IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

HUMAN ACTIVITY RECOGNITION IN EGOCENTRIC VIDEO USING GIST AND INTEREST POINTS DESCRIPTORS

Dr. S.Anu H Nair[1], Dr. B.Rajalingam[2], S.Sunanda[3], P.Krishna Reddy[4]

[1,3,4]Assistant Professor, [2]Associate Professor & HOD

[1]Department of CSE, Annamalai University, Chidambaram, India. (Deputed to WPT Chennai)

[2.3.4]St.Martin's Engineering College, Dhulapally, Secunderabad - 500100

anu_jul@yahoo.co.in, rajalingam35@gmail.com

ABSTRACT

In the contemporary world, recognizing the activity that assists old age people, differently able population has become a challenging task. This paper aims at development of a system for recognition of activities of human in egocentric video. In our proposal from the activity videos, we retrieved different features such as Generalized Search Tree (GiST) and interest point features like Scale Invariant Feature Transform (SIFT), Speeded up Robust Features (SURF) and Space Time Interest Points (STIP). Classification of activity is done by classifiers such as Probabilistic Neural Network (PNN), Support Vector Machine (SVM), k-Nearest Neighbor (kNN) as well as our combined SVM with k Nearest Neighbor (SVM+kNN) classifiers. Also, the input selected is multimodal egocentric activities. The results of the study show that SVM+kNN classifier has enhanced performance in comparison to other classifiers already in use.

Keywords: EgocentricActivity; Generalized Search Tree; Scale Invariant Feature Transform; Speeded up Robust Features; Space Time Interest Points; Probabilistic Neural Network;

1. INTRODUCTION

The latest trends in computer vision is identification of human activities from video. The novel improvements in wearables have caused the current interest in identification of human activities from egocentric that is self-centered. Egocentric means more self-centered rather than the community. Analyzing the activity of an individual is to help the geriatric population, differently abled people, etc., [1]. In the welfare field, activities of daily living (ADL) has number of categories in domestic behaviors. The user's lifestyle is monitored with home monitoring system to provide one day analysis of activities of daily living. A special health care is used to monitor the activities of daily that is utilized to improve a person's physical and mental strength. In recent days, for extending the application use on daily basis a glass type camera device is provided. All objects from the egocentric videos are identified on a study on ADL [2]. The behavior of users has been defined in this taxonomy. For example, the circumstance of "watching TV" is defined by the objects in the frame such as "a TV", "a sofa" and "a remote". This technique segregates every unique object as "active object" or "passive object" that is dependent on manipulation of object by the user. While estimating activities "Active object" acts as the key object. Still this technique has few shortcomings, there are restriction scenes and object types [3-13] that can be used. In this article, GiST, SIFT, SURF and STIP are extracted. Here, SIFT, SURF, STIP are interest point descriptors. For studies, multimodal egocentric activity dataset is applied that has 4 top level as well as 20 II level categories. The 4 top level categories include mobility, routines, office work and workout. Mobility has 8 II level categories such as walking, climbing up the stairs, climbing down the stairs, taking elevator upstairs, taking elevator downstairs, taking escalator upstairs and taking escalator downstairs. Routines has 4 II level categories such as eating, drinking, texting and calling. Office work has 4 II level categories such as working in computer, reading, writing and organizing files. workout has 4 II level categories such as running, push-ups, sit-ups and cycling.

2. LITERATURE RETROSPECT

Recognizing Human behaviors from video is a common question for exploration. Activity recognition are given importance on previous works [14–16]. In [14], identification of Complicated kitchen activities is done by the history of tracked features. In recent days, Activity of Daily Living analysis is applied using RGB-D sensors [15]. In this the capability is enhanced using traditional cameras. In [16] a kitchen environment is selected. The author [17– 21] demonstrated the problem with human behavior analysis with the data from wearable camera. The author [17] identified few features with the multiple object detector's output. Later, the author [18] employed a technique to separate social interaction in egocentric video that is got from social events. Here egocentric activity recognition is done on office environment where, motion descriptors are derived and merged with eye movement of the users. Few recent models have used varying surroundings such as kitchen, office, and so on. [19] used a kitchen environment codebook generation that can decrease the issue caused by varying styles and speed among various objects. In [20] a technique used for identification of anomalous events obtained through videos from chest mounted camera. The author [21] demonstrated a method for targeting FPV by video summarization. The author [22], enquired multi-channel kernels to merge details of local and global motion. This made a novel technique for identifying activities that designs the first-person video data's temporal structures. In this work [23] First Person Videos that is temporally classified to 12 hierarchical classes was

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

utilized. Multi-task learning framework analyses with FPV ADL analysis in this model. The author [24], the Speeded up Robust Features features were utilized for object identification. In his work the author [25] demonstrated extraction of SURF features for registering the images. [26] implemented feature extraction using GiST feature for character recognition. [27] illustrated the STIP feature for feature extraction.

3. PROPOSED WORK

ADL of people are tracked using egocentric video. There are 2 phases in the system. Initial phase is the training phase latter is testing phase. GiST and interest point descriptors or point features undergo extraction and analysis.

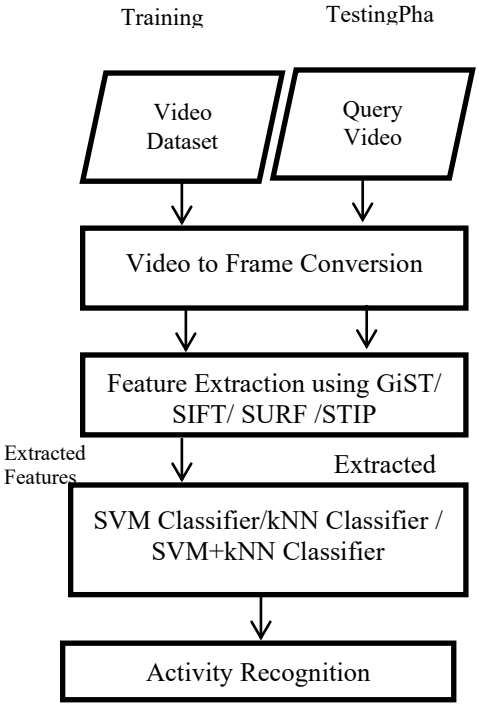


Fig. 1 Proposed System- Work Flow

The following steps are used during the training phase of the system

1. Application of feature extraction to the input video.
2. The features are GiST, SIFT, SURF and STIP.
3. The classifiers like PNN, SVM, k Neural Network as well as SVM+kNN are given with obtained features as input.
4. The output received is the activity recognition.

These are again done in testing phase. In this phase, query video is fed in. Figure 1 explains the work flow of both the phases.

4. FEATURE EXTRACTION

4.1 GiST

The Generalized Search Tree is the meaningful data which a person can recognize with a glimpse of an image. The Generalized Search Tree description contains the semantic label ofan image, and some objects and their surface characteristics, and finally the spatial layout and the semantic characteristics based on the function of an image. Hence, an explanation of semantic data the people comprehend and perceive are incorporated by GiST. GiST descriptors are utilized to demonstrate a low-dimensional image which has necessary data to distinguish a scenario of image. These descriptors can depict the dominant spatial structure of a scene with a bunch of perceptual dimensions which can be acquired by analysis of spatial frequency and orientation. Instinctly, GiST merges the gradient information (Scaleand Orientation) for various areas of an image, that gives an approximate description of the scene. The implementation used here first preprocesses the input image by converting it to gray scale, normalizing the intensities also locally scaling the contrast. The output image is then partitioned into a grid on number of scales and the response of each cell is calculated with a series of Gabor filters. All of the cell responses are interlinked to form the feature vector. Here,512 GiST features are extracted from the input video.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

4.2 SIFT

SIFT has enhanced capability of recognizing objects even if they are in clutter as well as partial occlusion. The reason is SIFT feature descriptor is not affected uniform scaling, orientation and is partly affected by affine distortion and illumination changes. SIFT extraction contains below steps:

1. **Scale-space extrema detection:** The principal phase of calculation starts searching every scales and image locations. It is applied in efficient way with the help of difference-of-Gaussian function to distinguish potential interest points. Function, $L(x, y,\sigma)$ defines the scale space of image. This is derived when a variable-scale Gaussian, $G(x,y, \sigma)$ convolves with input image $I(x, y)$:

$$L(x, y,\sigma) = G(x, y,\sigma) * I(x, y) \quad (1)$$

In this equation, * depicts x and y convolution operation,

$$G(x, y,\sigma) = \frac{1}{2\pi\sigma^2} - (x^2 + y^2)/2\sigma^2. \quad (2)$$

$D(x,y,\sigma)$ - difference of gaussian function convolved when is image is. The difference between the neighboring scales split up using k which is a constant multiplicative factor:

$$D(x, y,\sigma) = (G(x, y, k\sigma) - G(x, y,\sigma)) * I(x, y)$$
$$= L(x, y, k\sigma) - L(x, y,\sigma). \quad (3)$$

2. **Keypoint localization:** In every situation, a model fits to obtain the location and scale. Keypoints are identified based on measures of their stability. The scale of the keypoint is utilised to select the Gaussian smoothed image, L, which has the closest scale. Therefore, all computations are implemented in a scale-invariant manner.

3. **Orientation assignment:** Every key point is allotted with 1 or more key point location based on local image gradient directions. All operations are implemented on image data which has been transformed relative to the assigned orientation, scale and location for each feature, thereby providing invariance to these transformations. For each image sample, $L(x, y)$, the gradient magnitude, $m(x, y)$, and orientation, $\theta(x, y)$, is computed using pixel differences:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (Lx, y+1) - L(x, y-1))^2} \quad (4)$$

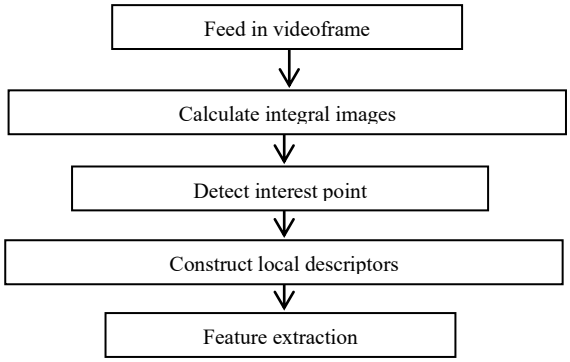
$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1))/(L(x+1, y) - L(x-1, y))) \quad (5)$$

Gradient orientation of the region of sample points around the keypoints gives an orientation histogram. It covers 360-degree orientation range with 36 bins.

4. **Keypoint descriptor:** The local image gradients are computed at chosen scale near each key point in the region. The above mentioned are converted into a presentation which permits considerable levels of distortion of local images and variation in illumination. The study results suggest that better performance is obtained using a 4x4 array of histograms and every array has 8 orientation bins. Hence, 4x4x8= 128 element feature vector has been used.

4.3 SURF

SURF is an enhanced feature detector that is utilized like computer vision tasks' components. SIFT descriptor propels it. SURF is faster and more vigorous in comparison to SIFT and many other image transformations than SIFT. SURF uses the 2D Haar wavelet responses' overall values which makes use of the integral images [28]. An integer approximation is used to the determinant of Hessian blob detector, this could undergo faster processing with an integral image. The overall value of Haar wavelet responses within a point of interest is used for features [29][30]. Figure 2 describes the workflow for SURF feature extraction.



International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

Fig. 2 workflow for SURF feature extraction

Step by step process to calculate Integral image:

1. Below equation defines the integral image of the input video frame "u":

$$U(x,y) = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le x} u(i,j) \qquad (6)$$

Integrating images assists in expediting the process of box type convolution filter.

2. 'u' is the input video frame and it convolves using 2-D Haar vertical as well as horizontal filters.

Step by step process to detect integral image:

1. Hessian matrix is used for identifying both the scale and the location of the interest point.

Let*x*be a point which is given as $(x, y)^T$ in an input image I. The Hessian matrix *H (x, σ)* is defined as follows:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \qquad (7)$$

In this equation, $L_{xx}(x, \sigma)$ depicts the convolution of X depicts mid-point using the Gaussian filter $\frac{\partial g(\sigma)}{\partial x^2}$

2. Hessian matrix determinant *H (x, σ)*, ΔH is given by

$$\Delta H = D_{xx}D_{yy} - (wD_{xy})^2 \qquad (8)$$

Trial and error method are used for weight allotment weight *w = 0.9* to obtain results.

3. The large change in scale among the first layers of every octave Scale is used to compute space interpolation.

An octave defines a series of filter response maps. This is achieved by convolution of the same input image with an increasing size of filter.

The octave index o and the interval index *i* and finally the scale sampling is calculated using the equation,

$$\sigma = \frac{1.2}{3}(2^{\circ} \times i + 1) = \frac{1.2}{3} l \qquad (9)$$

Procedure to construct local descriptors:

1. The applied non-maximum suppression 3x3x3 neighborhood consists of images and over scale with interest point
2. The intensity content distribution among the nearby interest point is calculated by descriptor.
3. Results of Haar wavelet in both x,y directions inside a circular radius 6s surrounding the interest point is calculated using scales at which interest point was identified.
4. To extract descriptors, a square region is built around the centre of the interest point as well as its orientation is selected earlier.
5. The region undergoes partitioning into 4x4 sub-regions that saves the vital spatial details.
6. For every partitioned area, the responses from Haar wavelet are computed at 5x5 sample points with equal spaces. Here Haar wavelet responses' of horizontal and vertical directions are represented by *dx, dy.*
7. After that, the entry of feature vector are made by addition of the wavelet outputs *dx* and *dy* for each partitioned area.
8. The sum of |dx| and |dy| gives the details related to the polarity of the intensity variation .

Therefore, every sub-region contains 4D feature vector *v* where $v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

We get a descriptor vector of length 64 by relating this for all 4 x4 sub-regions.

4.4 STIP

A lot of interest has been shown on STIP in video analysis. It successfully detects valid interest points in a video. Local structures in the spatio-temporal domain have been applied with Harris spatial interest points. The important points with image values have considerable local changes in both the space and time dimensions are identified. 2 histograms of gradient as well as flow are computed in a fixed sized spatial and temporal window around an interest point.

Step 1:

- Calculate the first set of interest points (C_σ) by applying Harris Corner detector.
- **Suppress Background Interest Points:** Compute inhibition term by calculating gradient difference $\Delta_{\Theta,\sigma}$ for each point in C_σ as : $\Delta_{\Theta,\sigma}(x, y, x-u, y-v) = |\cos(\Theta_\sigma(x, y) - \Theta_\sigma(x-u, y-v))|$
- Define $C_{\alpha,\sigma}(x, y) = H(C_\sigma(x, y) - \alpha t_\sigma(x, y))$, where t is the weighted sum of gradient weights in the suppression surround to the corresponding point.
- the strength of surround suppression is controlled by α. It relies on the numerical value of interest points in the neighboring texture of mentioned point.
- **Imposing Local Constraints :** (x, y) are the positions, $C_{\alpha,\sigma}(x', y')$ and $C_{\alpha,\sigma}(x'', y'')$ are two responses for the positions in consective positions $(x', y,)$ and (x'', y''). They are the intersection of a line crossing (x, y) with $\Theta_\sigma(x,y)$ and a square given with diagonals of an 8-neighbourhood are calculated using linear interpolation.
- In case value of $C_{\alpha,\sigma}(x, y)$ is more compared to the 2 adjacent points then it is a local maximum of the neighborhood, then a point is kept. Otherwise the value is set as zero.

Step 2: Temporal Constraint

- Consider 2 frames that consecutively occur at a time and exclude the shared interest points.
- $P_{\alpha,\sigma}^{T} = C_{\alpha,\sigma}^{T}\left\{C_{\alpha,\sigma}^{T} \cap C_{\alpha,\sigma}^{T-1}\right\}$
- The points that remain are the final set of STIP. These are utilized to derive local features.

The resulting descriptor contains 20 features.

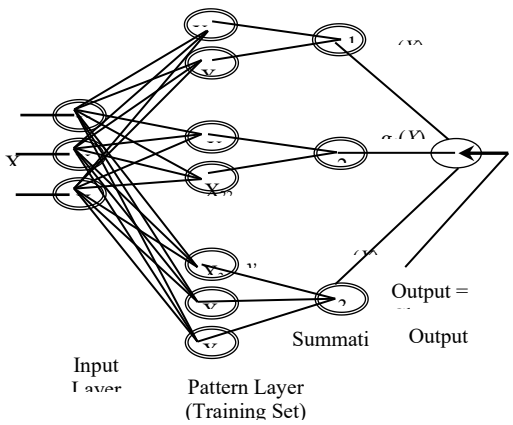
5. CLASSIFIERS

5.1. Probabilistic Neural Network:

The Probabilistic Neural Network was initially demonstrated by the author [31]. The PNN architecture has number of neurons in consecutive layers that are interconnected [32]. The input layer unit doesn't calculate and just disseminate the input values to the neurons in the pattern layer. The output of pattern x from the input layer x is calculated using the pattern layer neuron.

$$\phi_{ij}(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^{d}} \exp\left[-\frac{(x-x_{ij})^{T}(x-x_{ij})}{2\sigma^{2}}\right] \quad (10)$$

Where, d is the dimension of the pattern vector x. σ depicts the smoothing parameter and x_{ij}represents the neuron vector.



International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

Fig. 3 PNN architecture

The category c_i is obtained as the highest likelihood pattern x is computed with the summation layer neurons. The summary and average of output of all neurons are obtained [33].

$$P_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\frac{1}{N_i}\sum_{j=1}^{N_i}\exp\left[-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}\right] \quad (11)$$

where, N_i depicts the overall volume of samples in class C_i. If there are equal apriori probabilities for all classes having the similar value and the losses for taking a wrong choice, the decision layer segregates the pattern using the Bayes's decision rule that. This depends on the output of all summation layer neurons.

$$\hat{C}(x) = \arg\max\{P_i(x)\}, \qquad i = 1,2,.....n \quad (12)$$

where, $\hat{C}(x)$ depicts approximately calculated class of the pattern; x and m are the overall numerical quantity of classes in the training data. Figure 3 briefs that every image contains a specific value of the input that are combined. This is known as input pattern which explains the images' features. The input feature's quantity equals to quantity of neurons in the input layer. Overall numerical quantity of neurons is the added value of the neurons' numerical quantity that denotes the patterns for every category in pattern layer. There are nodes for every pattern layer which is in the output layer. The highest values among the propagated sum of each hidden node wins.

5.2 Support Vector Machine

SVM depict designs in machine learning which are related to learning algorithms for analysis and classification of information. A Support Vector Machine algorithm makes a design which allots a model to one category or another. The models that can be mapped using this algorithm[34]. Novel examples are given training for getting assigned to categories and they get predicted. Support vector machines have capability of performing a non-linear classification together with linear classification. We employed a multi-class support vector machine.

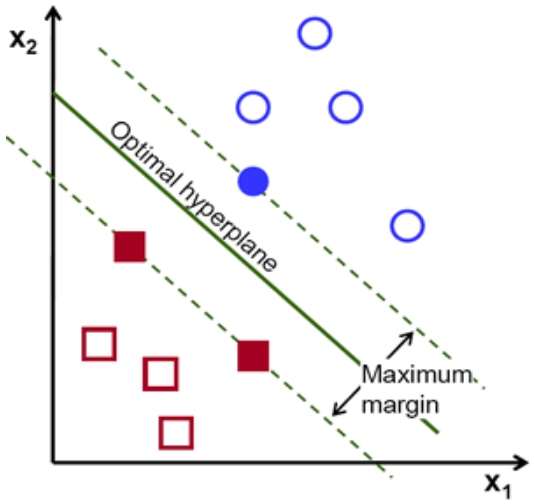


Fig. 4: Support Vector Machine

Above graph represents the separation of examples with optimal hyperplane that is referred from [35]. The operation of Support Vector Machine which depends upon the hyperplane which provides highest least distance between training samples. Recognizing activities for 4 top level categories and 20 next level categories are the outputs.

5.3 kNN Classifier

Generally, the kNN is a supervised classifier. In this, the majority of kNN categories are used to classify the data. This algorithm segregates a novel entity using attributes of the training examples. It predicts value for classification by using neighborhood. Figure 5 denotes kNN classifier that distinguishes the objects into various classes. The author [36] describes the kNN classifier.

kNN classifier algorithm steps:

1. Initialize k value
2. In the set of training objects measure the distance from the test object to every other object .
3. Then the nearest test object is chosen.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

4. Matched objects that has highest class is chosen.
5. Do this repeated till same class is obtained.

Here we used Euclidean distance function.

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, d_A(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (13)$$

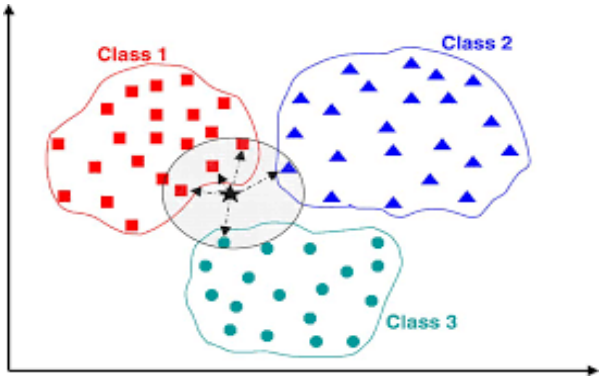


Fig.5: Classification using kNN(k=4)

Recognizing the activities of 4 top level categories and 20 II level categories are the outputs**.**

5.4 SVM+ kNN CLASSIFIER

SVM depicts a supervised learning algorithm. In this a hyperplane is made and a normal and abnormal data are isolated from one another. kNN classification algorithm is a machine learning algorithm. Support vector machine segregate data into various classes with the help of support vectors to make a hyperplane. k nearest neighbor algorithm identifies novel data. False positive rate are computed using SVM and kNN algorithms. This is the SVM+ kNN algorithm. In this SVM classifier, the Gaussian kernel can be computed by

L $$K(x,\acute{x}) = exp\left(\frac{-\|x-\acute{x}\|^2}{2\sigma^2}\right) \quad (14)$$

$\|x - \acute{x}\|^2$ - squared euclidean distance; σ - measure of expansion.

Algorithm: Combined SVM and kNN classifier

Step 1: Choose data from various class.

Step 2: Categorize data with the help of SVM classifier.

Step 3: Generate hyperplane to identify the classes among input data.

Step 4: Cluster data by applying kNN.

Step 5: Update dataset if any new addition of data;

These steps are repeated in all data in the data set.

The SVM+kNN algorithm functions along where SVM makes use of training data set for learning from data set, till novel information is included to its dataset. k nearest neighbor algorithm updates it.

6. PERFORMANCE EVALUATION

Accuracy is the ratio of sum of true positive and true negative rate to the total population.

Accuracy = Tp+Tn / (Tp+Tn+Fp+Fn) (15)

where, Tp represents those quantity that are perfectly classified as positive class

Tn denotes those quantity which is perfectly recognized as negative class

Fp represents those quantity that are perfectly classified as positive class

Fn denotes those quantity that are not correctly identified as negative class

7. EXPERIMENTAL RESULTS

Here we classified the videos into 2 levels that is 4 top level as well as 20 II level categories. We used a multimodal egocentric activity dataset that has 20 separate activities and clustered into 4 top level categories such as mobility, routines, office work and workout. Sample Frames are given in below images.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

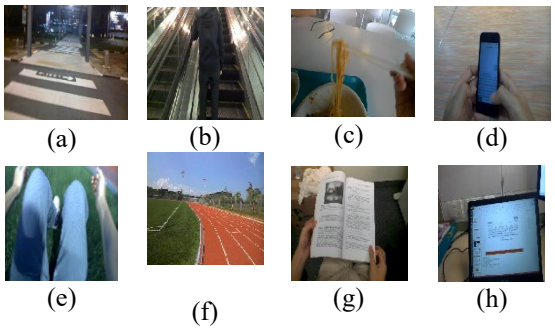


Fig. 6 Sample frames

(a) - Walking (b) - Riding Escalator Up, Daily activities are (c) – Eating (d) –Texting, workout activities are (e)- doing sit-ups (f) – Running, Office activities are (g) -Reading, (h)- Working at PC

Table 1: Accuracy for top level categories using GiST

Classifiers / Top Level Categories	PNN	SVM	kNN	SVM+ kNN
Mobility	77.10	84.04	81.86	**85.93**
Routines	78.25	83.71	80.81	**84.27**
Office Work	76.31	81.05	78.93	**82.25**
Workout	75.86	82.06	79.10	**83.37**

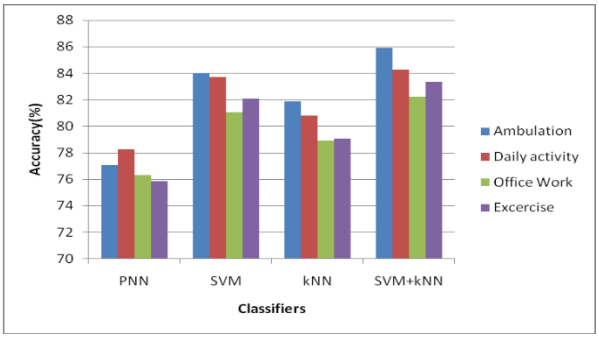


Fig.7: GiST Accuracy graph

Table 1 denotes the accuracy of top level categories obtained with GiST. The accuracy of top level categories with GiST is represented in Figure 7. The above diagrams represent SVM+kNN gives an accuracy of 85.93%, 84.27%, 82.25% and 83.37% are obtained for 4 top level categories as mentioned with GiST. This is high when compared with other classifiers. Hence SVM+kNN classifier has enhanced performance in comparison with other classifiers

Table 2: Accuracy values of top level categories using SIFT

Classifiers / Top Level Categories	PNN	SVM	kNN	SVM +kNN
Mobility	76. 67	83. 32	80.59	**84.45**
Routines	74. 38	82.23	81.38	**83.85**
Office Work	72.69	80.49	78.15	**81.58**
Workout	71.62	79.99	76.47	**80.41**

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

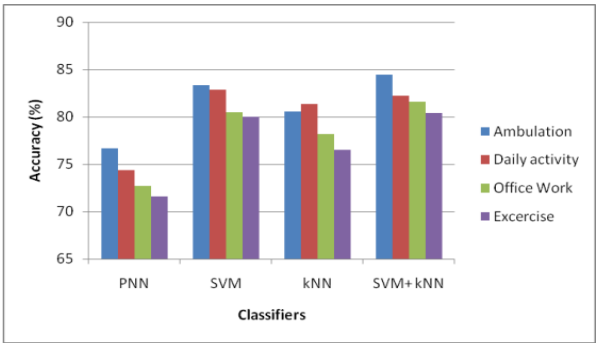


Fig.8: SIFT accuracy graph

Table 2 depicts the accuracy of top level categories with SIFT. Accuracy of top-level categories with SIFT is shown in Figure 8. The above diagrams represent SVM+kNN gives an accuracy of 84.45%, 82.23%, 81.58% , 80.41% for 4 top level categories as mentioned with SIFT. These rates are high when compared with other classifiers. Hence SVM+kNN classifier have enhanced performance in comparison with other classifiers.

Table 3: SURF accuracy table

Classifiers / Top Level Categories	PNN	SVM	kNN	SVM+ kNN
Mobility	74.08	80.26	78.52	**82.35**
Routines	76.50	79.54	77.41	**80.58**
Office Work	74.98	78.64	76.11	**79.84**
Workout	70.84	79.54	74.79	**81.16**

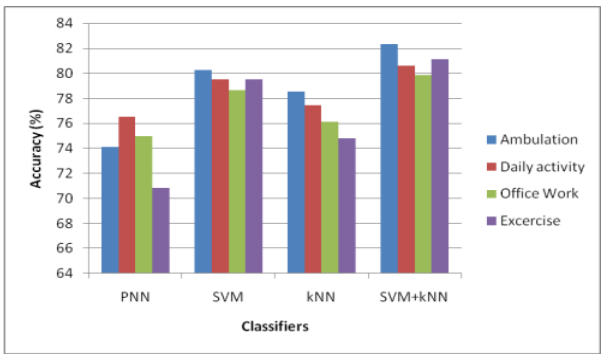


Fig.9: SURF accuracy graph

Table 3 shows the accuracy of various top level categories with SURF. The accuracy of top-level categories with SURF is shown in Figure 9. As per the given references in the form of tables and figures SVM+kNN gives an accuracy of 82.35%, 80.58%, 79.84%, 81.16% for 4 top level categories as mentioned with SURF. These rates are high when compared with other classifiers. Hence SVM+kNN classifier have enhanced performance in comparison with other classifiers.

Table 4: Accuracy for top level categories using STIP

Classifiers / Top Level Categories	PNN	SVM	kNN	SVM+ kNN
Mobility	73.57	79.05	77.93	**81.55**
Routines	75.42	78.13	76.25	**79.23**
Office Work	73.28	77.11	75.54	**78.47**
Workout	69.93	78.37	73.86	**80.91**

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

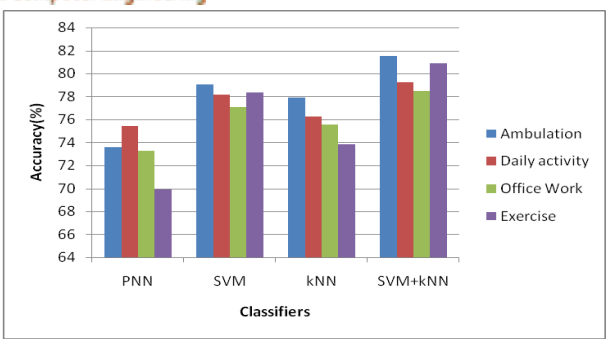


Fig. 10: STIP accuracy chart

Table 4 shows the accuracy of top level categories using STIP. The accuracy of top level categories with STIP is shown in Figure 10. The above diagram demonstrates SVM+kNN gives an accuracy of 81.55%, 79.23%, 78.47%, 80.91% for 4 top level categories as mentioned with STIP. These rates are high when compared with other classifiers. Hence SVM+kNN classifier have enhanced performance in comparison with other classifiers.

Table 5: II level categories accuracy

Classifiers / Features	PNN	SVM	kNN	SVM+ kNN
GiST	72.27	77.52	74.62	**78.43**
SIFT	67.39	72.32	71.75	**73.11**
SURF	67.59	71.08	69.31	**72.77**
STIP	65.87	69.52	66.44	**70.51**

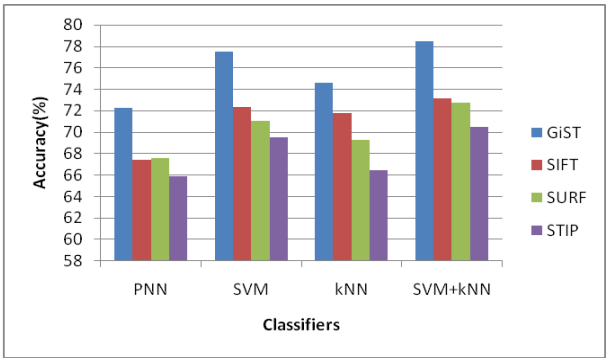


Fig. 11: II level category accuracy chart

The II level categories' accuracy has been depicted in Table 5. Figure 11 depicts the accuracy of 2nd level categories. When used in II level categories, SVM+kNN gives an accuracy of 78.43%, 73.11%, 72.77% and 70.51% for GiST, SIFT, SURF and STIP respectively. These rates are high when compared with other classifiers given in the given tables and graphs. Hence SVM+kNN classifier have enhanced performance in comparison with pre-existing classifiers.

8. CONCLUSION

In the proposed study, SVM+kNN classifier and GiST together gives enhanced outcomes. Support vector machine is advantageous in terms of accuracy and is effective even in the cased of biased samples. As the optimality problem is convex it gives unique solution. K nearest neigbor simple classifier. When both of them are merged as SVM+kNN the benefits of both the classifiers are added together to provide enhanced results in comparison with other classifiers.

9. REFERENCES

[1]. Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, Sei Naito (2014) 'An Attention-based Activity Recognition for Egocentric Video', *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

[2]. H. Pirsiavash and D. Ramanan (2012) 'Detecting activities of daily living in first-person camera views', IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3]. M. Cerf, J. Harel, W. Einhauser, and C. Koch (2007) 'Predicting human gaze using low-level saliency combined with face detection', Advances in Neural Information Processing Systems (NIPS), Vol. 20, pp. 241-248.

[4]. L. Itti, N. Dhavale, F. Pighin(2003) 'Realistic avatar eye and head animation using a neurobiological model of visual attention', SPIE 48th Annual International Symposium on Optical Science and Technology, Vol. 5200, pp. 64-78.

[5]. J. Harel, C. Koch, and P. Perona(2006) 'Graph-based visual saliency', Advances in Neural Information Processing Systems (NIPS), Vol. 19, pp. 545-552.

[6]. C. Koch and S. Ullman (1985) 'Shifts in selective visual attention: towards the underlying neural circuitry', Human Neurobiology, Vol. 4, No. 4, pp. 219-227.

[7]. L. Itti, C. Koch, and E. Neibur (1998)'A model of saliency-based visual attention for rapid scene analysis', IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259.

[8]. T. Avraham and M. Lindenbaum (2010)'Esaliency (extended saliency); Meaningful attention using stochastic image modeling', IEEE Transactions on Pattern Analysis and Machine intelligence (PAMI), Vol. 32, No. 4, pp. 693-708.

[9]. L. F. Coasta (2006) 'Visual saliency and attention as random walks on complex networks', ArXiv Physics e-prints.

[10]. W. Wang, Y. Wang, Q. Huang, and W. Gao (2010) 'Measuring visual saliency by site entropy rate',Computer Vision and Pattern Recognition (CVPR) IEEE, pp. 2368-2375.

[11]. T. Foulsham and G. Underwood (2008) 'What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition', Journal of Vision, Vol. 8, No. 2, pp. 1-17.

[12]. Kumar, A., Suman, S.K., Bhagyalakshmi, L., Sahu, A.K. (2022). IoT and Cloud Network Based Water Quality Monitoring System Using IFTTT Framework. In: Mahajan, V., Chowdhury, A., Padhy, N.P., Lezama, F. (eds) Sustainable Technology and Advanced Computing in Electrical Engineering . Lecture Notes in Electrical Engineering, vol 939. Springer, Singapore. https://doi.org/10.1007/978-981-19-4364-5_3

[13]. Tiwari, R., Shrivastava, R., Vishwakarma, S.K., Suman, S.K., Kumar, S. (2023). InterCloud: Utility-Oriented Federation of Cloud Computing Environments Through Different Application Services. In: Kumar, A., Mozar, S., Haase, J. (eds) Advances in Cognitive Science and Communications. ICCCE 2023. Cognitive Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-8086-2_8

[14]. R. Messing, C. Pal, and H. Kautz (2009) 'Activity recognition using the velocity histories of tracked keypoints', IEEE International Conference on Computer Vision.

[15]. J. Lei, X. Ren, and D. Fox (2012) 'Fine-grained kitchen activity recognition using RGB-D', ACM International Joint Conference on Pervasive and Ubiquitous Computing.

[16]. M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele (2012) 'A database for fine grained activity detection of cooking activities', IEEE Conference on Computer Vision and Pattern Recognition.

[17]. H. Pirsiavash and D. Ramanan (2012)'Detecting activities of daily living in first-person camera views',IEEE Conference on Computer Vision and Pattern Recognition.

[18]. K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato (2012) 'Coupling eye-motion and ego-motion features for first-person activity recognition',CVPR Workshop on Egocentric Vision.

[19]. L. Bhagyalakshmi, S. K. Suman and K. Murugan, "Corona based clustering with mixed routing and data aggregation to avoid energy hole problem in wireless sensor network," *2012 Fourth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2012, pp. 1-8, doi: 10.1109/ICoAC.2012.6416860.

[20]. S. K. Suman, L. Bhagyalakshmi and K. Murugan, "Non cooperative power control game for wireless ad hoc networks," *2012 Fourth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2012, pp. 1-6, doi: 10.1109/ICoAC.2012.6416822.

[21]. Suman S. K., Porselvi, S., Bhagyalakshmi L., and Kumar, D., "Game Theoretical Approach for Improving Throughput Capacity in Wireless Ad Hoc Networks",in proceedings of International Conference on Recent Trends in Information Technology (ICRTIT 2014, MIT Chennai), 10-12 April 2014. 10.1109/ICRTIT.2014.6996152

[22]. M. S. Ryoo and L. Matthies (2013) 'First-person activity recognition: What are they doing to me',IEEE Conference on Computer Vision and Pattern Recognition.

[23]. Y. Poleg, C. Arora, and S. Peleg (2014)'Temporal segmentation of egocentric videos', IEEE Conference on Computer Vision and Pattern Recognition.

[24]. Hari Prasad Bhupathi, Srikiran Chinta, 2024. "Battery Health Monitoring With AI: Creating Predictive Models to Assess Battery Performance and Longevity", ESP Journal of Engineering & Technology Advancements 4(4): 103-112.

[25]. Megha M Pandya, Nehal G Chitaliya, Sandip R Panchal (2013) 'Accurate Image Registration using SURF Algorithm by Increasing the Matching Points of Images', International Journal of Computer Science and Communication Engineering, Vol.2, No.1.

[26]. Mustapha Oujaoura, Brahim Minaoui and Mohammed Fakir (2013) 'Walsh, Texture and GIST Descriptors with Bayesian Networks for Recognition of Tifinagh Characters', International Journal of Computer Applications, Vol.81, No.12.

[27]. Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund and Jordi Gonzalez (2012) 'Selective spatio-temporal interest points', Computer Vision Image Understanding, Elsevier, Vol.116, No.3, pp.396-410.

[28]. N. Jayanthi, S. Indu (2016) 'Comparison of Image Matching Techniques', International Journal of Latest Trends in Engineering and Technology, Vol.7, No.3 pp. 396-401.