, S. Gong, and T. Xiang, Recognizing action as clouds of space-time interest points, *in CVPR*, 2009.
11. Dalal, N.,and Triggs, B,. Histograms of oriented gradients for human detection. *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
12. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
13. A. Klaser and M. Marszalek, "A Spatio-temporal descriptor based on 3Dgradients," in *Proc. 19th Brit. Mach. Vis. Conf.*, Leeds, British, 2008, pp. 995–1004.
14. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. 15th Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 357–360.
15. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Anchorage, AK, 2008, pp. 1–8.
16. A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

17. D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.
18. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse Spatio-temporal features," *in ICCV VS-PETS*, 2005.
19. H. Wang and C. Schmid, "Action recognition with improved trajectories," *in IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.
20. H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *IJCV*, 2015.
21. H.Wang, A. Klaaser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 60-79, 2013.
22. H. Wang, A. Kl¨aser, C. Schmid, and C. L. Liu, "Action Recognition by Dense Trajectories," *in IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176.
23. M. Jain, H. J´egou, and P. Bouthemy, "Better exploiting motion for better action recognition," *in CVPR*, 2013.
24. I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2, pp.107–123, 2005.
25. G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale invariant Spatio-temporal interest point detector," *in ECCV*, 2008.
26. S Kanagamalliga, and S. Vasuki "Contour-based object tracking in video scenes through optical flow and Gabor features", *Optik*, Volume 157, March 2018, Pages 787-797.
27. Jin Jiang, Ting Jiang, and ShijunZhai, "A novel recognition system for human activity based on wavelet packet and support vector machine optimized by improved adaptive genetic algorithm", *Physical Communication*, Volume 13, Part C, December 2014, Pages 211-220.
28. D. K.Vishwakarma, PrachiRawat, and RajivKapoor, "Human Activity Recognition Using Gabor Wavelet Transform and Ridgelet Transform", *Procedia Computer Science*,Volume 57, 2015, Pages 630-636.
29. Muhammad Hameed Siddiqi, Rahman Ali, Md. SohelRana, Een-Kee Hong , EunSoo Kim and Sung young Lee, "Video-Based Human Activity Recognition Using Multilevel Wavelet Decomposition and Stepwise Linear Discriminant Analysis", *Sensors* 2014, 14, 6370-6392.
30. Kodagoda, S.; Piyathilaka, J. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. *In Proceedings of the 2013 IEEE 8th International Conference on Industrial Electronics and Applications*, Melbourne, Australia, 19–21 June 2013; pp. 567–572.
31. Manish Khare and MoonguJeon, "Towards Discrete Wavelet Transform-based human activity recognition", Proc. SPIE 10443, Second International Workshop on Pattern Recognition, 19 June 2017.
32. EmanMohammadi, Q. M. Jonathan Wu, Yimin Yang, MehrdadSaif, "Effect of wavelet and hybrid classification on action recognition", *IEEE International Conference on Image Processing (ICIP*), Bejing China, 2017.
33. Selesnick, Ivan W.; Baraniuk, Richard G.; Kingsbury, Nick G., "The Dual-Tree Complex Wavelet Transform", *EEE Signal Processing Magazine*. 22 (6): 123–151, 2005.
34. Manish Khare, JeonghwanGwak, and MoonguJeon, "Complex wavelet transform-based approach for human action recognition in video", *International Conference on Control, Automation and Information Sciences (ICCAIS*), Chiang Mai, Thailand, 2017.
35. H. A. Moghaddam and Amin Zare, "Spatiotemporal wavelet correlogram for human action recognition", *International Journal of Multimedia Information Retrieval*, Vol.8, Issue 3, 2019, pp.167-180.
36. C. Sch¨uldt, I. Laptev, and B. Caputo, Recognizing human actions: A local SVM approach, *in IEEE ICPR*, 2004.
37. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, Actions as space-time shapes, *in Proc. ICCV*, 2005.

AUTHORS PROFILE

Mr. K. Ruben Raju is presently Research Scholar in JJT University, Rajasthan and working as Assistant Professor of CSE at Sphoorthy Engineering College, Nadergul, Hyderabad. His areas of interests are Image Processing, Computer Vision, and Machine Learning. He published more than 10 papers in various International Journals and Conferences.

Dr. Yogesh Kumar Sharma is presently working as Head of Department, and Associate Professor of CSE at JJT University, Rajasthan. He is also working as Research Coordinator at JJTU. His areas of interests areComputer Networks, Cloud Computing, Data Mining, Big Data, Machine Learning, and Data Science. He published more than 20 papers in various International Journals and Conferences and guiding several PhD Scholars

Dr. Birru Devender is presently working as an Associate Professor of CSE at Holy Mary Institute of Technology & Science, Hyderabad, Telangana, India. His areas of interests are Machine Learning, Cloud Computing, and Networking. He published several papers in various International Journals and Conferences and guiding several PhD scholars.