International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

View-Independent Discriminant Analysis with Gradient Self-Similarities for Action Recognition under Multiple Views

K Pradeep Reddy, G Apparao Naidu, B Vishnu Vardhan

Abstract*: In Multi-View Human Action Recognition, the actions are not of single view and hence to achieve an effective recognition performance under multi-view actions, there is a need of multi-view subclass discrimination analysis. Based on this inspiration, this paper proposed a novel action recognition framework based on the Subclass Discriminant Analysis (SDA), an extended version of Linear Discriminant Analysis (LDA). Further, a new key frames selection method is proposed based on Self-Similarity Matrix (SSM), called as Gradient SSM (GSSM). Once the key frames are selected, a composite feature set is extracted through three different set filters such as Gaussian Filter, Gabor filter and Wavelet Filter. Next, the SDA obtain an optimal feature subspace for every action under multiple Views. Finally the SVM algorithm classifies the action. The proposed framework is systematically evaluated on IXMAS dataset and NIXMAS dataset. Experimental results enumerate that our method outperforms the conventional techniques in terms of recognition accuracy.

***Keywords:** Multi-View Human Action Recognition, Self-similarity matrix, Gradients, Gabor filter, Discriminant Analysis, IXMAS dataset, Accuracy*

I. INTRODUCTION

From the past few years, the computer vision technology has gained a rapid development [1]. As an important and critical technology for the analysis and understanding of huge heterogeneous video data, the action recognition owns significant academic value, potential business value and also huge application prospect. Action Recognition has widespread applicability in various applications including Robotics, Human-computer Interactions field, e.g., gaming entertainment, robot navigation, video retrieval, intelligent traffic, and intelligent surveillance [2].For example, detection of falling actions of older aged people is very important and it is carried out through smart home system. Further, the detection of abnormal behavior of humans in time is a most important in intelligent video surveillance systems. In general, major of the earlier developed Human Action Recognition (HAR) techniques considered the single view of human action for recognition [3]. In these methods, the test video is of only a single sided view. However, the real world videos have so many challenges towards single view HAR, because, the visual appearance of actions are greatly affected by self-occlusion and view point changes. Hence there is a need to consider multiple views for an HAR, which is called as Multi-View HAR (MVHAR) [4].

In MVHAR, the recognition of human action is accomplished based on the multiple views. Every action is acquired with multiple cameras, representing each camera as one view. Since the self-occlusion can be handled more effectively through MVHAR, the multi-view frameworks can achieve more robust action recognition than the single view action recognition frameworks. In MVHAR, the action recognition is generally accomplished through the motion trajectories of camera viewpoints; the changes in the view point have significant effect in the understanding of action. Hence extraction of view invariant features is of most important. Temporal Self-Similarity Matrices (SSMs) [5] is one possible methodology to extract view invariant features. However, the SSMs are sensitive to the changes in the large view-point related appearances.

Furthermore, feature extraction is one more important step in HAR, which is accomplished filtering followed by dimensionality reduction. Principal Component Analysis (PCA) [6] and Linear Discriminant Analysis (LDA) [7, 8] are the two most significant dimensionality reduction techniques. In action recognition, compared to the features extracted through PCA, the LDA features have higher recognition performance due to the availability of class label. However, the assumption of class uni-modality in LDA limits its performance in the problems where classes form subclasses, i.e., classes represented by multiple disjoint classes, like an action class represented through multiple views. In MVHAR, action is considered as a class and its representations under multiple views are considered as subclasses. To achieve an efficient action recognition performance under multiple view-points, this paper proposed a novel HAR framework. In this framework, to extract key frames of action under multiple views, a new variant of SSM, called Gradient SSM (GSSM)is accomplished.

Manuscript published on January 30, 2020.
* Correspondence Author
K Pradeep Reddy*, Research Scholar, Dept. of Computer Science Engineering, Tirumala Engineering College, Hyderabad, Telangana, India.
Dr. G Apparao Naidu, Professor, Dept. of Computer Science Engineering, JB Institute of Engineering and Technology, Hyderabad, Telangana, India.
Dr. B Vishnu Vardhan, Professor, Dept. of Computer Science Engineering, Jawahar Lal Technological University College of Engineering, Manthani, Telangana, India.

Further to extract view invariant features of an action, this paper extracted three different features such as intensity features, orientational features, and contour features. Finally, a subclass information integrated inLDA, called as Subclass Discriminant Analysis (SDA) is accomplished to reduce the dimensionality of obtained feature vector. Simulation is conducted over two standard benchmark datasets, namely IXMAS and NIXMAS to test the performance of proposed framework.

Rest of the paper is organized as follows; Section II reveals the details of literature survey. Section III reveals the details SDA. Section IV discusses the details of proposed action recognition framework. Section V discusses the details of simulation results and finally the concluding remarks are discussed in section VI.

II. LITARATURE SURVEY

In this section, we review the related work on multi-view action recognition methods and Dimensionality Reduction methods.

A. Multi-view Action Recognition

Multi-view action recognition has gained much research interest recently; since a multi-view can overcome the self-occlusions problem and also can able to handle more robust recognition accuracy compared to the single view methods. For example, in the Action recognition method proposed by Weinland et al. [9], the action is represented through the 3D occupancy grids extracted from multiple view-points using an exemplar based Hidden Markov Model (HMM). Yen et al. [10] accomplished a 4D feature vector to recognize actions from arbitrary views. This feature vector encodes the motion and shape of actors under multiple views and constructs a 3D visual hull at every instant. However, the computational complexity of these methods is high, since discovery of a best match by searching a 3D model in 2D observations requires larger model parameter space. Next, based on the 3D Histogram of Oriented Gradients (HOG), Weinland et al. [11] developed a hierarchical action classification model. The robustness is achieved through the integration information from all views and trained the hierarchical classifiers. However, the integration of view point information at the training phase consequences to more computational burden at searching in classification.

A successful method to solve this problem of view-point related appearance differences on action recognition in 2D methods involves the design of view-invariant features. Temporal Self-Similarity Matrices (SSMs) is one possible methodology to extract view invariant features. Imran Junejo et al. [5] introduced the concept of SSM. Every action is represented with a set of SSMs. In this approach, initially, the action video is represented with low level features. Further the SSM is constructed based on the evaluation of Euclidean distance between the extracted features of all frames in a pairwise fashion. E. Shechtman and M. Irani [12] introduced a new image / video matching technique based on the internal self-similarities. This approach assumes that the images / videos are similar in their internal layouts even though the patterns generating those similarities are different. These internal self-similarities are effectively captured by Local Self-Similarity (LSS) descriptor. The LSS is applied densely throughout the video at multiple scales and extracted the matching entities even under various geometric distortions. Further, one more method is proposed by Stark et al. [13] in which the shape is considered as a local object and the similarity is accomplished between the shapes. Some more authors focused to combine the Bag of Visual Words (BoVW) with LSS and introduced a new similarity metric called as, bag-of-local-self-similarities (BOLSS) [14, 15].Inspired with the LSS, I. Junejo et al. [16] applied the local self-similarity surfaces for action recognition. These surfaces are constructed by performing the matching between patches, centered at a pixel. Once the surfaces are computed, it was proposed to transform these surfaces into HoG and then processed for training through Conditional Random Fields (CRFS). However, the LSS is never capture the global similarities in the entire image by which the image matching will be less effective.

Global Self-Similarity (GSS) is a new variant of self-similarity and can acquire the overall similarities within in the image. With GSS inspiration, T. Deselaers and V. Ferrari [17] introduced two different global descriptors, namely, self-similarity hyper cubes (SSH) and bag-of-correlation surfaces (BOCS). Jing Wang et al. [18] proposed a new HAR method based on SSM and Dynamic Time Warping (DTW). Here the SSM captured the Global Time information which was useful in the action recognition under viewpoints. DTW is applied for the full pledged utilization of SSM information. K-Nearest Neighbor Classifier (KNNC) is accomplished for classifications.

B. Dimensionality Reduction

Dimensionality reduction has an important contribution in the machine learning filed. A main advantage with dimensionality reduction is less computational complexity and computational time. PCA is the most popular technique for dimensionality reduction. Some more methods such as Local Linear Embedding(LLE) [19], Locality Preserving Projections (LPP) [20], are the two more techniques accomplished in earlier for dimensionality reduction in human action recognition system. However, these methods are unsupervised which does not guarantee good discrimination between classes. LDA is widely used technique which has an effective performance in the pattern recognition applications. Yuting Su et al. [21] accomplished LDA for open view action recognition which can learn a common space for action sample under different view by using different view using their category information. Further several variants of LDA are proposed for human action recognition such as Robust Linear Discriminant Analysis (RLDA) [22], Independent Component based LDA (IC-LDA)[23], and Regularized Discriminant Analysis (RDA) [24]. However, all these methods are assumed the uni-modality of action class which is not effective for the Multi-View Human Action Recognition which has subclasses for every action class.

III. DISCRIMINANT ANALYSIS

In the current MVHAR system, every action is captured with five cameras, each camera represents one view.

All the views carry almost same information, results in huge similar data. When the dimensionality of data is high, the algorithm can become susceptible to the well-known curse of dimensionality. It means that in the case of high dimensional data space, the action representation becomes sparse and hence larger amount of training data is required to estimate the parameters of machine learning method.

To solve this problem, several Dimensionality Reduction methods are proposed in recent years, also gained a lot of significance within the machine learning field. The main aim of any dimensionality reduction method is to determine the feature space, projection onto which results in lower dimensionality of data, while preserving the significant properties of data.

Based on machine learning algorithm associated with the Dimensionality reduction techniques are categorized into two categories such as unsupervised dimensionality reduction methods and supervised dimensionality reduction methods. The unsupervised methods only rely on the structure of data while the supervised methods rely on the structure of data as well as on the additional class label information provided by experts. PCA is the most common unsupervised method that projects the data onto the subspace where the data has highest variance. In the case of supervised dimensionality reduction methods, LDA has gained a significant research interest, which leads to an enhanced discrimination between classes due to an assumption that during training the data is given with class labels. In LDA, the optimal subspace is obtained by optimizing the Fisher-Rao's criterion which is defined as the ratio of within class scatter matrix to the between class scatter matrix, under the assumption of uni-modal classes and follows a normal distribution. Due to the incorporation of class label information only the LDA can define a subspace of at most 'd' dimensions, where d is the rank of the between class scatter matrix, which is equal to the *C-1* classes for case of *C* classes. The detailed process of LDA is illustrated below;

A. Linear Discriminant Analysis

Let's consider a dataset of N samples, $X = \{(x_i, l_i)\}_{i=1}^{N}$, where the term x_i represents the feature vector of the ith sample and l_i represents the associated class label of ith sample. Let *d* be the dimensionality of data and *C* be total number of classes. Let $(.)'$ denote the transpose operator. In the LDA, the optimal subspace is obtained through the following Fisher-Rao's criterion [26];

$$J(W) = \text{argmin}_W \frac{Tr(W^T S_w W)}{Tr(W^T S_b W)} \quad (1)$$

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;

$$S_w = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_{ij} - m_i)(x_{ij} - m_i)' \quad (2)$$

$$S_b = \sum_{i=1}^{K} (m_i - m)(m_i - m)' \quad (3)$$

Where N_i 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*.

However the main limitation of LDA is its assumption regarding the class uni-modality. Due to this limitation the LDA have very less performance when classes form subclasses, i.e., classes are represented by multiple subclasses which have different disjoint distributions. To solve this problem, the information of subclasses also needs to be incorporated. In the current Human action recognition system, every action is represented through five views captured with five cameras (CAM 0, CAM 1, CAM 2, CAM 3 and CAM 4) which have different distributions. In the case of actions considered as classes, their different view can be considered as sub classes. Similarly, if the cameras are considered as classes, the actions can be considered as sub classes. However, the LDA considers only the class label but not sub class label, which results in very much less recognition accuracy when there are subclasses derived from classes. Hence the subclass information also needs to be integrated to achieve better recognition accuracy in MVHAR.

B. Subclass Discriminant Analysis [25]

In the current human action recognition system, the descriptions of same actions from multiple view are available, resulting in multiple actions with multiple views, which can also be referred as multi-view problem. The nature of this problem is similar to the way the humans perceive the world and takes the decisions, as the real world data is not limited to one source, but consisting of different sources. When the action is perceived from different views, the decision is made by combining information from all views. Then only he HAR system can recognize more accurately. Multi-view dimensionality reduction method is the one which focuses to obtain an optimal subspace by combining the information coming from multiple views. Moreover the views have different modalities.

Subclass Discriminant Analysis is developed based on the multi-view or multi-modal Dimensionality reduction, to solve the uni-modality problem of LDA. Unlike the LDA, SDA represents the each action class with a subset of classes that are obtained by the accomplishment of clustering [27] on the action class. In the current HAR system, if one action is assumed as a class, then the five views through which it has captured are considered as sub classes. For instance, let X_i be the ith class action and $X_{ij,\ j=1,2,3,4,5}$ are subclasses. X_{i1} is the action captured with CAM1, X_{i2} is the action captured with CAM2, X_{i3} is the action captured with CAM3, X_{i4} is the action captured with CAM4, and X_{i5} is the action captured with CAM5. In SDA, the total scatter matrix S_t is minimized instead of within class scatter matrix, S_w, as $S_t = S_b + S_w$. The following mathematical expression accomplishes the SDA;

$$S_t = \sum_{q=1}^{N} (x_q - m)(x_q - m)' \quad (4)$$

$$S_b = \sum_{i=1}^{C-1} \sum_{l=i+1}^{C} \sum_{j=1}^{d_i} \sum_{h=1}^{d_l} p_{ij} p_{lh} (m_{ij} - m_{lh})(m_{ij} - m_{lh})' \quad (5)$$

Where *m*is the mean of data, *i* and *l* are the class labels, *j* and *h* are the sub class labels, p_{ij} and p_{lh} are subclass priories, $p_{ij} = N_{ij}/N$, where N_{ij} is the total number of samples in subclass *j* of class *i* and *N* is total number of samples in X. d_i and d_h are the total number of subclasses in class *i* and class *h* respectively.

The generalized solution to solve the problem (1) is formulated as a generalized Eigen decomposition as

$$S_t w = \lambda S_b w \quad (6)$$

The obtained Eigen vectors $[w_1, w_2, \ldots \ldots, w_d]$ are the minimal Eigen vectors of d dimensions and can form a projection matrix. Then based on these Eigen vectors, the input data x_i is projected as y_i as;

$$y_i = W^T x_i \quad (7)$$

IV. PROPOSED RECOGNITION METHOD

A. Overview of Framework

The overview of the proposed framework is shown in Figure.1. For a given input action sequence the proposed system finds key frames based on GSSM. Next, from the obtained frames, three different set of features are extracted. The novelty of the proposed HAR system is key frames extraction and it is done through a new SSM, called gradient SSM. Similar to our first contribution [28], this paper also focused on three set of features such as intensity features, Orientational features and contour features. After extracting individual features, a composite feature vector is formulated by concatenating all these features. Further an optimal feature subspace is obtained after applying the SDA over the composite feature vectors. Once the final set of features is extracted from test action video, they are subjected to classification through Support Vector Machine (SVM) Classifier.

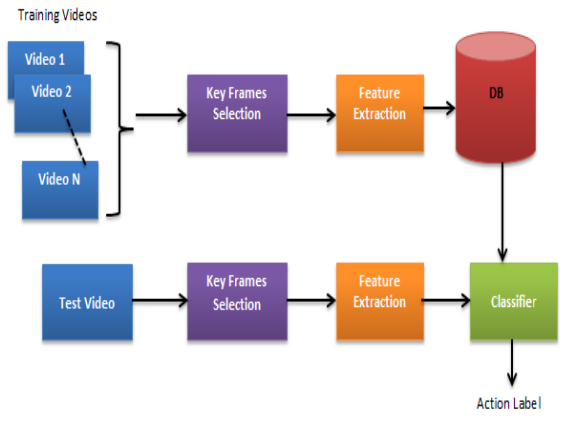

Figure.1 Block diagram of proposed HAR system

B. Key frames selection through GSSM

To extract the key frames, this paper considered the SSM as a base reference and proposed a new version of SSM, called as Gradient SSM (GSSM). Here, the main intention of GSSM is to extract the key frames with respect to the trajectories of an action sequence. For a given action, a continuous trajectory is extracted through gradients. Unlike the conventional SSM which is focused only on the exploration of temporal dynamics, the proposed GSSM focuses on the spatial dynamics also. The GSSM effectively discovers the similarity between two frames in the occlusion and moving background. The gradients evaluation for a frame att_{th} instant is shown in Figure.2.

Here, in the proposed GSSM evaluation, Laplacian operator is used to find the gradients. Since the Laplacian kernel is achieved a greater performance in edge enhancement in digital image processing, we adopt Laplacian operator to capture the spatial similarities. Mainly there are two reasons behind the consideration of Laplacian operator. (1) In the action image, the Laplacian operator can enhance the features with sharp discontinuities, highlights edges, and also can find the fine details. (2) Since the Laplacian operator is the second order derivate, it is more effective than the first order derivative in the analysis of finer details of image. Due to these two reasons, the Laplacian operator is considered here to extract the edge details and to highlight the finder details.

Consider a frame F_i from the available N frames $F = \{F_1, F_2, F_3, \ldots., F_N\}$ of an action sequence, it is represent with a set of feature as $F_i = \{f_1, f_2, f_3, \ldots., f_M\}$, where $f_i \in R^d$ is the ith feature. First apply gradient operator ∇ on the frame F, resulting in a first order gradient sequence as

$$G = \nabla F = \{G_1, G_2, G_3, \ldots., G_M\} \quad (8)$$

Where $G_i = dF_i/dt = f_i - f_{i-1}$. Since the input considered here is a 2-D image, the gradients are applied in the two directions, i.e., horizontal and vertical directions. Let G_{iH} and G_{iV} be the horizontal and vertical gradients respectively of an feature f_i, the final gradient can be obtained as

$$G_i = \sqrt{G_{iH}{}^2 + G_i} \quad (9)$$

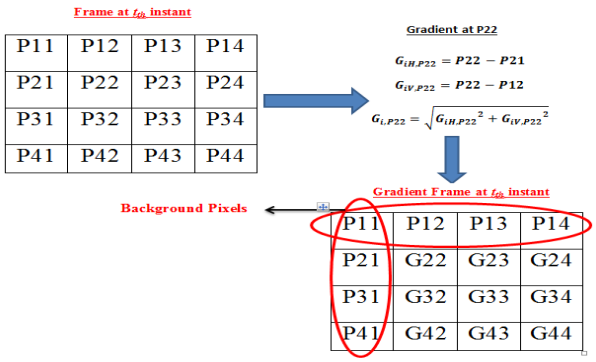

Figure.2 Gradients Evaluation of a sample frame

For every pixel/feature of a frame, this operation is performed and the entire frame is represented with first order derivatives. Next, apply the gradient operator ∇ to G, resulting in an another sequence as

$$L = \nabla G = \{L_1, L_2, L_3, \ldots., L_M\} \quad (10)$$

Where $L_i = dG_i/dt = G_i - G_{i-1}$. Since the input considered here is a 2-D image, the gradients are applied in the two directions, i.e., horizontal and vertical directions.

Let L_{iH} and L_{iV} be the horizontal and vertical gradients respectively of feature G_i, the final second order gradient can be obtained as

$$L_i = \sqrt{{L_{iH}}^2 + {L_{iV}}^2} \quad (11)$$

The resultant sequence L is the second order difference of a frame F. Simply L can be represented as $L = \nabla^2 F$. Initially, every frame is processed for gradients evaluation and then the resultant gradient frames are processed for SSM evaluation.
Similar to the case when the Laplacian operator applied over an image/frame, the steep changes, edges and finer details are enhanced which are more helpful in the detection of key frames from a larger set of frames. For a key frame, to discriminate between two actions, the system should have sufficient knowledge and it is provided by the enhanced edges, sharp points and steep changes.

Once the key frames are extracted from every view, they are processed for feature extraction according to the method described in [28].

C. Feature Extraction

After extracting the key frames through GSSM, they are processed for feature extraction and here totally three different set of features namely Intensity Features, Orientational Features and Contour Features are extracted and SDA is applied to reduce the dimensionality. Figure.3 shows the architecture of proposed feature extraction mechanism.

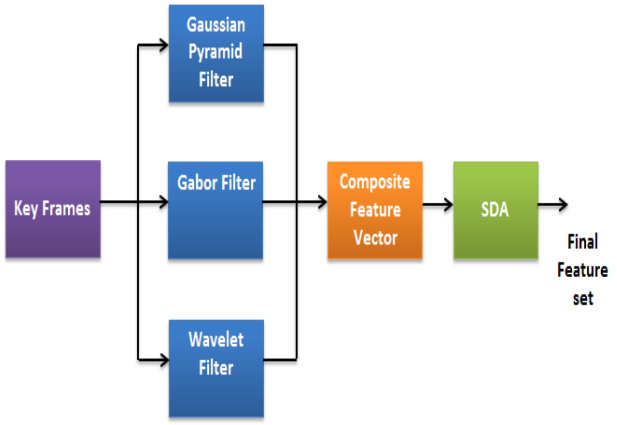

Figure.3 Schematic of feature extraction and dimensionality reduction

To extract the intensity features, this work applied Gaussian pyramid filtering. Here the Gaussian filter is applied in a pyramidal fashion up to seven levels and the intensity features are extracted by subtracting the high level Gaussian convolved features from low level Gaussian convolved features. For an every level, the input frame is down sampled and convolved with Gaussian filter and then subtracted from its previous level frame. Next, to extract the Orientational features, this paper accomplished Gabor filter in various orientations and scales. For a given action frame/image, the Gabor filter is applied at totally eight orientations such as 0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, and 315^0 and at different scales such as $5\times$, $7\times$, $9\times$, and $11\times$. Hence totally we will get 32 feature maps.
Only eight important feature maps are extracted from these 32 by applying max pooling over them. For every orientation, we will have four feature maps at different scale and among those four only one feature map is extracted through max pooling, hence totally we will get eight Orientational feature maps at this phase.
Further, in the contour features extraction phase, Discrete Wavelet Transform (DWT) is applied up-to five levels. Initially, the input action key frame is decomposed into four sub bands such as Approximations (A), Horizontals (H), Verticals (V) and Details (D). In the further decompositions, the approximation band is considered as input and further decomposed into four sub bands. In this manner, the DWT is accomplished up-to five levels. Similar to the intensity feature extraction mechanism, the contour features are extracted by subtracting the high level sub-band features from low level sub-band features. Here, to ensure the dimensionality equality, interpolation is applied on the high level sub bands. The subtraction is applied only between approximation bands. Finally, a 1-D feature vector is constructed by combining all the three features and applied SDA to reduce the dimensionality.

In the training phase, the obtained final set of features of every training action is trained through SVM algorithm. Next, in the testing phase, the videos of same action but captured under multiple views are tested one by one and here the SVM classifier is accomplished for classification. The SVM classifier classifies the input test action into one of the human action types and produces the label to which it belongs to. For SVM classification, we train non-linear SVMs using kernel and adopt one-against all approach for multiclass classification.

V. SIMULATION RESULTS

In this section, we assess the proposed HAR framework on two publicly available datasets namely, IXMAS and NIXMAS.The quantitative evaluation is done through various performance metrics under varying environments. To simulate the developed HAR model, MATLAB2014a software is used. Initially the training process is performed through different videos having different action sequences and also in different views. After the completion of training, testing is performed through different action sequences and with different views.

A. Dataset details

We consider totally two different datasets. The first dataset is INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset. This dataset consists of 12 action classes such as *point out (PO), pick up (PU), wave (WV), punch (P), turn around (TA), walk (WA), sit down (SD), get up (GU), Cross arms (CA), scratch head (SH), Kick (K)* and *check watch (CW)*. Each action is performed three times and 12 different subjects are recorded with five cameras, four are fixed at four sides and one is fixed on the top. These five cameras capture five views such as left, right front back and top. The frame rate is 23 frames per second and the size of frame is 390×291 pixels. Some action samples of this dataset are shown in Figure.4. The next dataset is NIXMAS having new videos with same action as of IXMAS dataset. The overall sequences present in this dataset are 1148.

The actions recorded under this dataset are with different camera, actors and viewpoints. Moreover, the two to third ratio of videos are having the objects that occlude the actors.

The major difference between IXMAS and NIXMAS is background only. In the IXMAS action videos, for all views, the background is constant and non-varying in nature, but in the NIXMAS dataset, the background is varying and consists of various objects. Some action samples of this dataset are shown in figure.5.

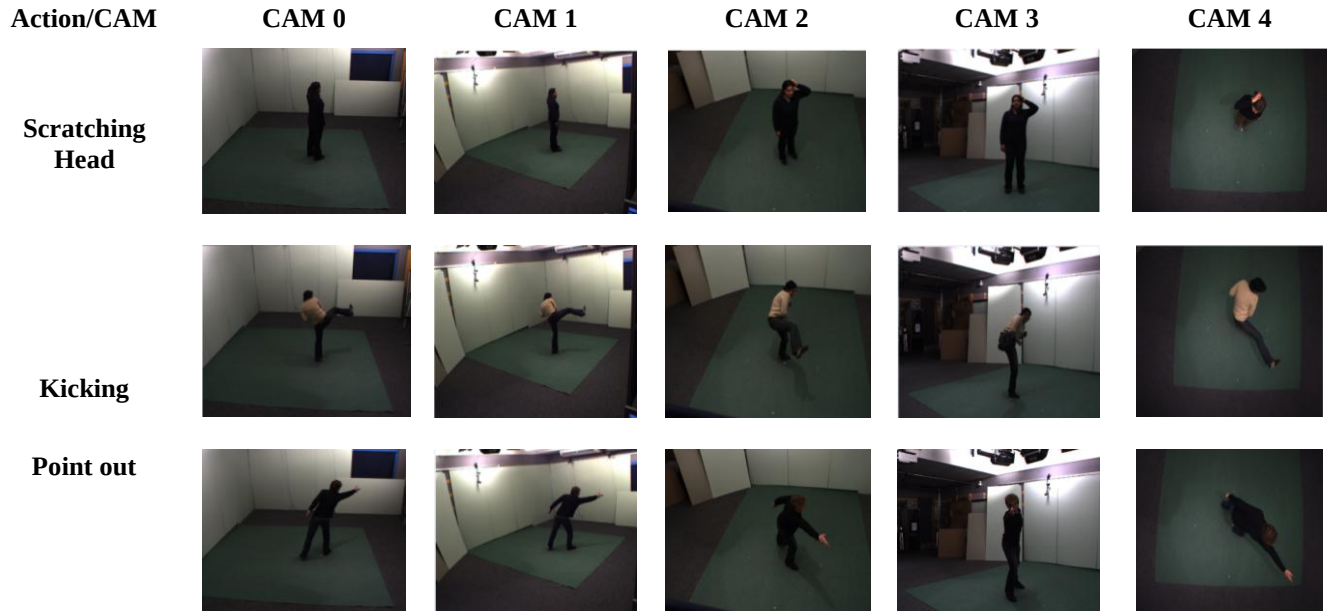

Figure.4 IXMAS dataset Action sample under multiple views

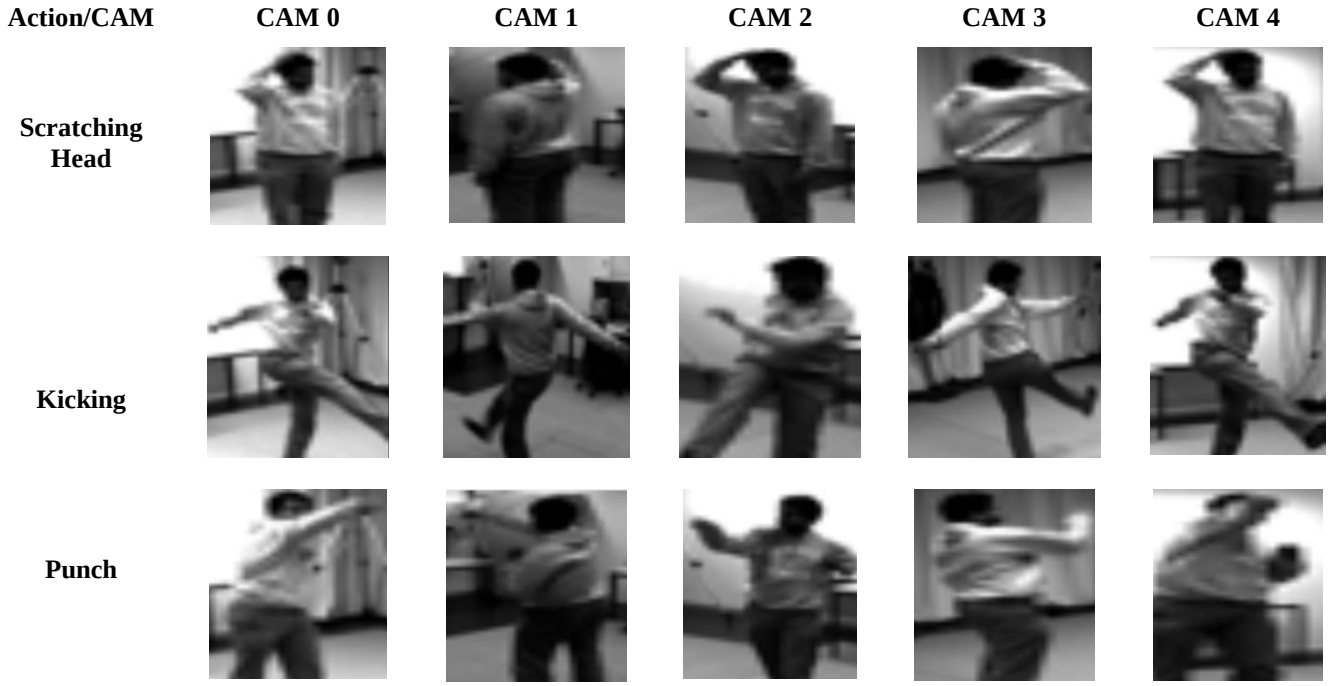

Figure.5 NIXMAS dataset Action samples under multiple views

B. Results

Here totally 12 actions are considered for every actor under different views. After the simulation of different actions sequences through the proposed HAR system, the obtained results are represented in this sub-section.

Table.1 shows the results of proposed action recognition system under for actions of IXMAS dataset. Under this simulation, the proposed approach considered the action as a class and Views as sub classes. For example, let's consider check watch as one class, then the five views captured through five cameras such as CAM 0, CAM 1, CAM 2, CAM 3 and CAM 4 are considered as subclasses.

Hence totally for 12 actions, we have considered 12 classes and 60 subclasses. The values shown in table.1 are obtained by averaging the values
obtained under individual views. For example, the TPR of check watch is 94.9669% and it is an average of five views of check watch action. Similarly, the remaining metrics such as TNR, PPV are also measured on the same basis.
Further the FNR is measured as 100-TPR and FPR is measured as 100-TNR. Finally, the F-score is measured as the harmonic mean of TPR and TNR. A higher value of TPR, PPV and TNR defines the better performance and vice versa. Since the F-score is a linear related with TPR and TNR, a higher value of F-score shows the good performance.

Unlike, the lower value of FPR and FPR depicts the better performance. From table.1, we can notice that the maximum TPR (95.0745%) is observed for Wave action, maximum TNR (96.1211%) is observed for Cross Arms, maximum PPV (95.1191%) is observed for Check Watch and maximum F-sore (95.4407%) is observed for check watch action. Further, the minimum FPR (3.8788%) is observed for Cross Arms and minimum FNR (4.9254%) is observed for Wave action.

Table.1 Performance Metrics for different actions under IXMAS dataset

Action/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
Check Watch	94.9669	95.9188	95.1191	4.0811	5.0330	95.4407
Cross Arms	94.7462	96.1211	94.9151	3.8788	5.2537	95.4286
Scratch Head	94.6392	94.9017	94.2905	5.0982	5.3609	94.7702
Sit Down	91.3276	92.9098	92.0880	7.0901	8.6723	92.1119
Get Up	92.2559	93.4091	92.8039	6.5908	7.7440	92.8289
Turn Around	90.5620	91.2800	91.2355	8.7199	9.4379	90.9195
Walk	94.0116	94.3019	93.8891	5.6980	5.9883	94.1565
Wave	95.0745	95.7146	94.5138	4.2853	4.9254	95.3934
Punch	88.4586	90.1684	88.7424	9.8315	11.541	89.3053
Kick	90.0540	90.8415	89.9676	9.1584	9.9459	90.4460
Point Out	90.7565	91.8331	91.7678	8.1668	9.2434	91.2916
Pick Up	91.5857	92.1001	91.2936	7.8998	8.4142	91.8421

Table.2 Performance Metrics for different actions of IXMAS under Different Views

Camera/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
CAM 0	93.1282	92.5346	92.0041	7.4654	6.8718	92.8305
CAM 1	92.2783	91.3348	93.3449	8.6652	7.7217	91.8041
CAM 2	91.1563	92.5043	94.7368	7.4957	8.8437	91.8254
CAM 3	94.5364	92.2794	92.2914	7.7260	5.4636	93.3943
CAM 4	88.5345	85.4678	89.4571	14.5322	11.4655	86.9741

Table.3 Performance Metrics for different actions under NIXMAS dataset

Action/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
Check Watch	92.5660	93.5034	93.6618	6.4966	7.4340	93.0323
Cross Arms	92.3231	93.6604	92.9540	6.3396	7.6769	93.1893
Scratch Head	91.9827	92.9196	92.0785	7.0804	8.0173	92.4487
Sit Down	89.5871	90.5240	89.6829	9.4760	10.4120	90.0531
Get Up	90.1111	91.0480	90.2069	8.9520	9.8889	90.5771
Turn Around	88.4744	89.4113	88.5702	10.588	11.5250	88.9403
Walk	91.5400	92.4769	91.6358	7.5231	8.4599	92.0060
Wave	92.5733	93.5102	92.6691	6.4898	7.4267	93.0393
Punch	86.5657	87.5026	86.8871	12.497	13.4343	87.0316
Kick	87.8582	88.7951	88.0040	11.204	12.1418	88.3241
Point Out	88.6049	89.5418	89.2214	10.458	11.3951	89.0708
Pick Up	89.2333	90.1702	89.3333	9.8298	10.7667	89.6993

Table.4 Performance Metrics for different actions of NIXMAS dataset under Different Views

Camera/Metric	TPR (%)	TNR (%)	PPV (%)	FPR (%)	FNR (%)	F-Score (%)
CAM 0	86.2836	86.6900	86.1595	13.3099	12.7163	86.9858
CAM 1	86.4337	85.4902	87.5003	14.5097	13.5662	85.9593
CAM 2	85.3117	86.6597	85.8922	13.3402	14.6882	85.9803
CAM 3	88.6918	86.4348	88.4468	13.5651	11.3081	87.5484

CAM 4	82.6899	79.6232	83.6125	20.3767	17.3100	81.1268

Table.2 shows the results of proposed action recognition system under for actions of IXMAS dataset under several views. Under this simulation, the CAM is considered as Class and the actions are considered as subclasses. Totally, five CAMs with fives views are considered as main classes and the 12 actions captured under these CAMs are considered as Sub classes. In this simulation, the input test action with captured with CAM 1 is given as input and at the output the class label is checked for CAM and sub class label is checked for action.

For instance, consider the check watch action with CAM 1 as input, the output label is checked for CAM (whether CAM 1 or other) and sub class label is checked for action (whether Check watch or other). In this manner, all the 12 actions with different views are tested and the average values are depicted in table.2.

From table.2, we can notice that the maximum TPR (94.5364%) is observed for CAM 3, maximum TNR (92.5346%) is observed for CAM 0, maximum PPV (94.7368%) is observed for CAM 2 and maximum F-sore (93.3943%) is observed for CAM 3. Further, the minimum FPR (7.4654%) is observed for CAM 0 and minimum FNR (5.4636%) is observed for CAM 3. In all the CAMs, the lowest performance is observed at CAM 4. Since CAM 4 is a top view, the entire actions is not covered and hence the performance is low.

Next, the proposed system is simulated through NIXMAS dataset and the obtained results for different actions and different views are shown in table.3 and table.4 respectively. Even though the Actions of NIXMAS dataset are occluded, the proposed feature extraction mechanism captured the more discriminative features of every action effectively. Due to the presence of Gabor filter and Wavelet filter, the contours of all actions which depict the action features are extracted perfectly. Moreover, the SDA ensures a perfect discrimination between all the actions thereby achieved an optimal recognition performance. From table.4, we can notice that the maximum TPR (92.5733%) is observed for Wave action, maximum TNR (93.6604%) is observed for Cross Arms, maximum PPV (92.9540%) is observed for Check Watch and maximum F-sore (93.1893%) is observed for Cross Arms action. Further, the minimum FPR (6.3396%) is observed for Cross Arms and minimum FNR (7.4267%) is observed for Wave action. Next, the performance metrics measured for NIXMAS under multiple views are depicted in table.4. From table.4, we can notice that the maximum TPR (88.6918%) is observed for CAM 3, maximum TNR (86.6900%) is observed for CAM 0, maximum PPV (88.4468%) is observed for CAM 3 and maximum F-sore (87.5484%) is observed for CAM 3. Further, the minimum FPR (13.3099%) is observed for CAM 0 and minimum FNR (11.3081%) is observed for CAM 3.

C. Comparative Analysis

Under this subsection, the proposed action recognition approach is compared with conventional action recognition methods such as IC-LDA + HMM [23] and SSM+SVM [28]. The comparative analysis is accomplished through the recognition accuracy .Initially, the recognition accuracy is compared with respect to actions under multiple views and then through datasets.

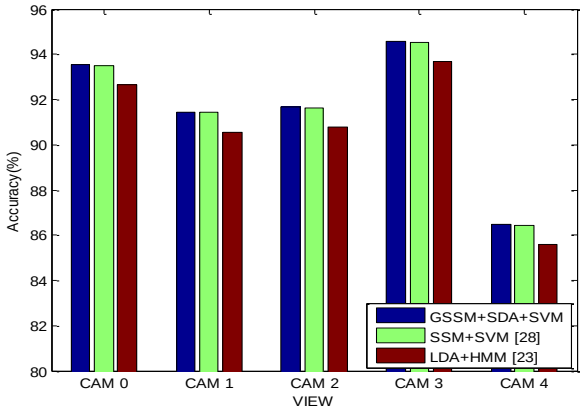

Figure.6 Accuracy Comparison under multiple views

Figure.6 describes the details of accuracy comparison between proposed and conventional approaches through multiple views. Here the accuracy is measured for individual CAMs. For this purpose, the actions under single CAM are trained and then processed for testing. For instance, the total 12 actions captured with CAM 1 are trained and tested. Similarly, the actions captured with remaining CAMs are also trained and tested. Under this case, the CAM is considered as Class and the actions are considered as sub class. According to figure.6, the actions captured through CAM 3 have gained a maximum accuracy compared to remaining CAMs. Since the CAM 3 is front view through which the actions can be viewed more clearly, the recognition system can acquire the exact features from every action. These exact features can provide a sufficient discrimination after processing through SDA. Moreover, the accuracy of actions captured with CAM 4 have less accuracy due to the top view. In the top view, some actions like sit down, get up, turn around, and walk can't be viewed because they have no strict motions in hands and leg movement. Hence the top view has less accuracy.

Furthermore, the proposed GSSM + SDA + SVM has achieved better accuracy compared to the conventional approaches such as IC-LDA + HMM [23] and SSM + SVM [28]. IC-LDA is a dimensionality reduction method which doesn't consider the subclasses and SSM + SVM has no dimensionality reduction method.

Hence these two approaches have less accuracy. On an average, the proposed GSSM + SDA+SVM has gained an accuracy of 91.5346%, whereas the accuracy of conventional approaches is observed as 91.0254% and 90.6141% for IC-LDA + HMM and SSM + SVM respectively. The novelty of proposed approach SDA which can effectively discriminate the same action under different views, which is not with conventional approach, SSM + SVM.

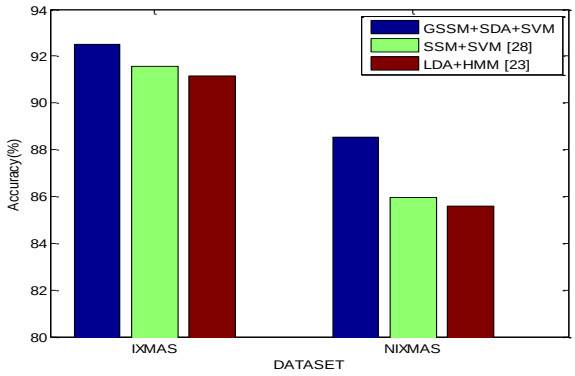

Figure.7 Accuracy Comparison under multiple Datasets

Figure.7 describes the details of accuracy comparison with respect to two datasets such as IXMAS and NIXMAS. As it can be seen form above figure, the accuracy gained for IXMAS dataset is high compared to the accuracy of NIXMAS dataset. Since the actions of NIXMAS dataset are occluded, the exact discriminative features can't be extracted ad hence the system recognize one actions as other action. The videos of IXMAS dataset are more qualitative and hence the motion features are extracted through the proposed feature extraction technique.

These features are more helpful in the provision of an efficient discrimination between actions and hence the proposed approach gained more accuracy. Moreover, the proposed feature extraction technique is composed of three different set of filters such as Gaussian, Gabor and Wavelet and hence the proposed system has gained more recognition accuracy even in the case of Occluded dataset, NIXMAS. Here the accuracy is average accuracy for all camera views. On an average, the accuracy of proposed GSSM+ SDA + SVM is observed as 90.5395% whereas the accuracy of conventional approaches is 88.7613% and 88.0065% for IC-LDA + HMM and SSM + SVM respectively.

VI. CONCLUSION

In this paper, a novel multi-view human action recognition is proposed based on Self-Similarity matrix and Linear Discriminant analysis. A novel extension of the SSM, Gradient SSM is introduced in this paper, which finds the similarities between the frames of a same action through the gradients. GSSM helps in the reduction of unnecessary frames which have no significant motion information. Only key frames are extracted through the GSSM. Next, to extract all the possible features by which the motion shape of body can be extracted, a composite feature extraction method is proposed which is composed of three different set of features. Further to ensure a sufficient discrimination between different actions as well as between different views, SDA is accomplished. Simulation experiments are conducted over two standard publicly available benchmark datasets, IXMAS and NIXMAS. The obtained results had proven the superior performance of proposed system when compared to the state-of-art methods.

REFERENCES

1. Q. Ke, J. Liu, M. Bennamoun, S. An, F. Sohel, F. Boussaid, Computer vision for human-machine interaction, Comput. Vision Assist. Healthc. (2018) 127–145.
2. Teddy Ko, "A survey on behavior analysis in video surveillance for homeland security applications", In: Proc. 37th IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, pp. 1–8, 2008.
3. S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, Visual Computer 29 (10) (2013) 983–1009.
4. T. Syedamahmood, M. Vasilescu, and S. Sethi, Recognizing action events from multiple viewpoints, in Proc. Eventvideo, 2001, pp. 64–72.
5. I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. Pre-Prints, 2010.
6. R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd Edition, Wiley,New York, NY, USA, 2000.
7. J. Ye, Least squares linear discriminant analysis, International Conferenceon Machine Learning, (2007) 1087–1093.
8. H. Wang, X. Lu, W. Zheng, Fisher discriminant analysis with l1-norm,IEEE transactions on cybernetics, 4 (2014) 828–842.
9. D. Weinland, E. Boyer, and R. Ronfard, "Action recognition fromarbitrary views using 3D exemplars," in Proc. IEEE Int. Conf. Comput.Vis., Oct. 2007, pp. 1–7.
10. P. Yan, S. M. Khan, and M. Shah, "Learning 4D action feature modelsfor arbitrary view action recognition," in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Jun. 2008, pp. 1–7.
11. D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robustto occlusions and viewpoint changes," in Proc. 11th Eur. Conf. Comput.Vis., 2010, pp. 635–648.
12. E. Shechtman and M. Irani.Matching local self-similarities across images and videos.In CVPR, 2007.
13. M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In ICCV, 2009.
14. K. Chatfield, J. Philbin, and A. Zisserman. Efficient retrieval of deformable shape classes using local self-similarities. In NORDIA Workshop at ICCV 2009, 2009.
15. C. H. Lampert, H. Nickisch, S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer.In CVPR, 2009.
16. I. Junejo, "Self-similarity based action recognition using conditionalrandom fields," in Information Retrieval Knowledge Management (CAMP), 2012 International Conference on, march 2012, pp. 254 –259.
17. T. Deselaers and V. Ferrari, "Global and efficient self-similarity forobject classification and detection," in Computer Vision and patternrecognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp.1633–1640.
18. jing wang, and huicheng zheng, "view-robust action recognition based on temporal self-similarities anddynamic time warping", ieee international conference on computer science and automation engineering (csae), zhangjiajie, china, 2012.
19. Tat-Jun Chin, Liang Wang, Konrad Schindler, David Suter, Extrapolatinglearned manifolds for human activity recognition, in: Proceedings of theinternational Conference on Image Processing (ICIP'07), vol. 1, San Antonio,TX, September 2007, pp. 381–384.
20. Liang Wang, David Suter, Visual learning and recognition of sequential datamanifolds with applications to human movement analysis, Computer Visionand Image Understanding (CVIU) 110 (2) (2008) 153–172.
21. Yuting Su, Yang Li and Anan Liu, "Open View human action recognition based on Linear Discriminant Analysis", Multimedia tools and applications, Vol.78, Issue 1, pp.767-782, 2019.
22. Ming Guo and Zhelong 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), Calabria, Italy, 2015.
23. Md. Zia Uddin, J. J. Lee, T. S. Kim, "Independent Component feature-based human activity recognition via Linear Discriminant Analysis and Hidden Markov Model", 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 2008.
24. B Mandal, How-Lung Eng, "Regularized Discriminant Analysis for Holistic Human Activity Recognition", IEEE intelligent systems, 2012.
25. M. Zhu, A. Martinez, Subclass discriminant analysis, IEEE transactions onpattern analysis and machine intelligence 28 (2006) 1274–1286.
26. R. Fisher, The statistical utilization of multiple measurements, Annals ofeugenics 8 (1938) 376–386.

27. X. Chen, T. Huang, Facial expression recognition: a clustering-based approach,Pattern Recognition Letters 24 (2003) 1295–1302.
28. K. Pradeep Reddy, G.Apparao Naidu, and B.vishnuvardhan, "View-Invariant Feature Representation for Action Recognition under Multiple Views", International Journal of Intelligent Engineering and Systems, Vol.12, No.6, 2019.

AUTHORS PROFILE

Mr. K.PradeepReddy is presently Research Scholar in JNTUH and working as Associate Professor of CSE at Tirumala Engineering College,Bogaram,hyderabad. He has submitted one patent and one book chapters with international publishers. His areas of interests are Image Processing, Data mining, Information security and Machine Learning. Published more than 20 papers in various International Journals and Conferences. He was Conducted 5 Workshops and Attended 2 Workshops. He was Guided 65 Projects for UG and PG.

Dr. G.Apparao Naidu is presently working as Dean and Professor of CSE at JBIET,Moinabad.Under his guidance 8scholars' research work is in progress. He has submitted one patent and two book chapters with international publishers. His areas of interests are Data mining, Information security and Machine Learning. Published more than 60 papers in various International Journals and Conferences. He was Conducted 12 Workshops and Attended 28 Workshops. He was Guided 88 Projects.

Dr. B Vishnu Vardhan is presently working as Vice Principal and Professor of CSE at JNTUH College of Engineering Manthani, Peddapally. Under his guidance 13 scholars were awarded with PhDs and another 8 scholar's research work is in progress. He has submitted four patents and two book chapters with international publishers. His areas of interests are Linguistic processing, Data mining, Natural language processing, Information security and Machine Learning. Published more than 100 papers in various International Journals and Conferences. He has completed one UGC project worth 9 lakhs and one state funded project worth 5 lakhs. He is acting as a Chairman, Board of Studies for Computer Science Department for Sathavahana University.