International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195217

Recurrent Neural Network for Human Action Recognition using Star Skeletonization

Anantha Prabha P[1], Srimathi R[2], Srividhya R[2], Sowmiya T G[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

[2]UG Scholars, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Human Action Recognition has been an active research topic since early 1980s due to its promising applications in many domains like video indexing, surveillance, gesture recognition, video retrieval and human-computer interactions where the actions in the form of videos or sensor datas are recognized. The extraction of relevant features from the video streams is the most challenging part. With the emergence of advanced artificial intelligence techniques, deep learning methods are adopted to achieve the goal. The proposed system presents a Recurrent Neural Network (RNN) methodology for Human Action Recognition using star skeleton as a representative descriptor of human posture. Star skeleton is the process of jointing the gross contour extremes of a body to its centroid. To use star skeleton as feature for action recognition, the feature is defined as a five-dimensional vector in star fashion because the head and four limbs are usually local extremes of human body. In our project, we assumed an action is composed of a series of star skeletons overtime. Therefore, images expressing human action which are time-sequential are transformed into a feature vector sequence. Then the feature vector sequence must be transformed into symbol sequence so that RNN can model the action. RNN is used because the features extracted are time dependent

Keywords : Human Action Recognition, Video Streams, RNN, Star skeletonization.

I. INTRODUCTION

Human Activity recognition aims to recognize common human activities in real time.Recognizing human activity from video streams is one of the most challenging research topics in computer vision,image processing and pattern recognition fields.The reason for video based activity recognition becoming more sensational in computing field because of its emerging applications viz, Automated surveillance systems in public places like airports and railway stations which requires detection of abnormal and suspicious activities which may cause security threat, as opposed to normal activities. Traditionally, the process of perceiving and analysing human activities demanded human operators, for instance in vault security systems or in patient monitoring systems. However, this manual intervention not only becomes more challenging for the operators, but also exorbitant,in particular, since it needs interminable operation.With the boon of evolving technologies and fields in computer world like Artificial Intelligence(AI), it is possible to reduce the human intervention and increase the automation.

Further, the primary goal of human activity recognition is to automatically analyse on going activities from an unknown video (i.e. a sequence of

image frames). In a simple case where a video file is parted to contain only one execution of a human activity,the objective of the system is to correctly classify the video into its respective activity category. In a more general case, the recognition of human activities must be performed continuously by detecting the starting and ending times of all performing activities from an input video.

II. TYPES OF HUMAN ACTIVITIES

There are various types of human activities. Depending on their complexity, they can be conceptually categorized into four different levels: gestures, actions, interactions, and group activities. Gestures are rudimentary movements of a person's body part, and are the irreducible components describing the meaningful motion of a person. "Stretching an arm" and "shrugging the shoulders" are good examples of gestures, whereas Actions are single-person activities that may be composed of one or more gestures organized temporally, such as "walking" and "jogging". Interactions are human activities that comprises of two or more persons and or objects. For example, "two persons talking" is an interaction between two human beings and "a person stealing a wallet from another" is a human-object interaction involving two humans and one object. Lastly, group activities which involves multiple persons performing the conceptual activity.For instance, "A group of persons dancing" and "a group having a business meeting". In this research, the main focus is given to improve the recognition accuracy of single human activities from real-time video sequences.

Nowadays, more and more people record their daily activities using digital cameras,and this brings the enrichment of video content on the internet, and also causes the problems of categorizing the existing video, and classifying new videos according to the action classes present. Categorizing these videos is a time-consuming task. Our dataset consists of 10 different types of actions. The task are time-consuming, if it is done manually, and recognizing certain actions from scenes of interest in real movies is impossible to accomplish through manual effort. For these reasons, the area of human action recognition has attracted considerable attention. Existing approaches aimed at solving this problem have focused on a pattern recognition system, which is trained using feature descriptors extracted from the training videos, and enables the computer to identify the actions in new videos automatically.

III. RELATED WORKS

In the previous research, a method called Action-Fusion is used to recognize the human action from depth maps and posture data using Convolutional Neural Networks(CNN) using two input descriptors. The first input descriptor is a depth motion image(DMI) and the second input is a moving joints descriptor(MJD).The MJD descriptor represents the motion of body joints with time. Three CNN channels are trained with different inputs. The first channel is trained with depth motion images (DMIs), the second channel is trained with both DMIs and moving joint descriptors together, and the third channel is trained with moving joint descriptors only. The action predictions generated from the three CNN channels are fused together for the final action classification[1].Action recognition has been a challenging problem for many years.Many algorithms were used over time to improve the results.Majority of the techniques focuses on traditional classification algorithms like Naive Bayes (NB) [2], Decision Trees [3], HMM [4], CRF [4], Nearest Neighbor (NN) [5], Support Vector Machines (SVM) [6] and different boosting techniques.

The Naive Bayesian classifier which is a simple probabilistic classifier among the machine learning

techniques produces good accuracy when the data is large but does not model any temporal information.The HMM, HSMM, and CRF are the most popular approaches for including such temporal information. However, these approaches sometimes discard pattern sequences that convey information through the length of intervals between events.The above mentioned facts are the reason for delving deep in the study of Recurrent Neural Networks(RNN) which promises the recognition of patterns that are defined by temporal distance[7].

The proposed system primarily does Human silhouette extraction, contour extraction, followed by star skeletonization, the above three methods extracts the features. Finally, RNN is trained to accurately recognize the human actions.

IV. PROPOSED METHODOLOGY

In comparison with the four stages of human activity recognition, due to variations in environmental factors and differences in actor's activities due to the variation in both environment and actor behaves, action description is still one of the greatest challenges. The changes occuring in the environment include illumination variations, camera view angle difference, and resolution of the image. These changes highly affects the performance of the human activity recognition. Primarily,the variations in the illumination cause serious problems to non-robust background subtraction, and many research approaches fail due to the lighting problem.

Secondly, videos captured under different view angles of the camera appear differently.

Applying the same approach to different videos having different view angles causes deviation in the results. Thirdly, image scales can influence the recognition accuracy and high High-level resolution video needs more computation time and obtains more noise than low resolution video. Since, People look different in different videos and perform similar actions differently, there occurs more difficulty for action recognition because similar actions can be easily classified into two different categories. Difficulty arises in differentiating between some actions such as walking and jogging.

Nowadays, more and more people record their daily activities using digital cameras, and this brings the enrichment of video content on the internet, and also causes the problems of categorizing the existing video, and classifying new videos according to the action classes present.

After reading from different resources following are the different methods, we have implemented in this project of Human Action Recognition. This section explains each steps briefly. First it discusses about the method of extracting human body contour from a frame of the given video. Then it explains about the single star skeleton method to represent the human body extremities which will represent feature vectors forcorresponding training examples while implementing classification process. After that it discusses about the training model implemented for the classification of data.The overall project flow is depicted below in fig(1).

The first target in this project is to extract human body silhouette from given image. To achieve so, we first need to extract the human body from the given frame. We have two types of videos in our data-set. First set of videos are those in which the person stays at the same place and performs the actions like bending, doing jacks, waving with single or double hands, etc(In frame videos). The second set includes those videos in which the human enters the video in one frame and exits from video in coming frames i.e. the whole human body is moving from one place to another (out Frame Videos). So according to the type

of video we used two methods to extract human body from a frame of the video.

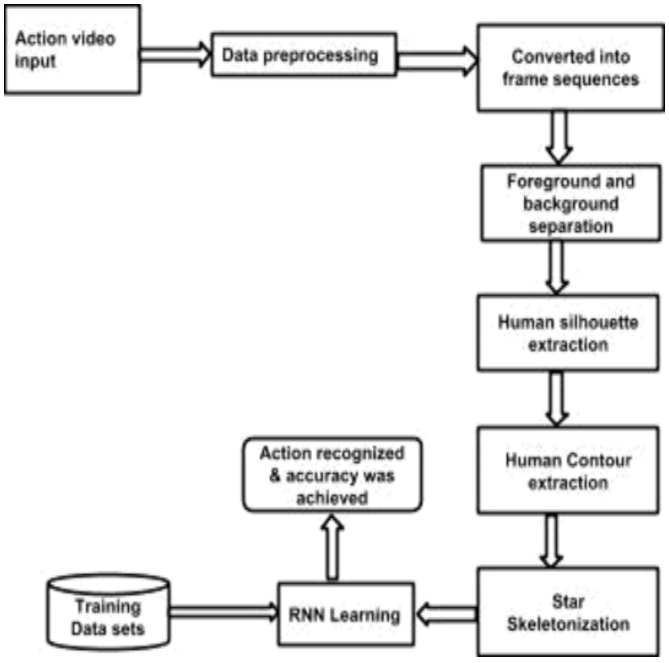

Fig:1.Human Silhouette Extraction

- **In Frame Videos**: For this type of videos direct differencing method was used to extract the human body i.e.direct difference between the background and the current frame was taken to obtain the human body. We had one background image corresponding to each video. The method gave good results for human silhouette extraction.
- **Out Frame Videos**: For this type of videos, inbuilt Gaussian Mixture Model based Foreground detection method was used to extract the human body from frames of the videos.

The color spaces are used for various purposes in image processing domains. Most commonly used color spaces are RGB and HSV because it is directly supported by many scanners.The RGB space which stands for Red,Green and Blue is converted into HSV (Hue, Saturation, and Value) and HLS (Hue, Lightness, and Saturation) color models respectively.The HLS is used to estimate the background of an image,where the H is a hue value in HSV space, and L and S components are defined as Lightness and Saturation values in modified HLS space. The conversion process from RGB color space to modified HLS space is given by,

$$m=\min(g,b)$$

$$H=\begin{cases} -1(\text{undefined}) & \text{if } r=m \\ 60\times(b-r)/(m-r) & \\ \text{ifg}=60\times(r-g)/(m-r), & \text{if } b=m \\ H+360 & \text{,if } H<0 \end{cases}$$

$$L=(m+r)/2$$

$$s=\begin{cases} 0 & \text{if } m=r((m-r)/(m+r)) \\ \text{,if } L\leq 0.5\ ((m-r)/(2-m-r)) & \text{otherwise} \end{cases}$$

where r, g, and b stands to the normalized RGB ranges from 0 to 1. The modified HLS space is slightly more robust to noise but very similar to original HLS space . To allow controlled lighting conditions, the chroma-key laboratory is used here and subject is captured against a uniform background.

$$\xi = {}^{\text{mod}}x\in m,y\in n\ ({}^{h}x,y)$$

where $h_{x,y}$ is defined as the hue value present at coordinate (x, y) in m×n region of the image. In this data, a better estimation of hue component i is offered by HSV space.The conversion is also done from RGB color space to HLS color space.

$$\psi = [\sigma L\ (i)\ ,\ \sigma S\ (i)]$$

The above formula defines the range of lightness and saturation components in the HLS space. here σ describes variance of each color component. The background feature (ψ) can be calculated by color clustering method using lightness (L) and saturation (S) components. The background features can be removed by using Equation 2 and 3, namely, the pixel values in HLS space can be re-defined by background estimation as

H, L, S = {0 if ($p_{k-H} \subset \xi$)Λ($p_{k-LS} \subset \psi$) p_k
otherwise

where p_{k-H} is H component of pixel k in HSV space, and p_{k-LS} is L and S components of pixel k in modified HLS space.

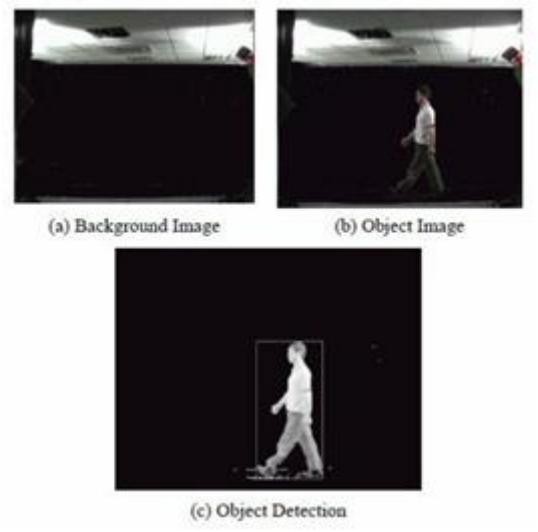

Figure 2. Background Subtraction

As seen in the figure, the object image still has some noise (background components), and the object has also lost some foreground components. Noise filtering and histogram logarithm methods are applied to the background subtracted image to remove this noise and to recover the lost components, . The noise filter is defined by

$$g(x,y) = \{g(x,y), if \sum_{k=1}^{} \sum_{j=1} f(x-j,y-k)|_{(j,k)\neq(x-j,y-k)} > 2$$

If f(x,y) is greater than a threshold value Tn, then

f(x, y) is set to 1 otherwise it is set to 0. Histogram logarithm is defined as

$p'k = c \times \log\log(1 + pk)$

where pk describes the pixel value at index k, and c is a constant. To enhance detail in the darker region of the image, histogram logarithm is used which increases the dynamic range of grayscale via contrast stretching. However, histogram projection method can be used to detect the object for grayscale images.

B.Human Contour Extraction

In the process of extracting the contour of a detected human body from the video stream, a thresholding and morphological method is used. Fig.3 shows the depiction of thresholding and extraction of a human body contour. Thresholding is used because it is one of the most important methods in the field of image segmentation, and under irregular illumination,choosing a correct threshold is difficult. However, improvement can be done using the background information in an image. Accordingly, a thresholding method which is based on similarity (or dissimilarity) measures between the the object image and the background is used. Let Ib(x, y) and Io(x, y) be the feature (or brightness in grayscale image) of a pixel with coordinate (x, y) in the background image (Ib) and the object image (Io). The below formula is used to compute the similarity Θ(x, y) at coordinate (x, y) :

$$\Theta(x,y) = |I_b(x,y) - I_0(x,y)|$$

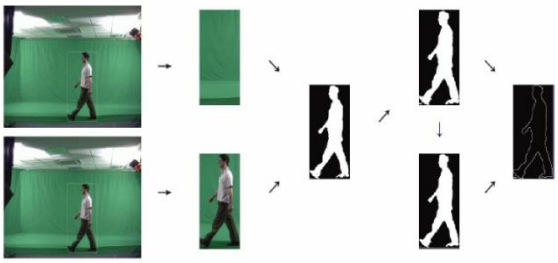

Figure 3. Extraction of Human Body Contour

gx,y={gx,y,i k=11j=11fx-j,y-k|(j,k)≠(x-j,y-k)>2
otherwise gx,y={gx,y,i
k=11j=11fx-j,y-k|(j,k)≠(x-j,y-k)>2
otherwise

Similarity values nearing to zero imply a high probability of being background. In Converse, greater values of similarity imply high probability as an object. Therefore, the binary image I_{bi} (x, y) is thresholded as

$$Ibix,y = \begin{cases} 1, & \text{if}((x,y)>\tau)\wedge(I0(x,y)>\lambda) \\ 0, & \text{otherwise} \end{cases}$$

By theory, thresholding is a very simple image segmentation method, which is very effective and greatly used for small and low-resolution images, but suffers from difficulty due to the change in illumination. To overcome the problem and to improve this method, a more effective algorithm using the probability density of the similarity for determining appropriate values of and λ is required.

Adding to it, morphological filtering is used to remove the noise and to extract the human body contour, by the dilation and erosion because the binary image can have some noise inside the object which is nothing but a human body part and some noise outside the object. Mathematically, the dilation of a set A by a structuring element B is denoted by A $\oplus$B and is defined as

$$A \oplus B = \{x | B_x \cap A \neq \phi\}$$

Ultimately, the human body contour is obtained by subtracting the dilation and erosion images using the formula,

$$C = (A \oplus B) - (A \ominus B)$$

If dilation adds pixels to an object or thickenns, then erosion will make an image smaller or performs thinning. In a nutshell,binary erosion will remove the outer layer of pixels from an object. object contour can be obtained easily just by using the subtraction operation of the dilation and erosion image.

C. star skeletonisation

The concept of star skeleton is to connect from centroid to gross extremities of a human contour. To find the gross extremities of human contour, the distances from the centroid to each border point are processed in a clockwise or counter -clockwise order. Extremities can be located in representative local maximum of the distance function. Since noise increases the difficulty of locating gross extremes, the distance signal must be smoothed by using smoothing filter or low pass filter in the frequency domain. Local maximum are detected by finding zero-crossings of the smoothed difference function. The star skeleton is constructed by connecting these points to the target centroid.

The algorithm takes human contour as input and produces a skeleton in star fashion as the output of the algorithm.

- Determine the centroid of the target image border.
- Calculate the distances from the centroid to each border point.
- Smooth the distance signal for noise reduction by using linear smoothing filter or low pass filter in the frequency domain.
- Take local maximum of the distance signal as extremal points, and construct the star skeleton by connecting them to the centroid. Local maximum are detected by finding zero-crossings of the difference function.

As a feature, the dimension of the star skeleton must be fixed. The feature vector is then defined as five-dimensional vectors from centroid to shape extremes because head, two hands, two legs are usually local maximum. The final cut-off frequency of star skeleton is determined automatically. The cut-off frequency is first set to a higher frequency, and gradually decreases until the dimension of star skeleton is within five. For postures with more than five contour extremes, we adjust the low pass filter to lower the dimension of star skeleton to five. On the other hand, zero vectors are added for postures with less than five extremes. Since the used feature is

vector, its absolute value varies for people with different size and shape, normalization must be made to get relative distribution of the feature vector. This can be achieved by dividing vectors on x-coordinate by human width, vectors on y-coordinate by human height.

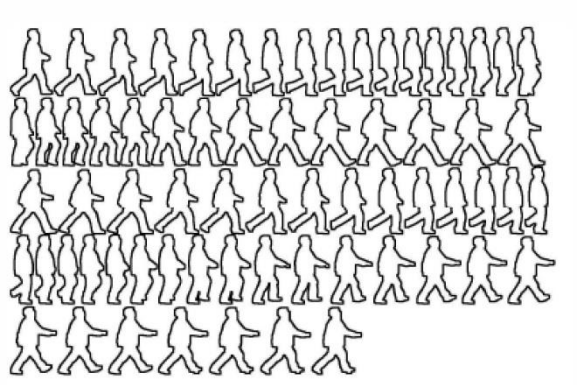

Fig: 4 A walk action is a series of postures over time

The concept of star skeleton is to connect from centroid to gross extremities of a human contour. To find the gross extremities of human contour, the distances from the centroid to each border point are processed in a clockwise or counter-clockwise order. Extremities can be located in representative local maximum of the distance function. Since noise increases the difficulty of locating gross extremes, the distance signal must be smoothed by using smoothing filter or low pass filter in the frequency domain. Local maximum is detected by finding zero-crossings of the smoothed difference function. The star skeleton is constructed by connecting these points to the target centroid.

Star skeleton Algorithm

Input: Human contour
Output: A skeleton in star fashion

1. Determine the centroid of the target image border (x_c, y_c)

$$x_c = \frac{1}{N_b} \sum_{i=1}^{N_b} x_i$$

$$y_c = \frac{1}{N_b} \sum_{i=1}^{N_b} y_i$$

$$y_c = \frac{1}{N_b} \sum_{i=1}^{N_b} y_i$$

where N_b is the number of border pixels, and (x_c, y_c) is a pixel on the border of the target.

2. Calculate the distances d_i from the centroid (x_c, y_c) to each border point

$$(x_c, y_c) \quad d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$$

These are expressed as a one-dimensional discrete function $d(i) = d_i$

2. Smooth the distance signal d (i) to d(i) for noise reduction by using linear smoothing filter or low pass filter in the frequency domain.

3. Take local maximum of d(i) as extremal points, and construct the star skeleton by connecting them to the centroid (x_c, y_c) . Local maximum is detected by finding zero-crossings of the difference function.

$$\delta(i) = d(i) - d(i-1)$$

The star skeleton of each videos will be used as its features for training the model. There will good number of frames for each video, therefore it will be difficult to use the whole skeleton image of each frame as the feature of a single video.We will reduce the number of features representing a video. We have got five points for each frame which will be used as features for that frame. Directly we cannot use these points as the features so the angle of line joining the

points from the centroid is calculated and used these angles as the features for that frame. Also, the distance of the centroid from one of the vertical boundaries of the image to keep track of the position of human body from center of the image, which will help us to differentiate from p-jump action (in place jump) and jump action (jumping from end of the image to another end of the image). In this way every frame is represented with 6-dimensional vector.

D. KEMA(Kernel Manifold Alignment)AND RBF KERNELS:

The KEMA method aims to construct domain-specific projection functions, to project the data in Hilbert space from all domains to a new common latent space, on which the preservation is done on instances topology of each domain. The instances which belongs to the same classes will locate nearly, and the ones belonging different classes will be distant from each other. To perform so, KEMA desires to find a data projection matrix that minimizes the following cost function where TOP, SIM, and DIS denote the topology, class similarity, and class dissimilarity, respectively.In KEMA, the RBF kernels were used with the bandwidth fixed as half the distance of the median between the samples of the particular video (labeled and unlabeled), where **radial basis function kernel**, or **RBF kernel**, is a popular kernel function used in various pattern analysis learning algorithms.By doing so, different kernels are allowed in each domain, thus tailoring the similarity function to the data structure observed.

V. V.DATA SET

The data set consists of 10 videos representing each action.The actions are Bend, Jack, Jump in place, Waving with one hand, Waving with two hand, Run, Walk, Side Run, Skip, Jumping from one end of image to another. Each video runs for about 3-6 seconds in time.

VI. IMPLEMENTATION AND RESULTS

We used different methods for silhouette extraction. First method uses the thresholding on the saturation part of the HSV image. As the number of epochs are increased the model tries to overfit the training data hence the accuracy on testing data reduces. The figure shows the confusion matrix for the RNN model trained.We build a visible light action dataset, named MSR3D ACTION, following the approach to construct an action recognition dataset.

In all experiments, each of the dataset is randomly categorized into training and testing sets. The average precision (AP) is used for evaluating purposes, which is the average of recognition precisions of all actions. For each evaluation,the experiments are repeated with the same setting 5 times and the average accuracy is reported.In KEMA, RBF kernels are used with the bandwidth fixed as half of the median distance between the samples of the specific video (labeled and unlabeled).so that, different kernels are allowed in each domain, thus tailoring the similarity function to the data structure observed

Table :1 Comparison of proposed for MSR3D dataset in terms of algorithm and iterations

Algorithm	No.of iterations
RNN	
Fold-1	636
Fold-2	975
Fold-3	753
Fold-4	854
Fold-5	1043

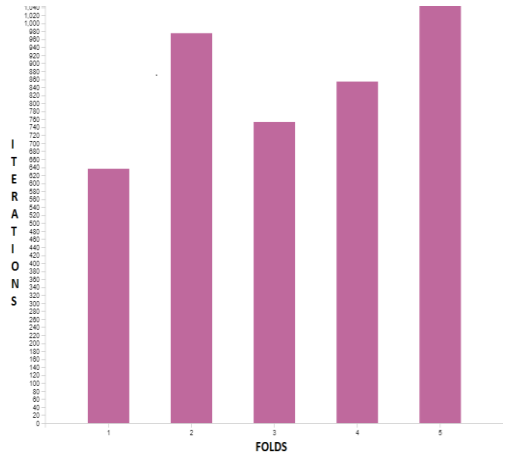

Fig 5:graphical representation of the table1

Table : 2 Comparison of proposed method with the existing methods for MSR3D dataset in terms of iterations and training data.

No. of iterations	Train(min)
636	5.32
975	8.43
753	4.65
854	10.32
1043	7.54

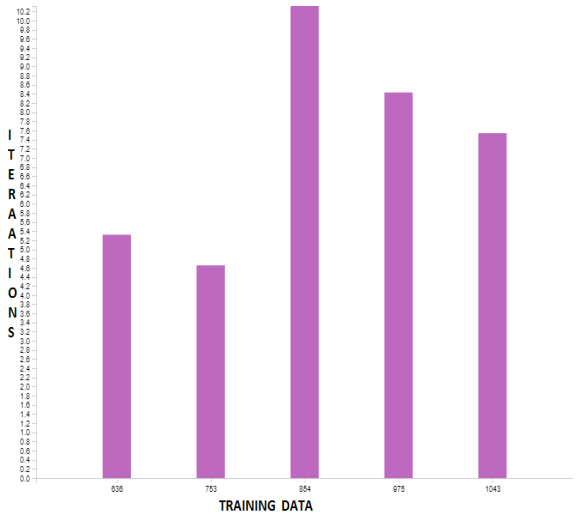

Fig 6:graphical representation of the table2

Table :3 Comparison of proposed method in terms of iterations and testing data time in seconds.

No. of iterations	Test (sec)
636	0.54
975	1.43
753	2.65
854	1.54
1043	6.43

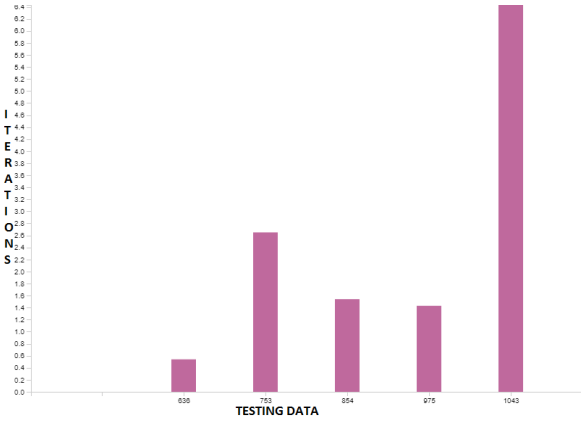

Fig 7 : Comparison of proposed method with the existing methods for space time shape dataset.

VII. REFERENCES

[1]. Aouaidjia Kamel, Bin Sheng ,Bin Sheng,Po Yang, Ping Li,Ruimin Shen, and David Dagan Feng, "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures", IEEE transactions on system,man and cybernetics systems.

[2]. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) Pervasive 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004). doi:10.1007/ 978-3-540-24646-6 10

[3]. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) Pervasive 2004. LNCS,

vol. 3001, pp. 1–17. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24646-6 1 14.

[4]. Van Kasteren, T., Noulas, A., Englebienne, G., Kr¨ose, B.: Accurate activity recognition in a home setting. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 1–9. ACM (2008).

[5]. Wu, W., Dasgupta, S., Ramirez, E.E., Peterson, C., Norman, G.J.: Classification accuracies of physical activities using smartphone motion sensors. J. Med. Internet Res. 14, e130 (2012).

[6]. Zhu, Y., Nayak, N.M., Roy-Chowdhury, A.K.: Context-aware activity recognition and anomaly detection in video. IEEE J. Sel. Top. Sig. Proces. 7, 91–101 (2013).

[7]. Duckworth, M. Alomari, Y. Gatsoulis, D. C. Hogg, and A. G. Cohn, “Unsupervised activity recognition using latent semantic analysis on a mobile robot,” in Proc. Eur. Conf.Artif. Intell., 2016, pp. 1062–1070.

[8]. M. H. Kolekar and D. P. Dash, “Hidden Markov model based human activity recognition using shape and optical flow based features,” in Proc. IEEE Region 10 Conf.,2016, pp. 393–397.

[9]. G. Liang, X. Lan, J. Wang, J. Wang, and N. Zheng, “A limb-based graphical model for human pose estimation,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 48, no. 7, pp.1080–1092, Jul. 2018

[10]. J. Sriwan and W. Suntiamorntut, “Human activity monitoring system based on WSNs,” in Proc. Int. Joint Conf. Comput. Sci. Softw. Eng., 2015, pp. 247–250.

[11]. Y. Guo, D. Tao, W. Liu, and J. Cheng, “Multiview Cauchy estimator feature embedding

Cite this article as :

Anantha Prabha P, Srimathi R, Srividhya R2, Sowmiya T G, "Recurrent Neural Network for Human Action Recognition using Star Skeletonization", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 335-344, March-April 2019.

Available at doi :

https://doi.org/10.32628/CSEIT195217

Journal URL : http://ijsrcseit.com/CSEIT195217