International Journal on Emerging Technologies **11**(3): 1033-1040(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Dyadic Human Interaction Recognition from Videos using Multi-layer 3D CNN

Subetha T.[1], Chitrakala S.[2] and Uday Theja M.[3]
[1]*Assistant Professor, Department of Information and Technology, BVRITH, Hyderabad, India.*
[2]*Professor, Department of Computer Science and Engineering, CEG Anna University, India.*
[3]*Department of Computer Science and Engineering, IITM, Chennai, India.*

(Corresponding author: Subetha T.)
(Received 24 March 2020, Revised 27 May 2020, Accepted 28 May 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Human Activity Recognition aims in recognizing and interpreting the activities of human automatically from videos. In general, the major issues of Human Activity Recognition are dynamic backgrounds, similar action recognition, dyadic human interaction recognition, incremental learning etc. Among these issues, identifying the interactions of human within minimal computation time and reduced misclassification rate is a cumbersome task. Hence, a Dyadic Human Interaction Recognition system is proposed in this paper that utilizes a multi-layer 3D CNN and a kernel-based rv-tSNE transformation algorithm for better classification. Preprocessing techniques like gray scale conversion, pixel normalization, and one-hot encoding have been deployed to handle enormous amount of irregularities in the dataset. The model is trained using multi-layer 3D Convolution Neural Networks (CNN). Additionally, a transformation model named kernel-based rv-tSNE has been incorporated for dimensionality reduction and easier visualization of the data. The requirement of the transformation module is justified by comparing the end results of the system with and without the transformation algorithm. The results demonstrate that the proposed system recognizes the interactions of human with reduced misclassification rate and minimal processing time compared to the benchmarking datasets for various interactions. The proposed system achieves 94.78% and 93.394% for SBU kinect interaction dataset and for the newly recorded AU-Interaction dataset respectively .The proposed system finds its applications in sports event analysis, video surveillance, content-based video retrieval, robotics and others.

Keywords: Human Activity Recognition; Convolution Neural Networks; multi-layer3D Convolution Neural Networks; Stochastic Neighbor Embedding; t-Stochastic Neighbor Embedding; kernel-based rv-tSNE

Abbreviations: HAR, Human Activity Recognition; CNN, Convolution Neural Networks; DHIR, Dyadic Human Interaction Recognition; RNN, Recurrent Neural Network; DBN, Deep Belief Network; t-SNE, t-Stochastic Neighbor Embedding; rv-tsne, reduced variantt-Stochastic Neighbor Embedding;

I. INTRODUCTION

In recent times, it has been found that Human Activity Recognition (HAR) is playing an indispensable role in various essential domains like surveillance, video annotation, video indexing, and robotics [1]. The intention of activity recognition is an automated interpretation of ongoing events that are captured in a video. Understanding and identifying the activities of human enable us to extend the system to many important applications such as robots training, abnormal event recognition, vandalism detection, fall detection, and so on. The automatic detection of human activity from the video is not limited to detecting the abnormal activity in surveillance, ATM fraud detection, and abnormal crowd behavior. It can be applied to even behavioral biometrics that involves understanding methods and their algorithms to identify he humans uniquely based on their behavioralcues.

In general, an activity recognition system analyzes the extracted key frames to learn about the paradigm for each activity and uses these patterns for recognizing the activity during testing. The early handcrafted features [2] [3] make use of shape features, texture features, spatio-temporal features, silhouette features, and given to the classifier for training with the features obtained. Though the features can be extracted automatically, they are dataset dependent and cannot be applied to real-world problems. Thus, deep learning based approaches can be adopted for better solution.

Deep learning [4, 5] is arising as a family of learning models in recent times due its ability of learning highly discriminative features. The most popular learning models are Recurrent Neural Network (RNN) [6-8], Convolutional Neural Network(CNN) [9], Deep Belief Network (DBN) [10, 11], and so on. The shifting paradigm from handcrafted features to deep learning features led to various advantages such as pulling out high-level features, learning complex information from raw data without much preprocessing. It is gaining much attention nowadays, as it takes advantage from unlabeled data, building distributed representation, and learning high-level complex data. The main challenge in incorporating deep learning technology is its complex hyper parameters, local minima, stochastic gradient descent performance, over fitting, and massive amount of training data [12]. Hence in this paper, the deep learning based approach is incorporated by overcoming these challenges.

As hierarchical level approaches have been exhibited to outperform the single-level approach, the proposed system utilizes the hierarchical level approach. In order

to provide a robust framework of the Dyadic Human Interaction Recognition (DHIR) system that can recognize the out-of-sample interactions and to reduce the processing time, a multi layer 3D CNN with kernel-based rv-tSNE architecture is proposed. The proposed system takes video as input and performs preprocessing techniques like grey scale conversion, pixel normalization, and one-hot encoding. The preprocessed input is given to the proposed multi-layer 3DCNN for recognizing the interactions. The robustness of the proposed system is validated by creating and testing using a new dataset named AU-Interaction dataset.

The main contribution of this paper is the development of multi-layer 3D CNN that automatically takes in the spatio-temporal orientation of two humans and classifies their actions as activities and further more classifies it as interaction in one go, optimizing the space and time complexity by scaling down the dimensionality of frames using kernel based rv-tSNE.

This paper is structured as follows. Section II depicts the survey of previous works. Section III discusses the complete system development with a detailed description of the algorithms used along with implementation details. Section IV highlights the results and experimental evaluation, which is followed by the conclusions drawn from the work besides exploring future work in Section V.

II. RELATED WORKS

Several HAR systems are proposed for CNN as well as deep learning. In the existing architecture for traditional neural networks, each input layer is mapped to all the hidden layers and the hidden layer maps to the output layer. Here, the image is made into one-dimensional array and fed as input for the neural network. The limitation is that it cannot incorporate spatial dimensions. The interconnection between all nodes make the network computationally expensive. Convolutional Neural Network [9] can solve this problem by using convolution and pooling layers. Image classification (2012) [14] is the first achievement of deep learning based CNN in computer vision. An efficient model is trained with 1.2 million images to group into 1000labels and it is two-dimensional which cannot be used for videos, as it has an additional z-axis named temporal axis. So 3d-CNN is availed where three dimensional kernel is convolved transversely in all three axis. However, the research on video-based HAR is not explored much due to challenges in processing temporal information from the video stream. This section shows how different methodologies are incorporated in video-based HAR based on processing color videos [4].

Discriminative features are extracted from the raw inertial data, which is a critical and challenging task for HAR [15].Most of the existing work depends on heuristic handcrafted features, which are also known as shallow features .Ensemble classification methods [16], which blend multiple learning algorithms, can achieve better ecognition rate. Jalal *et al.,* [17] and Lin *et al.*, [18], for example, combine decision trees, multilayer perceptron, and logistic regression for HAR.

As 2D CNN [19] yields either spatial or temporal features, 3D CNN [20] makes use of both spatiotemporal features. But, it requires large number of training samples. In the CNN based method [19], the spatial and temporal information are captured in separate CNN and fused to recognize the actions. This system does not take pooling into account. Mo et al. [21] applied additional maxpooling layers after feature detection from the raw input to produce scale invariant features, which is then introduced to a 1024 neuron hidden layer to merge features from multiple channels, and another additional soft-max layer to generate the classification result. Baccouche *et al.*, [22] and Chen *et al.* [23] both use CNNs with multiple iterations of convolution and subsampling layers, or convolution and pooling layers being applied for feature extraction. Some authors have claimed that ensemble learning [39] for feature selection will increase the accuracy.

Temporal 3D ConvNet [24] is employed to make use of the temporal cues for action recognition. The computational complexity of two-stream CNN [21] is reduced by incorporating the motion vector [25] in video stream instead of optical flow and then it is trained using CNN for recognizing the actions. A dense correspondences between RGB image and a surface-based representation [26] of the human body is established and then trained the system using CNN for final recognition. Recognizing action is performed by intricately fusing CNN and LSTM [27]. A deep learning framework is developed with temporal scale invariant features [28] by drawing out spatial and motion features with the help of two CNN. The convolutional fusion layer reveals information about the association between features. Action recognition is performed by combining the prediction score and linear weight summation of LSTM network. A deeply coupled ConvNet [39] for human activity recognition that utilizes the RGB frames at the top layer with bi-directional long short-term memory (Bi-LSTM) is developed for better recognition. Skeleton analysis and RGB data streams [40] is combined and processed together to improve the recognition rate.

A three-stream CNN [29] is developed with sequential deep trajectory descriptor for recognizing actions. Here, CNN is employed to identify spatial features and LSTM for temporal features. Gowda [10] demonstrates activity recognition using a method based on DBNs. A deep hybrid feature model [11] is dealt using unsupervised training. This takes both local and deep feature models. Semisupervised learning is combined with active learning to reduce the manual class labeling of incoming videos. The main advantage is that it handles continuous video streams. A combination of deep belief network, which combines a modified version of weber descriptor and local binary patterns (LBP) descriptor is used by Gowda [10]. The extracted features are fed into CNN for labeling the actions. Human activity from continuously streaming videos is detected by binding deep learning networks and active learning. Multi-Modality Multi-Task Recurrent Neural Network (MM-MT RNN) [30] incorporates both RGB and Skeleton networks to recognize the actions. Two-Level Fusion Strategy [31] is employed to combine features from high

level handcrafted strategy and machine learning techniques to address the problem of large variety of actions.

III. PROPOSED DHIR SYSTEM

The survey results in the succeeding issues.
•Though state-of-art techniques possess higher accuracy, they exhaust a large amount of computational resources.
• The prevailing techniques use high-end handcrafted features to build a model, which will be extremely difficult to carry out in the real- world.
• The prevalent transformation algorithms could not resolve data variance, data crowding, and data discrimination problems precisely.
• There is a sharp difference in error that occurred in the training data set and the error encountered in anew unseen data set.
To resolve the above aforementioned issues, a Dyadic-Human Interaction Recognition (DHIR) system is developed by scaling down the dimensionality of frames using kernel-based rv-tSNE and also flawlessly predicting the interactions of human by utilizing the principle of deep learning. The initial preprocessing consists of key frame extraction, gray scale con version, pixel normalization, and one-hot encoding. The preprocessed image is given to the multilayer 3D CNN for model construction. The major issues in deep learning are the computational time needed for training and to know what is going on inside the model. This can be solved by giving the feature map to the transformation module before the dense layer for ease of computation and better visualization. The entire system is made adaptive to handle the data, that is, out-of-sample training by extensively tuning the parameters.

A. Preprocessing module

In the preprocessing module, video is given as input and it does key frame extraction, grayscale conversion, pixel normalization, and one-hot encoding. Initially, the frames are extracted from the video using set fps. Then on the extracted frames, grayscale conversion [32] is done to convert the number of channels from 3 to 1 for easier processing and less complexity.

Fig. 1. Pre-processing Module.

The gray scale converted frames are then resized by pixel normalization [33], as all the frames are in different dimensions and it is required to maintain same dimensions. Finally, one-hot encoding is performed and it is passed into multi-layer 3D CNN and the output of preprocessing module is shown in Fig. 1.

B. Multi-layer 3D CNN model construction

CNN is best suited for video recognition tasks. Hence, a multi-layer 3D CNN is designed and constructed in this paper. The 3D convolution formula is defined as shown in equation1.

$$b_{abt} = f(\sum_i \sum_j \sum_k w_{ijk}\, v_{(a+i)(b+j)(c+t)} + e \qquad (1)$$

where b_{abt} is a feature map value at (a,b,t), f is the activation function, w_{ijk} is kernel weight and v (a+i)(b+j)(t+k) is an input value at (a+i, b+j, t+k).This results in learning in the particular spatial region.3D CNN utilize both spatial and temporal features, whereas multi-layer 3D CNN, after the first convolutional layer, only the spatial dimensions are downsized retaining the temporal information for the deeper layers. This is done so the intrinsic movement of the human is captured correctly thus resulting in high recognition accuracy.

Network architecture: Designing neural network is time consuming and difficult because of its complex set of hyper parameters and the synergy between them. The optimal network architecture may be different for different datasets and even for different subsets. Thus, the architecture is selected based on own experiments with data in hand.

Base model network architecture: The network architecture for the base model consists convolution, maxpooling and softmax classification layer. These convolution layers are generated with 16, 64, and 256 filters respectively. The kernel size is chosen as 5 × 1 × 1 and 3 × 1 × 1. Small kernels are chosen as it is proven efficient with deep architectures. Each of the convolution is followed by a max pooling layer to reduce the spatial and temporal dimensions to half. The graph constructed for the base model as shown in Fig. 2 clearly depicts the overfitting problem and there is a need to overcome this by enhancing the model, which is discussed below.

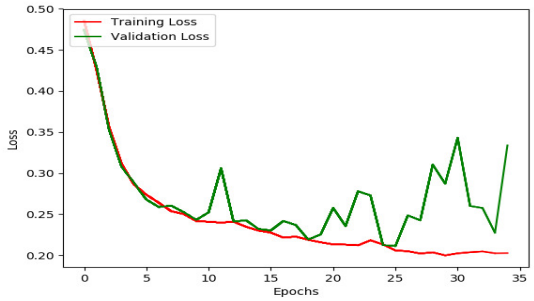

Fig. 2. Base model learning graph.

Multi-layer 3D CNN Model: The network architecture of the multi-layer 3D CNN Model consists of convolution layer, statistics pooling, kernel-based rv-tSNE and softmax classification layer. Since the base model designed was observed to be overfitted and having high-dimensional data points, which are difficult for

visualization, a multi-layer 3D CNN with transformation using kernel-based rv-tSNE is proposed.

In model construction, after convolution layer, statistics pooling is performed with the following equation 2, equation 3, and equation 4.

$$\mu_U = mean(\sum_n \alpha_n^s V^n) \quad (2)$$

$$\sigma_U = std(\sum_n \alpha_n^s V^n) \quad (3)$$

$$I_u = concatenate(\mu_U, \sigma_U) \quad (4)$$

Where μ_u, σ_u, and I_u are the mean, standard deviation and final statistics pooling output vector. The vector is flattened and it is given to kernel-based rv-tSNE which is explained in detail in below section. Dropout [34] is adopted to hinder overfitting and transformation is applied after a dense layer. In Dropout, a part of the neurons are withdrawn from active service. This causes the networks to utilize and rejuvenate the weights of residue neurons. The dropout is applied to the fully connected layers. In addition, it incorporates NADAM as an optimizer in place of ADAM. The default values in keras for ADAM optimizer learning rate is 0.001, whereas it is 0.002 for NADAM and also there is a scheduled decrease in the learning rate. The learning model obtained for both SBU kinect interaction dataset and AU Interaction dataset is shown in Fig. 3.

kernel-based rv-tSNE: As fully connected layers have hundreds of neurons, the dimensionality reduction technique is adopted to visualize these features. The major issues that affect the performance of the system are computational complexity, data variance, data discrimination, and data crowding problem. Hence, a kernel-based rv-tSNE is proposed to overcome these issues. The videos are passed through multi-layer 3D CNN and the output is extracted before the dense layers of the network architecture and utilized for visualization.

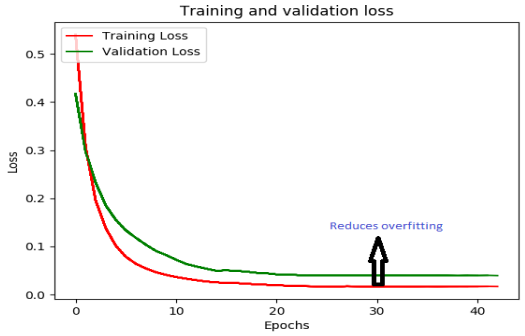

Fig. 3. Proposed model learning graph.

SNE[35] is a leading dimensionality reduction algorithm that utilizes pairwise Euclidean distance between data points for converting higher dimensional data points to lower dimensional data points using conditional probability distribution. The higher dimensional representation is calculated as equation 5.

$$p(j|i) = \frac{\exp(-||a_i - a_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||a_i - a_j||^2 / 2\sigma_i^2)} \quad (5)$$

Where σ_i^2 is the Gaussian variance. The low dimensional equation is calculated as in equation 6.

$$q(j|i) = \frac{\exp(-||b_i - b_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||b_i - b_k||^2 / 2\sigma_i^2)} \quad (6)$$

SNE computes the optimum low-dimensional B by diminishing C(B) and it is given by equation 7.

$$C(B) = \sum_i KL(P_i||Q_i) = \sum_{i,j} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad (7)$$

But the computational cost of SNE is high and it is negatively impacted by the crowding problem. This problem is alleviated by another variant of SNE named t-SNE [36]. It utilizes the joint probability distribution in-contrast to SNE as equation 8.

$$p_{ij} = \frac{\exp(-||a_i - a_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||a_k - a_l||^2 / 2\sigma_i^2)} \quad (8)$$

It makes use of student t-distribution to model b_i to b_j and is given by equation 9.

$$q_{ij} = \frac{(1+||b_i - b_j||^2)^{-1}}{\sum_{k \neq i} (1+||b_k - b_l||^2)^{-1}} \quad (9)$$

By utilizing t-SNE, the crowding problem is eliminated, but the variance acquired after the distribution seemed to be high that leads to the data discrimination problem. rv-tSNE [37] solves this problem, but fails in capturing the intrinsic structure of data and non-linear projection of data. The divergence of the data points is solved by introducing a new algorithm named kernel-based rv-tSNE. The proposed algorithm 1 captures the intrinsic structure of data and non-linear projection of data. The computational complexity is further reduced using this algorithm.

A kernel is picked and a covariance matrix is constructed using the following equations 10, 11, and 12.

$$V_{ij} = E((c_{ij} - \mu)(c_{ij} - \mu)^T) \quad (10)$$

$$\left[\frac{1}{n}\sum_{i=1}^n (a_i - \mu)(a_i - \mu)^T\right].v = \lambda.v \quad (11)$$

$$v = \sum_{i=1}^n \alpha_i (\alpha_i) \quad (12)$$

where $\alpha_i$is Nx1 matrix. By utilizing the kernel function and re-arranging we obtain equation 13.

$$K^2 \alpha_j = n\lambda_j K \alpha_j \quad (13)$$

where k is NxN kernel matrix. Then, a normalized kernel matrix is performed on the data using equation 14.

$$K = k - 2_{\frac{1}{n}} K + 1_{\frac{1}{n}} K 1_{\frac{1}{n}} \quad (14)$$

Next eigenvalue is obtained by orthogonal basis from the sample $a_1, a_2, a_3, \ldots a_n$ as in equation 15.

$$K\alpha_i = \lambda_i \alpha_i \quad (15)$$

This is transformed to low data points and is given as equation 16.

$$b_j = \sum_{i=1}^n \alpha_{ij}\, k(a, a_i) \quad (16)$$

where j=1,2,3,.....d. The low transformed points are then given to the fully connected layer which is then mapped into the softmax classification layer for recognizing the final interaction label. The predicted outputs for the interactions class label is shown in Fig. 4.

Algorithm 1: Proposed Kernel-based rv-tSNE

Input: High-dimensional descriptor points in each frames $X = (x_1, x_2, x_3, \ldots x_n)$; Each frame's number $n(x_i)$ and class label $c(x_i)$; Low-dimensionality d; and Iteration parameters: Iterations *t*, learning rate η, the momentum α(t);kernel K

Output: The corresponding frame's co-ordinate $Y(0) = (y_1, y_2, y_3, \ldots y_n)$ in a lower-dimensional space

Sample initial low dimensional data as $Y(0) = (y_1, y_2, y_3, \ldots y_n)$ where $N \epsilon 0, 10^{-4}$

Iterate for all the points

for i= 1 to N do

Compute the joint probability Pi,j , from equation (8)
Covariance matrix is constructed from

$$V_{ij} = E\left(\left(c_{ij} - \mu\right)\left(c_{ij} - \mu\right)^T\right)$$

$$\left[\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)(a_i - \mu)^T\right].v = \lambda.v$$

$$v = \sum_{i=1}^{n}\alpha_i\,(\alpha_i)$$

$$K^2\alpha_j = n\lambda_j K\alpha_j$$

$$\tilde{K} = k - 2_{\frac{1}{n}}K + 1_{\frac{1}{n}}K1_{\frac{1}{n}}$$

Low dimensional data points:

$b_j = \sum_{i=1}^{n}\alpha_{ij}\,k(a, a_i)$where j=1,2,3,.....d.

end

Fig. 4. Output Interaction class label.

IV. EXPERIMENTAL RESULTS

This section discusses the results obtained at various stages, different models constructed and performance comparison of the proposed system with the existing system. Multi-layer 3D CNN is implemented using keras. The model is trained for 200 epochs and the weights that give the best performance on the validation data is loaded. This constructed model is then tested on test data. The whole data is randomly split into training and test data, test data is chosen to be one third of the total data. Multi-layer 3D CNN is constructed by performing convolution and pooling alternatively. Convolution is done by configuring the filters, kernel size, strides, padding, and activation. The activation functions used to built the model is relu. The convolution layer is followed by a statistics pooling layer. In addition a transformation is incorporated to produce a better result and is depicted in the comparison Table 1.

This combination of alternating convolution and pooling layer is followed by a global pooling layer, which converts its input to a 1d-vector. This 1d vector is given as input for kernel-based rv-tSNE for better visualization which is an extended algorithm of t-SNE that alleviates the problem of crowding and high variance. The transformed points of the proposed transformation algorithm are found more discriminative and the 2D embedding is given in Fig. 5 and corresponding graph is plotted with the comparison of the existing algorithms in Fig. 6. To evaluate the performance of the proposed DHIR system, benchmarking SBU Kinect Interaction dataset is used for training and a new dataset named AU Interaction dataset with 215 videos was recorded and used for testing. This proves the robustness of the system as it can perform classification independent of the actors and the environment. Various parameters like precision, recall, training time, recognition time, adaptation time are used to evaluate the interaction recognition techniques.

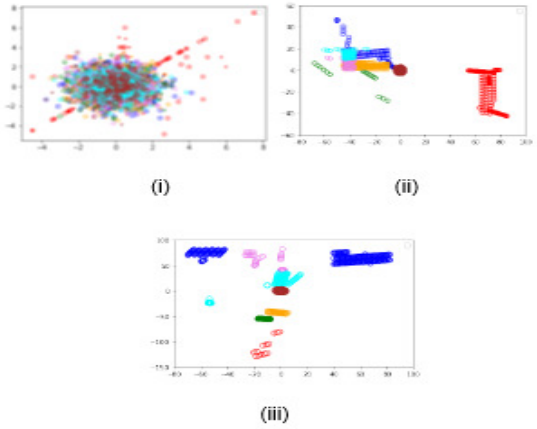

Fig. 5. 2D embedding results of (i) SNE,(ii) rv-tSNE,(iii) kernel-based rv-tSNE.

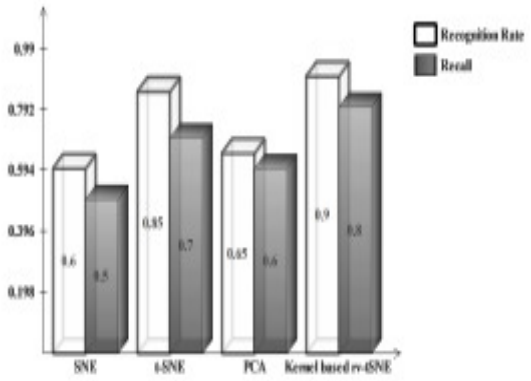

Fig. 6. Comparison chart for various transformation algorithms.

The confusion matrix obtained for proposed multi-layer 3D CNN is given in Table 2.

The graph depicting the performance metrics for each interaction is shown using the Fig. 7.

The results have also been compared against other existing algorithms, such as 2D CNN [14] and 3D CNN [20] approaches which is carried out and approximated with our selected datasets and average performance of the proposed system is compared with the existing algorithms and plotted in bar chart and shown in Fig. 8. The graph shows that the proposed DHIR system recognizes the interaction more accurately than the existing baseline systems.

This proves that the proposed algorithm is more effective when compared to other existing transformation algorithm. The fully connected network will have 6 neurons for each class label and for the final interaction layer, softmax activation function is utilized that maps the input layer to a particular class by

calculating the probability of input data. The input data with the maximum probability is assigned to a particular class label.

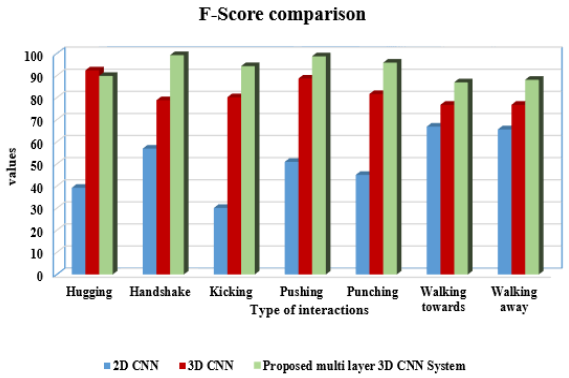

Fig. 7. Interaction comparison graph with existing System.

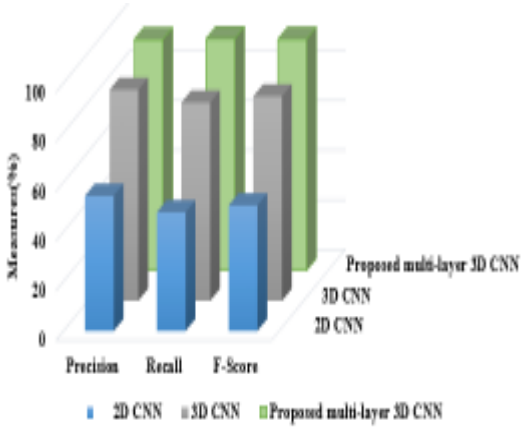

Fig. 8. Average comparison graph of proposed multi-layer 3D CNN with existing system.

Table 1: Performance metrics for models constructed.

Model	SBU Dataset			AU Dataset		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Base Model	58.21	57.22	57.71	62.14	60.22	61.16
Proposed model	93.88	94.78	94.32	93.56	93.23	93.394

Table 2: Confusion matrix.

Interactions	Hugging	Handshake	Kicking	Punching	Pushing	Walking towards	Walking away
Hugging	0.14	0.83	0.03	0	0	0	0
Handshake	0	1	0	0	0	0	0
Kicking	0	0	0.8	0.2	0	0	0
Punching	0	0	0	0.9	0.1	0	0
Pushing	0	0	0	0.1	0.9	0	0
Walking towards	0	0	0.1	0	0	0.85	0.05
Walking away	0	0	0.11	0	0	0.05	0.84

V. CONCLUSION

In this paper, a new multi-layer 3D CNN is implemented to recognize the interactions between human efficiently. Preprocessing techniques such as grayscale conversion and normalization are carried out to remove the inconsistencies in the datasets. The proposed kernel-based rv-tSNE transforms the high-dimensional feature to low dimension, therefore enabling effective embedding of all the information obtained from each frame of the video. By reducing the dimensions, the processing of data is reduced and hence, the classifier can easily group the interaction in different classes, enabling faster processing and classification. Multiple multi-layer 3D CNN models are constructed with different structures, hyper parameters and the best among those are chosen as the final model, the results of each model are then compared. The proposed system is tested using the benchmark SBU Kinect dataset and our own AU Interaction dataset, and proves that the proposed system is better than the existing system. In future, we are planning to extend the system for real-time human activity recognition by incorporating human behavior understanding in our system.

REFERENCES

[1]. Subetha, T., & Chitrakala, S. (2016). A survey on human activity recognition from videos. In 2016 International Conference on Information Communication and Embedded Systems (ICICES) (pp. 1-7). IEEE.

[2]. Xiao, X., Xu, D., & Wan, W. (2016, July). Overview: Video recognition from handcrafted method to deep learning method. In 2016 International Conference on Audio, Language and Image Processing (ICALIP) (pp. 646-651). IEEE.

[3]. Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. Image and vision computing, *60*, 4-21.

[4]. Zhang, Z., Ma, X., Song, R., Rong, X., Tian, X., Tian, G., & Li, Y. (2017, October). Deep learning based human action recognition: A survey. *In 2017 Chinese Automation Congress (CAC)* (pp. 3780-3785). IEEE.

[5]. San, P. P., Kakar, P., Li, X. L., Krishnaswamy, S., Yang, J. B., & Nguyen, M. N. (2017). Deep learning for human activity recognition. *In Big Data Analytics for Sensor-Network Collected Intelligence* (pp. 186-204). Academic Press.

[6]. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
[7]. Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110-1118).
[8]. Wang, H., & Wang, L. (2017). Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 499-508).
[9]. Prabhu. Understanding of Convolutional NeuralNetwork (CNN)-deep learning.
[10]. Gowda, S. N. (2017). Human activity recognition using combinatorial Deep Belief Networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1-6).
[11]. Hasan, M., & Roy-Chowdhury, A. K. (2015). A continuous learning framework for activity recognition using deep hybrid feature models. *IEEE Transactions on Multimedia, 17*(11), 1909-1922.
[12]. Koohzadi, M., & Charkari, N. M. (2017). Survey on deep learning methods in human action recognition. *IET Computer Vision, 11*(8), 623-632.
[13] Nielsen, M. A. (2015). Neural networks and deep learning. San Francisco, CA: Determination press.
[14]. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in Neural Information Processing Systems* (pp. 1097-1105).
[15]. Ravi, D., Wong, C., Lo, B., & Yang, G. Z. (2016). Deep learning for human activity recognition: A resource efficient implementation on low-power devices. *IEEE 13th international conference on wearable and implantable body sensor networks (BSN)* (pp. 71-76). IEEE.
[16]. Bagheri, M. A., Hu, G., Gao, Q., & Escalera, S. (2014). A framework of multi-classifier fusion for human action recognition. *22nd International Conference on Pattern Recognition* (pp. 1260-1265). IEEE.
[17]. Jalal, A., Uddin, M. Z., & Kim, T. S. (2012). Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE Transactions on Consumer Electronics, 58*(3), 863-871.
[18]. Lin, W., Chen, Y., Wu, J., Wang, H., Sheng, B., & Li, H. (2013). A new network-based algorithm for human activity recognition in videos. *IEEE transactions on circuits and systems for video technology, 24*(5), 826-841.
[19]. Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1933-1941).
[20]. Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence, 35*(1), 221-231.
[21]. Mo, L., Li, F., Zhu, Y., & Huang, A. (2016). Human physical activity recognition based on computer vision with deep learning model. *IEEE International* Instrumentation and Measurement Technology Conference Proceedings (pp. 1-6). IEEE.
[22]. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In International workshop on human behavior understanding (pp. 29-39). Springer, Berlin, Heidelberg.
[23]. Chen, Y., & Xue, Y. (2015). A deep learning approach to human activity recognition based on single accelerometer. IEEE international conference on systems, man, and cybernetics (pp. 1488-1492). IEEE.
[24]. Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., & Van Gool, L. (2017). Temporal 3d convnets: New architecture and transfer learning for video classification. arXiv preprint arXiv:1711.08200.
[25]. Zhang, B., Wang, L., Wang, Z., Qiao, Y., & Wang, H. (2018). Real-time action recognition with deeply transferred motion vector cnns. *IEEE Transactions on Image Processing, 27*(5), 2326-2339.
[26] Alp Guler, R., Neverova, N., & Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7297-7306).
[27]. Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen (2017). Beyond frame-level cnn: Saliency-aware 3-d CNN with LSTM for video action recognition. *IEEE Signal Processing Letters, 24*(4), 510–514.
[28]. Chen, H., Chen, J., Hu, R., Chen, C., & Wang, Z. (2017). Action recognition with temporal scale-invariant deep learning framework. *China Communications, 14*(2), 163-172.
[29]. Shi, Y., Tian, Y., Wang, Y., & Huang, T. (2017). Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia, 19*(7), 1510-1520.
[30]. Liu, J., Li, Y., Song, S., Xing, J., Lan, C., & Zeng, W. (2018). Multi-modality multi-task recurrent neuralnetwork for online action detection. *IEEE Transactions on Circuits and Systems for Video Technology, 29*(9), 2667-2682.
[31]. Das, S., Thonnat, M., Sakhalkar, K., Koperski, M., Bremond, F., & Francesca, G. (2019). A new hybrid architecture for human activity recognition from rgb-d videos. *In International Conference on Multimedia Modeling* (pp. 493-505). Springer, Cham.
[32]. Saravanan, C. (2010). Color image to grayscale image conversion. Second International Conference on Computer Engineering and Applications, *2*, 196-199). IEEE.
[33]. Sane, P., & Agrawal, R. (2017). Pixel normalization from numeric data as input to neural networks: For machine learning and image processing. International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2221-2225). IEEE.

[34]. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15*(1), 1929-1958.
[35]. Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. *In Advances in neural information processing systems* (pp. 857-864).
[36]. Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*, 2579-2605.
[37]. Subetha, T., & Chitrakala, S. (2017). Silhouette Based Human Action Recognition Using an Efficient Transformation Technique. In International Conference on Data Science Analytics and Applications (pp. 153-162). Springer, Singapore.
[38]. Singh, T., & Vishwakarma, D. K. (2020). A deeply coupled ConvNet for human activity recognition using dynamic and RGB images. Neural Computing & Applications.
[39]. Gopalakrishnan, N., Krishnan, V., & Gopalakrishnan, V. (2020). Ensemble Feature Selection to Improve Classification Accuracy in Human Activity Recognition. *In Inventive Communication and Computational Technologies* (pp. 541-548). Springer, Singapore.
[40]. Franco, A., Magnani, A., & Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognition Letters, 131*, 293-299.

How to cite this article: Subetha T., Chitrakala S. and Uday T. M. (2020). Dyadic Human Interaction Recognition from Videos using Multi-layer 3D CNN. *International Journal on Emerging Technologies*, *11*(3): 1033–1040.