HUMAN ACTION RECOGNITION THROUGH FUSED FEATURE VECTOR AND KERNEL DISCRIMINANT ANALYSIS

K Ruben Raju¹, Yogesh Kumar Sharma², Birru Devender³

¹Research Scholar, Dept. of Computer Science Engineering, Shri JJT University, Rajasthan, India.

²Head & Associate Professor, Dept. of Computer Science Engineering, Shri JJT University, Rajasthan, India.

³Associate Professor, Dept. of Computer Science Engineering, Holy Mary Institute of Technology & Science, Hyderabad, India

Email: rubenrajk@gmail.com

Abstract

Aimed at the problems of Intensity, Contour and orientation information, a Human Action Recognition (HAR) method based on Fused Feature Vector (FFV) is proposed in this paper. The FFV is constructed based on three different features such as Intensity features, Gradient features, and Orientation features. These three set of features are obtained through three different feature extraction methods based on Gaussian Filter, Gradient Filter and Gabor filter. Further to ensure optimal discriminant subspace, Kernel Discriminant Analysis is employed as a dimensionality reduction technique. Given the FFV of each action image, Support Vector Machine (SVM) is employed for classification. The proposed recognition model is evaluated systematically on the three public datasets such as KTH dataset, Weizmann dataset and the challenging UCF YouTube action dataset. Experimental results prove that our method outperforms the conventional approaches in terms of recognition accuracy.

Keywords: Human action recognition, Gaussian, Gradient, Gabor, Kernel Discriminant Analysis, Support vector Machine, Recognition Accuracy.

I. Introduction

In the field of computer vision, a substantial amount of work has been done in the pattern recognition that involves the detection and identification of the objects in the video sequences. In this area, due to the addition of temporal expression, the power of camera is increased drastically and employed to solve a large variety of complex and composite problems. Human Action Recognition (HAR) is one of such
Recognizing human actions from a video is a challenging task in many practical applications. Basically, the HAR is accomplished under three phases such as Pre-processing, Feature extraction and classification. Under pre-processing phase, the action video is image is subjected to initial processing like quality enhancement, action unit detection etc. The most important stage of HAR is feature extraction. To achieve a better performance in HAR, feature extraction need to be more effective and this is the key solution for HAR problems. Under this phase, the all possible features are need to be extracted from an action image/video and trained such that the action present in the test image/video under any variation must be recognized. Based on this inspiration, several feature extraction approaches are developed in earlier. Based on the methodology employed, they are categorized as Global based method and local based methods [XXXIV].

Under global feature extraction, the entire action image is employed for feature extraction. In this category, the feature of actions is extracted as Spatio-temporal shapes [XXIV], Key-posed [XIII], motion history [I] and action templates [XLIII]. However, these methods are very sensitive to quality of input action image, i.e., for a low quality image, these methods cannot describe the action shape effectively. Moreover, these approaches are also sensitive to illumination variations, different viewpoints, background clutter, and camera movements and had shown a poor performance in such circumstances. Next, the local feature extraction methods employed over a local Spatio-temporal location and these methods explode static or motion information. Some examples of this category are Hessian 3D descriptor [XV], Harris Corner Detector [VII], Gabor filter [XXXII], Space-time Interest points (STIPs) [XVI], and Histogram of Gradients [II].

Even though both of these methods have achieved a remarkable performance in the action recognition, they are computationally ensuring much burden. Instead of such complex techniques, some simple methods are there like Gaussian filtering which have less complexity. They can also capture the local and global appearance information and also the shape of body. Hence they can overcome of the computational burden. Based on this inspiration, in this paper, we have proposed to develop a composite feature extraction technique by integrating different and simple feature extraction methods. Towards such aspect, we have employed three different filters such Gaussian Filter, Gradient Filter and Gabor filter and extracted three different feature maps. And a Fused Feature vector (FFV) is formulated by
integrating these three feature maps. Further to achieve the dimensionality constraints, we have employed a Kernalized Discriminant Analysis (KDA). Finally for recognition, we have used Support Vector Machine (SVM) algorithm.

Remaining paper is organized as follows; Section II explores the Literature survey details. Section III explores the complete details of proposed action recognition framework. The details of simulation experiments are stipulated in section IV and the concluding remarks are stipulated in section V.

II. Literature Survey

From the past decades, several approaches have been developed, proposing a variety of methods for Human action recognition. As discussed earlier, among the two possible categories, the Local feature extraction techniques are more effective in recognition the human action; here we have done a brief literature survey over several local feature extraction techniques. Among the local feature extraction techniques, STIPs [VI], Gabor features, and Histogram of Gradients [XX], [XLIV] is the two methods which have gained a better performance in the action recognition under varying circumstances. Inspired with HOG, I.C. Dutta et al., [XII] proposed Histograms of Motion Gradients (HMG) based on the spatial derivation, which captures the changes between two consecutive action frames. Further for feature extraction, this work employed Shape Difference Vector of Locally Aggregated Descriptors (SD-VLAD) which brings the complementary information using the shape information. Further, Jin Wang et al., [XVIII] employed Pyramid Histogram of Oriented Gradient (PHOG) and two state-space models such as Hidden Markov Model (HMM) and Conditional Random Field (CRF) to characterize human figures for action recognition. Further, Bo Lin and Bin Fang [IV] proposed Spatio-Temporal Pyramid Histogram of Gradients (SPHOG) which is based on the gradient changes between successive frames. Further to incorporate the local information distribution into VLAD, a Gaussian Kernel is implanted to measure the weighted distance histograms of local descriptors. Further, combining the Distance mean histogram of gradients with segmented block of mean image with normalization for generating action descriptor. Random forest algorithm is employed for classification. Considering the gradient of motion,

V. Thanikachalam and K.K. Thyagarajan [XLII] proposed an action recognition based on Accumulated Motion Image (AMI) in which the histograms are built based on the energy distributions. After the evaluation of AMI, Discrete Fourier Transform (DFT) is employed and mean and variance are measured. Finally the Dynamic Time Wrapping (DTW) is employed for training. The Gradient descriptor is much effective in the extraction of edge structure or the appearance of an image because it is computed from the local distribution gradient. However, the gradient based algorithms are sensitive to noise. To overcome this problem, D.K. Vishwakarma and C. Dhiman [IX] proposed to extract the gradient features along with difference of Gaussian kernel [XXVIII] features. Spatial distribution gradients are computed at different levels of resolution, resulting different shape variations. Further these Difference of Gaussian features are combined with STIPs and created a
fused vector for every action class. Similarly, parularora et al., [XXXI] proposed new descriptor called Gradient Histogram Gaussian Image (GHGI) for action recognition. GHGI have much effectiveness in the reduction noise due to the incorporation of Gaussian function over HOG [XXX]. Nearest Neighbor classifier is employed for action classification.

Next, one more local descriptor, Gabor feature descriptor [XXV] and Log-Gabor features [XXIX] also gained significant performance in the action recognition. Based on this fact, S. Kanaganammigaand S. Vasuki [XXXV] employed Gabor filter and Optical flow features for movement estimation and object tracking. An Expectation Maximization based Effective Gaussian Mixture Model (EMEGMM) is accomplished for background subtraction and Adaboostalgorithmis employed for classification. Further, D. K.Vishwakarma et al. [X] combined the Gabor Wavelet Transform (GWT) with RidgeletTransform (RT) to recognize human action with scale, rotation and invariant features extraction. Here the GWT extracts the scale and rotation invariant features while the RT extracts orientation invariant features. Next, S. Maheswari and P. ArockiaJansiRani [XXXVI] combined the features obtained through Gabor filter and Haar filter to perform HAR. Before the feature extraction, they accomplished Adaboostfor frame splitting and mean shift algorithm us employed for object tracking. Gabor feature explores the orientation and frequency information while Haarfeature explores pixel variations and finally Relevance Vector Machine (RVM) is adopted for classification.

More number of feature extraction methods employed over an action image consequences to an excellent feature set but with larger size. The large size feature vector creates an extra computational burden for classification algorithm due to multiple time comparison. To solve this some, some authors tried to reduce the size of feature vector through standard dimensionality reduction algorithms like Independent component analysis (ICA) [XXVI], Principal Component Analysis (PCA) [XXIII, XVII], Linear Discriminant Analysis (LDA) [XXXVII] and their subsequent [XL], [XXVII]. Yuting et al., [XXXVII] employed LDA for open view action recognition through which a common discriminant subspace is obtained for every action class. However, the LDA achieve optimal space by projecting linearly separated instances which is a not practical scenario. Further some subsequent discriminant analysis method are proposed such as Robust Linear Discriminant Analysis (RLDA) [XXVII], Independent Component based LDA (IC-LDA) [XL], and Regularized Discriminant Analysis (RDA) [III]. However, all methods are assumed that the features are linearly related and tried to reduce the dimensionality by deriving only linear discrimination.

III. Proposed Framework

Overview

This section describes the details of proposed action recognition framework in detail. The architecture of proposed framework is shown in figure.1. Accordingly, the proposed framework is carried out in three phases. (1) Feature extraction, (2) Dimensionality reduction and (3) Classification. The main contribution of this paper...
is done in the feature extraction phase by developing three different feature extraction techniques and a new feature vector is constructed by integrating the features obtained at individual techniques. Next, at dimensionality reduction technique, we have focused to reduce the dimensions of feature vector, because it is very larger sized vector due to multiple features. Finally the obtained feature vector is fed to classification and at this phase, we have employed support vector machine algorithm to classify the actions.

**Feature Extraction**

In this phase, we have focused to employ three different feature extraction techniques such as Gaussian features, Gradient features and Gabor features. After the extraction of all these features through individual techniques, they are fused into a single feature vector called as Fused Feature Vector (FFV). This FFV is resilient to noises, edge discontinuities, scaling and rotations. In this approach, the Gaussian features make the recognition system resilient to noise; gradient features makes resilient to edge discontinuities and Gabor features makes resilient to scaling rotational variations. The detailed description about the feature extraction techniques is described in the following subsections.

1. Gaussian Features

   The main intention of Gaussian features is to make the recognition system robust to noises. For a given action image/frame contaminated with noise, the noise added part is also visualized as small edges and this creates an unnecessary confusion for the classifier. Moreover, the Gaussian features also helps in the enhancement of edges of an action boundary in the image. Actually the proposed Gaussian feature map extraction is inspired with Center Surround (CS) theory [VIII, XI] which have been identified as better edge enhancement techniques in human visual system. The CS theory has achieved better performance in edge enhancement that provides the detection, location, and tracking of small objects. In action images, due to noises at some instances, the boundaries will get visualized as discontinuities and will appear
like small objects. These small objects have different characteristics at different scales and they are highlighted only at particular scales. After CS operation, these features with different scales like edges, boundaries will get enhanced and aggregated into a separate sub band images.

With this inspiration, in this paper, we have proposed to extract the edge, and boundary features through Gaussian filter at different scales. For this purpose, initially we construct a seven level Gaussian pyramid for a given input action image. At first level, the Gaussian feature map is constructed by convolving the original action image with a Gaussian filter with variance \( \sigma = 2 \). Next, to obtain the Gaussian at second level, the original action image is down sampled and then convolved with same Gaussian filter. In this way, the Gaussian Pyramid is constructed up to seven levels by convolving the action image with different scales with the Gaussian filter. To construct Gaussian map, we have employed the following mathematical expression;

\[
G_L(i,j) = \sum_{x=1}^{X} \sum_{y=1}^{Y} f_{g}(x,y)G_{L-1}(2i+x, 2j+y)
\]

(1)

Where

\[
f_{g}(x,y) = \frac{1}{(\sqrt{2\pi} \sigma)^2} e^{-\frac{x^2+y^2}{2\sigma^2}}
\]

(2)

Here \( L \) denotes the level of Gaussian Pyramid, \( f_g(x,y) \) denotes Gaussian filter, and \( G_L(i,j) \) denotes the pixel at \( (i,j) \) position in the Gaussian Filtered image \( G_L \). Here \( X \) and \( Y \) denotes the size of Gaussian filter and in our work we have fixed these value to 5, i.e., \( X=Y=5 \).

Next, the final Gaussian feature map is constructed by the subtraction of Gaussian Features Maps at different levels in the Gaussian Pyramid. According to the CS theory, i.e., the subtraction is done between center levels (CL) and surrounding levels (SL). In our Gaussian pyramid, we have picked up the Gaussian feature map at fourth level as center level and surrounding levels are depicted as \( SL = CL \pm d \) (where \( d = 1,2 \)). Hence totally we have obtained four surrounding levels such as 4-1=3, 4+1=5, 4-2=2 and 4+2=6. Based on this, we have obtained four Gaussian feature maps by subtracting the Gaussian feature map at fourth level from the Gaussian feature maps at second, third, fifth and sixth as \( G_4 - G_2, G_4 - G_3, G_4 - G_5 \) and \( G_4 - G_6 \). As a result four levels of feature maps are calculated as Gaussian feature maps. An example of Gaussian Feature map construction is depicted in figure.2.

Note: during the subtraction process, the size of both Gaussian feature maps must be equal. Here the size of Gaussian feature maps at different levels is not equal. With an increase in the level of pyramid, the size of Gaussian feature map gets recued due to down sampling. At this case, the subtraction is not possible. Hence to make it possible, the Gaussian maps at levels 5 and 6 are interpolated to the size of Gaussian map at fourth level such that the subtraction \( G_4 - G_5 \) and \( G_4 - G_6 \) becomes feasible. Similarly, to accomplish the subtractions such as \( G_4 - G_2 \), and \( G_4 - G_3 \) possible the
Gaussian map at fourth level is interpolated to the size of Gaussian maps at level 2 and level 3.

2. Gradient Features

In HAR, gradient features have more importance because the gradient of an image explores the fine details such as edges and sharp discontinuities. The main intention of gradient features is to explore the action with respect to its direction of movement. Actually for a given action image, the pixel intensities vary with action movements and if such direction of movements are captured at the feature extraction phase, then the recognition system will become more effective. In this paper, for a pixel in the action image, to derive the direction of movements, we have considered its neighbor pixels in horizontal and vertical directions. And a difference between the current pixel and its neighbor pixels gives an information about the direction of movement. This flexibility can be gained through gradient operators. Laplacian of gradient is one of the most powerful and effective gradient operator which derives the fine coarse details from the image. This fine detail helps in the detection of sharp discontinuities in the action image boundary. Figure 3 shows the gradient features of an action image.

Fig.2: Gaussian feature map construction

With this inspiration we have adopted a one-dimensional Laplacian operator to capture the differences in an action image. There are two reasons behind this adoption of Laplacian operator for gradient features extraction. (1) Within an action image, the Laplacian operator can detect the fine details, highlight the edges, and also...
enhance the features with sharp discontinuities. (2) Laplacian is a second order derivative and hence it has a strong response towards the fine details than the first order derivatives, like gradient operator [XXXIII]. Due to these two reasons, the Laplacian operator over an action image can highlight regions of spontaneous changes in the pixel intensities and have it was used in several applications for blob and edge detection. For a given action image, the Laplacian operator will theoretically highlight the edges and boundaries.

Let’s consider an action image \( A \) of size \( M \times N \), where \( M \) is the row size and \( N \) is column size, initially it was reformulated into a 1-D signal and let it be \( A = \{a_1, a_2, \ldots , a_p\} \) where \( a_i \in \mathbb{R}^d \) is a pixel. First apply gradient operator \( \nabla \) on the action image, resulting in a first order gradient image, as

\[
G_I = \nabla A = \{G_{i1}, G_{i2}, G_{i3}, \ldots , G_{ip}\}
\] (3)

Where \( G_{ii} = \frac{dI}{dt} = a_i - a_{i-1} \). Since the given input is a 2-D image, the gradient operator is employed in both horizontal and vertical directions. Let the horizontal gradient is \( G_{iH} \) and vertical gradient is \( G_{iV} \), the overall gradient magnitude is computed as

\[
G_I = \sqrt{G_{iH}^2 + G_{iV}^2}
\] (4)

Next, apply the gradient operator \( \nabla \) on the gradient image \( G_I \), resulting in a second order gradient image \( G_{II} \), as

\[
G_{II} = \nabla G_I = \{G_{i1}, G_{i2}, G_{i3}, \ldots , G_{ip}\}
\] (5)

Where \( G_{ii} = \frac{dG_I}{dt} = G_{i1} - G_{i-1} \). Here \( G_I \) is the first order gradient, \( G_{i,i} \) is the \( i \)th gradient feature in the \( G_I \) and \( G_{i,i-1} \) is the \((i-1)\)th gradient feature of \( G_I \). Since the second order gradient is also a 2-D object, the gradient operator is employed in both horizontal and vertical directions. Let the horizontal gradient is \( G_{iIH} \) and vertical gradient is \( G_{iIV} \), the overall gradient magnitude is computed as

\[
G_{II} = \sqrt{G_{iIH}^2 + G_{iIV}^2}
\] (6)

The resultant \( G_{II} \) is a second order derivative of an action image \( A \). This is more helpful in the provision of sufficient discrimination between different actions. For example in the horizontal hand waving action, the movements are along horizontal direction and the gradient of such type of action highlights the edges along horizontal direction only. In such case the horizontal gradients such as \( G_{iH} \) and \( G_{iIH} \) have higher magnitudes compare the vertical gradients \( G_{iV} \)and\( G_{iIV} \). Similarly, for another hand waving action (upwards) in the KTH dataset, the movements are along vertical direction. In such case the vertical gradients such as \( G_{iV} \) and \( G_{iIV} \) have higher magnitudes compare the vertical gradients \( G_{iIH} \)and\( G_{iIV} \). Furthermore, the boundaries with sharp discontinuities are also enhanced giving a more clarity whether it is belongs to external edge or a part of action boundary.
3. Gabor Features

The main intention of Gabor features extraction is to make the recognition system resilient to scale and rotational variations. Similar to our first contribution [XXI], to extract the Gabor features, this paper also employed Gabor filter in multiple orientations and multiple scales. For a given action image, Gabor filter is employed at totally eight ordinations such as $0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$, $180^\circ$, $225^\circ$, $270^\circ$, and $315^\circ$ and at different scales such as $5 \times 5$, $7 \times 7$, $9 \times 9$, and $11 \times 11$. Hence totally we will get 32 feature maps, which are too many in number. From these 32 maps, we extract only eight significant feature maps. These eight maps are obtained through max pooling; means one feature map for one orientation. The obtained eight feature maps are scale invariant.

![Fig.3: (a) Original Action image (b) Laplacian gradient of an action image](image)

4. Fused Feature Vector (FFV)

Once the features are extracted from individual feature extraction techniques such as Gaussian Filter, Gradient Filter and Gabor filter, then they are fused and formulated as a single feature vector and such vector is called as FFV. This FFV consists of intensity, gradient, and Gabor features and hence the action recognition system is more effective and recognized any type of action given as input. To fuse the features one by one, we have employed vertical concatenation process and the FFV is a 1-D vector. The process of fusion is depicted in figure 4.

Dimensionality Reduction

The dimensionality reduction is applied over the FFV to reduce its dimensions. Since the proposed feature extraction mechanism is composed of three different feature extraction techniques, the size of FFV is too large, resulting in a heavy computational burden for classifier. Hence this paper tends to reduce the dimensionality of FFV. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the two most popular dimensionality reduction techniques. PCA is an unsupervised and LDA is supervised methods. In these two methods, LDA have better performance compared to PCA because the Principal components obtained through PCA have high variance which won’t give effective
results in the recognition of actions, especially when the actions have similar trajectories like running and jogging.

In the case of supervised algorithms, LDA has perceived an excellent performance in the action recognition. In LDA, the optimal subspace is obtained by the optimization of fisher-rao’s criterion which is defined as the ratio of within class scatter matrix to the between class scatter matrix. Mathematically the optimal subspace is defined as;

\[ J(W) = \arg \min_W Tr(W^T S_W W) / Tr(W^T S_B W) \]  \hspace{1cm} (7)

Where \( S_W \) is a within class scatter matrix and \( S_B \) is a between class scatter matrix, that are both are symmetric and positive definite matrices. The mathematical expressions for \( S_W \) and \( S_B \) are defined as;

\[ S_W = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_{ij} - m_i)(x_{ij} - m_i)' \]  \hspace{1cm} (8)

And

\[ S_B = \sum_{i=1}^{K} (m_i - m)(m_i - m)' \]  \hspace{1cm} (9)

Where \( N \) denotes the total number of samples in class \( i \), \( m_i \) is the mean of data in class \( i \), \( m \) is the mean of total class data and \( x_{ij} \) is the jth sample of class \( i \).

LDA tries to maximize the separation between different classes and minimize the separation within the class simultaneously. However, LDA captures the linear spaced features only but not focused on the non-linear spaced features. Kernel Dimensionality Analysis (KDA) is a non-linear extraction of LDA which was used in this paper to obtain non-linear discriminant features through kernel techniques [V]. In KDA, the input data is mapped to the low dimensional feature space by non-linear mapping. In KDA, the within and between class scatter matrices are defined as;

\[ J(W) = \arg \min_W Tr(W^T S_W W) / Tr(W^T S_B W) \]  \hspace{1cm} (7)
\[ S_{\text{w}} = \sum_{i=1}^{C} \sum_{x \in X} (\phi(x) - \delta_i)(\phi(x) - \delta_i)' \]  

And

\[ S_{B} = \sum_{i=1}^{K} N_i(\delta_i - \delta)(\delta_i - \delta)' \]  

Where \( N_i \) is the number of samples in the action class \( i \), \( \delta_i \) is the centroid of class \( i \), \( \delta \) is the global centroid, \( C \) is the number of classes, \( x \) is a vector of specific class, and \( X_i \) is the set of samples in the class \( i \). In Eq.(10) \( S_{W} \) determines the scattering degree within the class of actions and is measured as the summation of covariance matrices of each class. Next, In the Eq.(11), \( S_{B} \) determines the scattering degree between the class of actions and is measured as the summation of covariance matrices of means of each class. Finally the optimal subspace is obtained as;

\[ J(W) = \arg\min_{W} \left( \frac{\text{Tr}(W^T S_{W} W)}{\text{Tr}(W^T S_{B} W)} \right) \]  

The major difference between LDA and KDA is the computation process of scattering matrices. In the LDA, the scattering matrix is measured through the computation of mean deviations. For within class, the class samples are discriminated by measuring the deviation of sample with the mean and in between class, the discrimination is computed by measuring the deviation of mean from overall mean of data. Unlike LDA, in KDA, the discrimination is computed based on centroids. For within class discrimination, initially one centroid is chosen and the samples in that class are discriminated by measuring the deviation of samples with centroid of that particular class. Next, for between classes, the discrimination is measured by computing the deviation of a centroid of particular class with overall centroid. This evaluation has one man advantage, i.e., it can measured the samples which are non-linearly related and this is the most realistic scenario in real time applications, because, all the action are not linearly related.

IV. Simulation Results

To evaluate the performance of developed HAR system, we used three different and standard benchmark datasets that include KTH dataset, UCF YouTube action dataset and Weizmann Dataset. The simulation is accomplished through MATLAB software. Initially, we have discussed the details of datasets and then the results obtained after the deployment of proposed approach over hem is discussed. Further a detailed comparative analysis is stipulated between proposed and conventional approaches.

Datasets

1. 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 [XXIV]. 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 $180 \times 144$. Some examples frames of this dataset are shown in figure.5.

2. UCF YouTube Action Dataset

This dataset consists of totally 11 actions such as horseback riding, golf swinging, diving, biking/cycling, basketball shooting, swinging, soccer juggling, trampoline jumping, tennis swinging, walking with dog, and volleyball spiking [XIX]. This is a very challenging dataset because it has so many variations in illuminations, cluttered background, viewpoint, object scale, pose, appearance and camera motion. For every action category, the videos are grouped into 25 groups and every group is composed of more than four clips. All the videos are in mpeg4 format. Some examples frames of this dataset are shown in figure.6.

3. KTH Dataset

This dataset consists of totally six different actions such as Walking, Jogging, Running, Boxing, Hand waving, and Hand clapping [VII]. The actions of this dataset are captured under four different environments such as Outdoors, Indoors, Outdoors with several scales and with several clothes. Totally the actions are captured with the help of 25 subjects and the total number of videos present in this dataset is 600. All the videos are in AVI format. The frame rate of camera used to capture the actions present in this dataset is 25 frames per second and are captured under homogenous background. Some examples frames of this dataset are shown in figure.7.

![Figure 5: Weizmann samples](image5)

![Figure 6: UCF YouTube Action dataset samples](image6)

Copyright reserved © J. Mech. Cont. & Math. Sci.
K Ruben Raju et al
Results

To measure the performance of developed HAR system, this paper considered several performance metrics like Detection Rate or Recall, Precision, F-Score and Accuracy. Under the simulation, we have created three different sets for training such as \( S_1 \), \( S_1 + S_2 \) and \( S_1 + S_2 + S_3 \). Under \( S_1 \), we have trained the system with only one set of actions (for each action only one actor is considered). Next, for \( S_1 + S_2 \), we have trained the system with two sets of actions, i.e., for every action we have considered two different actions and they are trained. Similarly, for \( S_1 + S_2 + S_3 \), we have trained the system with three sets of actions, i.e., for every action we have considered three different actions and they are trained. At every case, we have tested the proposed HAR system with different sets of features. The obtained results are shown in the following figures.

![Fig.7: KTH samples, (a) Walking, (b) Jogging, (c) Running, (d) Boxing, (e) Handwaving, (f) Handclapping](image)

**Fig.7:** KTH samples, (a) Walking, (b) Jogging, (c) Running, (d) Boxing, (e) Handwaving, (f) Handclapping

![Fig.8: Detection rate of Weizmann Dataset](image)

**Fig.8:** Detection rate of Weizmann Dataset

Figure 8 shows the details of Detection Rate (DR) measured after the simulation of proposed approach over weizmann dataset. Under this, the simulation is done totally in three different environments. As specified in the above figure 8, the HAR is employed with only one set of features (either Ga (Gaussian), Gb (Gabor) or Gr (Gradient)). In Ga+SVM, we have employed only Gaussian pyramid features, in Gb+SVM, we have employed only Gabor features and in Gr+SVM, we have
employed only gradient features, at both training and testing phases. Finally in FFV, SVM, we have employed Proposed Fused Feature Vector as a feature descriptor and different actions were tested from Weizmann dataset. As it can seen the detection rate is high for FFV+ SVM compared to the remaining techniques. This is due to the provision of sufficient knowledge to the system about all possible variations such as Noise, Edge discontinuities, Scale and rotations. Next, viewing the DR in the point of Training sets, the maximum DR is achieved for \{S_1 + S_2 + S_3\} because more actions are trained under this set. Due to this training, the system will gain more knowledge about the actions. For example, an action can be performed in different ways by different actors. Due to this advantage, the recognition system can recognize the action even though the input action image is having several side effects like scalings and rotations. The DR shown in the above figure is average DR, as it is measured by averaging the DRs obtained for individual actions.

The results shown in figure 9 and figure 10 are obtained after the simulation or proposed approach over KTH dataset and UCF YouTube action dataset respectively. As it can be seen from these two figures, in both simulation cases, the FFV has gained more detection rate compared to the remaining techniques. Next, the Detection rate obtained is high for KTH dataset compared to the UCF Dataset. Because, compared to the KTH dataset, the UCF is more challenging dataset and it consists of action videos with so many challenges. Moreover, the action images in the UCF dataset are composed of some more moving objects (ex. Dog in walk with dog and horse in horse riding). In such type of actions, the extra included objects are also moving which creates much confusion to the recognition system. Whereas in the KTH dataset, the videos are composed of a homogeneous background with only Human action. Further we can observe that the DR at S1 training set is very less compared to the remaining training sets. For instance for UCF dataset, the DR obtained through FFV+SSM at S1 is approximately 72.5123% whereas for the same case, at KTH and Weizmann datasets it is of 78.5423% and 79.8653% respectively. The deviation is approximately 6% and this deviation is reduced with an increase of action videos in training sets such as \{S_1 + S_2\} and \{S_1 + S_2 + S_3\}. For \{S_1 + S_2\}, the average DR of Weizmann, KTH and UCF is observed as 90.0124%, 88.4512% and 85.8856% respectively. At this case, the average deviation is noted as 2%. This is further reduced at \{S_1 + S_2 + S_3\}, as it is of only 1%. Based on this analysis, we can understand that as the number of different features trained to the database are increased, the recognition system will gain more knowledge and hence it can recognize any type of action given as input.
Comparative Analysis

To alleviate the performance or proposed action recognition framework, we have compared the performance of proposed FFV + SVM with several earlier studies like Gabor + Haar [XXXVI], HoG + SVM [XXX] and Gabor + RT [X]. In the first conventional method, i.e., Gabor + Haar, the action recognition is done based on the Gabor features and Haar features. For a given input image, this method employed Gabor filter and Haar filter for feature extraction and they are fed to classifier and RVM is employed for classification. In this method, the Gabor filter is accomplished to extract the orientation invariant features while the Haar filter is employed to
discover the pixel variations. Even though this approach have used Gabor filter which has a great significance in the action recognition, they didn’t focused over the action boundaries with sharp discontinuities which is more important in the dataset like UCF.

Next, the Wavelet transform can solve this problem by discovering the edge feature as high frequency bands.

Based on this aspect, Gabor filter is combined with Ridgelet Transform and proposed an action recognition system. However, due to the similar structural aspects between wavelet and Ridgelet, it is susceptible to shift invariance. Next, compared to the Gabor filter and Wavelet filter, the Histogram of Gradients have gained more recognition performance in HAR. Based on this P.A. Dhulekar, S. T. Gandhe [XXX] applied only HOG for action recognition. But the main problem with HoG is it can’t provide the information about the scaling and rotational variations if present in the action image. For a given input image with some tilt or scaled, the HoG features will change and they don’t match with the features in the dataset accurately. This is more specific in some actions like Running and Jogging which have similar motion trajectories. Hence, the feature descriptor of an action must in such a way that it should provide all possible variations those tend to happen in the action image, with less feature space. To do this, we have developed a composite feature extraction technique which has integrated three different feature extraction techniques followed by an effective dimensionality reduction technique. The proposed FFV has more significance in the derivation of an effective features and the KDA has effectiveness in the reduction of dimensionality.

The comparison is done with respect to four performance metrics such as Detection Rate, Precision, F-Score and Recognition Accuracy. These metrics are computed after the deployment of proposed as well as earlier approaches over three datasets one-by-one. The obtained performance metrics are depicted in the following figures.

![Detection rate vs. datasets](image.png)

**Fig.11:** Detection rate vs. datasets
Recall or DR is measured as the ratio of total number of actions classified correctly (TP) to the sum of TP and False Negative results (FN). As the total number of correctly classified results increases, DR also increases. Figure 11 shows the comparison between proposed and conventional approaches through detection rate for three datasets such as KTH, Weizmann and UCF. From this figure, we can observe that the higher detection rate is for Weizmann dataset and low DR is for UCF dataset. This is due to the presence of several complex constraints such as extra moving objects, illumination, camera movements etc., in UCF dataset. To achieve a better DR even under such type of constraints, the feature extraction technique must be more effective such that the HAR system can recognize any type of action given as input. The proposed FFV has such capability and hence the DR of proposed FFV+SVM is high compared to the conventional approaches. The FFV can cover all the advantages of conventional approaches and hence it has achieved better DR and on an average it is approximated as 94.4033% whereas as it is of 81.5400%, 85.0023%, 90.0255% for Gabor + Haar, HoG + SVM and Gabor + RT respectively.

![Fig.12: Precision vs. datasets](image)

Precision is measured as the ratio of total number of correctly classified results to the sum of total number of correctly classified results and the total number of false positive results (FP). Precision determines the recognition system’s precise performance, i.e., for a total number given input actions, how many actions are precisely classified (excluding the results those are classified as given input action class). Figure 12 shows the comparison between proposed and conventional approaches through precision for three datasets such as KTH, Weizmann and UCF. From this figure, we can observe that the higher precision is for Weizmann dataset and low precision is for UCF dataset. In UCF dataset, some actions are there like Tennis Swing and Golf swinging which have similar motion trajectories, makes the recognition system more confused. This confusion results in more FPs, resulting in less precision. To overcome such constraints, the feature descriptor should be able to track the motion trajectories more effectively even for the actions with similar

*Copyright reserved © J. Mech. Cont.& Math. Sci.*

*K Ruben Raju et al*
trajectories. This is achieved in the proposed FFV+SVM due to the proposed Gradient filter with Laplacian kernel and hence the proposed approach has higher precision for all datasets. On an average the precision is approximated as 94.1633% whereas as it is of 85.4067%, 86.8867%, 91.0034% for Gabor + Haar, HoG + SVM and Gabor + RT respectively. For, UCF dataset, we can observe that the precision of Gabor + Haar is high compared to the HoG+SVM, while it is low for other datasets. This is due to the fact that the only HoG feature descriptor cannot ensure a perfect discrimination between different actions, resulting a higher FPs.

F-Score the harmonic mean of DR and precision. For higher values of DR and precision, the F-score also high. Figure 12 shows the comparison between proposed and conventional approaches through F-Score for three datasets such as KTH, Weizmann and UCF. From this figure, we can observe that the higher precision is for Weizmann dataset and low precision is for UCF dataset. On an average the F-Score is approximated as 94.2748% whereas as it is of 83.4272%, 85.9152%, 90.5064% for Gabor + Haar, HoG + SVM and Gabor + RT respectively.

![F-Score vs. datasets](image1)

![False Negative Rate vs. datasets](image2)

**Fig.13:** F-Score vs. datasets

**Fig.14:** False Negative Rate vs. datasets
FNR is simply related to the DR as $\text{FNR} = 100 - \text{DR}$. Simply FNR can be measured with this formula or it can also be measured as the ratio of total number of FNs to the sum of TPs and FNs. A higher FNR defines poor performance and vice versa. Figure 14 shows the comparison between proposed and conventional approaches through FNR for three datasets. From this figure, we can observe that the higher FNR is for UCF dataset and lower FNR is for Weizmann dataset. Further we can also observe that the proposed FFV+SVM have gained a less FNR while it is more for Gabor + Haar, for all datasets. On an average the FNR is approximated as 5.5967% whereas as it is of 18.4623%, 15.2031%, 9.9767% for Gabor + Haar, HoG + SVM and Gabor + RT respectively. The main reason behind this achievement is the accomplishment of composite feature descriptor and KDA. In this regard, the KDA also is much helpful in the provision of a perfect discrimination between different actions and hence the proposed approach have less FNR.

FDR is related to precision and they are related as $\text{FDR} = 100 - \text{Precision}$. For a higher value of precision, the FDR is less and vice versa. The FDR is also measured as ratio of FPs to the sum of TPs and FPs. For a given input action images, the FDR defines the total percentage of action image those are falsely discovered. For any recognition system, the higher value of FDR defines its poor performance. As shown in the figure 15, the lower FDR is attained for Weizmann dataset and higher FDR is for UCF dataset due to the complex nature of actions videos in UCF. Even though, the proposed FFV+SVM has gained less FDR compared to the conventional approaches. On an average, it is approximated as 5.8345% whereas as it is of 14.6933%, 13.1133%, 9.0047% for Gabor + Haar, HoG + SVM and Gabor + RT respectively.

Figure 16 shows the comparison between proposed and conventional approaches through Recognition accuracy for all the three datasets. From this figure, we can notice that the proposed FFV+SVM have gained higher recognition accuracy compared to the conventional approaches. The main reason behind this success is the deployment of integrated feature descriptor which is composed of several features.

Fig.15: False Discovery Rate vs. datasets

Fig.16: Recognition accuracy vs. datasets

Copyright reserved © J. Mech. Cont.& Math. Sci.
K Ruben Raju et al
those alleviate different side effects such as illumination, scaling, rotation, and discontinuities in the boundaries of motion trajectories.

Fig.16: Average Recognition Accuracy vs. datasets

Whereas in the conventional approaches only one set of feature are considered by which the system cannot recognize if an action with different variation is given as input. For instance, to the Gabor + Haar system, if the action ‘walk with dog’ is given as input, then it can’t recognize because, the discontinuity in the motion trajectory between human and dog is not learned. This flexibility is ensured in the proposed FFV+SVM due to the consideration of gradient features. Along with this flexibility, the proposed KDA model also provided a more discriminant subspace for every class of actions, resulting in an improved recognition performance at individual actions. Hence the proposed approach has gained higher recognition accuracy and on an average it is approximated as 97.8267%, while it is of 91.3323%, 93.2433% and 95.7412% for conventional approaches Gabor + Haar, HoG + SVM and Gabor + RT respectively.

V. Conclusion

In this paper, such a recognition approach for the recognition of human action based on Fused Feature Vector and Kernel Discriminant Analysis is proposed. Though the proposed model resembles with some previous developed models of action recognition, it differs substantially in some important aspects resulting in a significant improved performance. Most importantly, unlike to the conventional Gaussian models, the Difference of Gaussian in this model is obtained through Gaussian pyramid construction at different scale and made the system to effectively track the motion object at different scales. Next, the proposed Gradient filter based on Laplacian kernel made the system resilient to discontinuities in action boundaries. Further the Gabor filter made the system resilient to scale and rotational variations. Finally the KDA has achieved less computational burden followed by an improved
recognition performance. Simulation experiments conducted over three public and benchmark datasets revealed the effectiveness and outperformed the conventional approaches.

References

I. A. Bobick and J. Davis, “The recognition of human movement using temporal templates,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

II. A. Klaser and M. Marszalek, “A Spatio-temporal descriptor based on 3dgradients,” in *Proc. Eur. Conf. Comput. Vision*, 2008.

III. B. Mandal, How-Lung Eng, “Regularized Discriminant Analysis for Holistic Human Activity Recognition”, *IEEE intelligent systems*, 2012

IV. Bo Lin and Bin Fang, “A new spatial-temporal histograms of gradients descriptor and HOD-VLAD encoding for human action recognition”, *International Journal of Wavelets, Multi-resolution and Information Processing*, Vol. 17, No. 02, 2019.

V. B. Sch¨olkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT press, 2002.

VI. Chakraborty, B.; Holte, M.B.; Moeslund, T.B.; Gonzalez, J. Selective Spatio-temporal interest points. *Comput. Vis. Image Underst*. 2012, 116, 396–4100

VII. C Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, 2004, pp. 32–36.

VIII. C.Yang, M. Schmalz, W. Hu, and G. Ritter, “Center-surround filters for the detection of small targets in cluttered multispectral imagery: Background and filter design,” in *Proc. SPIE*, 1995, vol. 2496, pp. 637–648

IX. D.K. Vishwakarma and C. Dhiman, “A unified model for human activity recognition using spatial distribution of gradient and difference of Gaussian kernel”, *Vis Comput*. 35, 1595-1613, 2019.

X. 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.

XI. D. Song and D. Tao, “Biologically inspired feature manifold for scene classification,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.
XII. Duta. I. C, Uijlings, J. R, Ionescu B, et al. Efficient Human Action recognition using Histograms of motion gradients and VLAD with descriptor shape information. Multimed Tools Appl. 76, 22445-22475, 2017.

XIII. D. Weinland and E. Boyer, “Action recognition using exemplar-based embedding,” in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2008, pp. 1–7.

XIV. G.Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, “Advances in human action recognition: A survey,” New J. Phys., vol. 17, no. 8, pp. 1_30, 2015.

XV. G.Willems, T. Tuytelaars, and L. Van Gool, “An efficient dense and scale-invariant Spatio-temporal interest point detector,” in Proc. Eur. Conf. Comput. Vision, 2008, pp. 650–663.

XVI. I Laptev and T. Lindeberg, “Space-time interest points,” in Proc. IEEE Int. Conf. Comput. Vision, 2003, pp. 432–439.

XVII. J. Arunnehru1 and M. KalaiselviGeetha, “Motion Intensity Code for Action Recognition in Video Using PCA and SVM”, In: prasath R., Kathirvalavakumar T. (eds) Mining Intelligence and knowledge Exploration Lecture notes in computer science, Vol.8284, Springer, Cham.

XVIII. Jin Wang et al. “Human action recognition based on Pyramid Histogram of Oriented Gradients”, IEEE International Conference on Systems, Man, and Cybernetics, AK, USA, 2011.

XIX. J. Liu, J. Luo and M. Shah, Recognizing realistic actions from videos "in the wild", CVPR 2009, Miami, FL.

XX. Kishore K. Reddy, NareshCuntoor, AmithaPerera, Anthony Hoogs, “Human Action Recognition in Large-Scale Datasets Using Histogram of Spatiotemporal Gradients”, IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, Beijing China, 2012.

XXI. K. Ruben Raju, Yogesh Kumar Sharma, BirruDevender, “Composite Feature Vector Assisted Human Action Recognition through Supervised Learning”, International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-6, March 2020.

XXII. K. Soomro and A. R. Zamir, “Action recognition in realistic sports videos," in Advances in Computer Vision and Pattern Recognition, vol. 71. Cham, Switzerland: Springer, 2014, pp. 181_208.

XXIII. M. B. Abidine, B. Fergani, “Evaluating a new classification method using PCA to human activity recognition”, ICCMA 2013: IEEEInternational Conference on Computer Medical Applications, At Sousse, Tunisia.
XXIV. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 2, 2005, pp. 1395–1402.

XXV. M. Bregenzio, S. Gong, and T. Xiang, Recognizing action as clouds of space-time interest points, *in CVPR*, 2009.

XXVI. Md. Zia Uddin, J.J. Lee, and T.-S. Kim, “Shape-Based Human Activity Recognition Using Independent Component Analysis and Hidden Markov Model”, In: Nguyen N. T., Borzemski L., Grzech A., Ali M. (eds) New frontiers in applied artificial intelligence. IEA/AIE 2008. Lecture notes I computer science, vol. 5027, Springer, Berlin, Heidelberg.

XXVII. M. Guo and Z. Wang, “A feature extraction method for human action recognition using body-worn inertial sensors”, IEEE 19th *International Conference on Computer Supported Cooperative Work in Design* (CSCWD), Itlay, 2015.

XXVIII. Michael W. Davidson, Mortimer Abramowitz, “Molecular Expressions Microscopy Primer: Digital Image Processing – Difference of Gaussians Edge Enhancement Algorithm”, *Olympus America Inc.*, and *Florida State University*.

XXIX. Ning Li, De Xu, “2D Log-Gabor wavelet based action recognition”, *IEICE Trans. Inf. & Sys.* Vol.E92-D, No.11, 2009.

XXX. P A. Dhulekar, S. T. Gandhe, “Action recognition based on Histogram of Oriented gradients and Spatio-temporal interest points, *IJET*, 7(4), 2018, 2153–2160.

XXXI. Parul Arora, Smruti Srivastava, Kunal Arora, Shreya Bareja, “Improved Gait Recognition using Gradient Histogram Gaussian Image”, Procedia Computer Science 58 (2015) 408 – 413.

XXXII. P. Doll’ar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proc. Joint IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 65–72.

XXXIII. R. Gonzalez and R. Woods, *Digital image processing*. Pearson/Prentice Hall, 2008.

XXXIV. R. Poppe, “A survey on vision-based human action recognition”, Image and Vision Computing 28 (2010) 976–990.

XXXV. 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.
XXXVI. S. Maheswari and P. ArockiaJansi Rani, “RVM-based human action classification through Gabor and Haar feature extraction”, *Int. J. Computational Vision and Robotics*, Vol. 6, Nos. 1/2, 2016.

XXXVII. Su, Y., Li, Y. & Liu, A., “open-view human action recognition based on Linear Discriminant Analysis”, Multimedia tools Appl, 78, 767-782, 2019

XXXVIII. T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no.2-3, pp. 90–126, 2006.

XXXIX. T. Ko, “A survey on behavior analysis in video surveillance for homeland security applications,” in *Proc. 37th IEEE Appl. Imagery Pattern Recog. Workshop*, Washington, DC, 2008, pp. 1–8.

XL. Uddin M, Lee JJ, Kim T.S., “Independent component feature-based human activity recognition via Linear Discriminant Analysis and Hidden Markov Model.”, In: *ConfProc of IEEE Eng Med Biol Soc*. 2008;2008:5168-71.

XLI. VikasTripathi, DurgaPrasadGangodkar, AnkushMittal, VishnuKanth, “Robust Action Recognition framework using Segmented Block and Distance Mean Histogram of Gradients Approach”, Procedia Computer Science, Volume 115, 2017, Pages 493-500

XLII. V. Thanikachalam and K.K. Thyagarajan, “Human Action Recognition using Accumulated motion and gradient of motion from video”, ICCCNT 2012.

XLIII. Y. Ke, R. Sukthankar, and M. Hebert, “Event detection in crowded videos,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2007, pp. 1–8.

XLIV. Yuan Shen, Zhenjiang Miao, “Oriented Gradients for Human Action Recognition”, *ICIMCS’10*, December 30–31, 2010, Harbin, China.