International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 14 (2022) pp. 110-118

A Deep Learning Based Hybrid Approach for Human Physical Activity Recognition in Thermal Imaging

P.Srihari[1], Dr.J.Harikiran[2]

[1,2]School of CSE, Vellore Institute of Technology, VIT-AP University,
Amaravathi, A.P., India.
psrihari851@gmail.com

Abstract: **One of the tough topics of research is the recognition of human action in the supervision video presently. To classify using traditional algorithms of image processing the human actions comprised of the same patterns sequence that is hard. Video analysis is a significant field of research that implies analytics to the camera. It screens the contents of the video and abstracts intelligent data from it. The majority of the tasks in this field are subjected to constructing the classifying techniques on complex properties that are handcrafted or modelling DL-based CNNs, which work on inputs that are raw and take out important data along with the video directly. In this paper, for the segmentation of human activities in video sequences, k-means clustering is used. To classify and detect various human activities like boxing, carrying, digging, robbing, etc the hybrid combination of ResNet50 and 3D-CNN is utilized. The (Resnet-50) Pre-trained technique is utilized as a DL technique in this article. In order to capture the information of motion among the adjoining frames, the 3D-CNN extracts the features in the dimension of temporal together with the dimension of spatial. The performance measures are evaluated for various metrics such as precision, recall, f-score and accuracy.**

Keywords: Deep learning, Human action recognition, thermal imaging, ResNet50, 3D-CNN

I. Introduction

In several applications namely animation, automated observation, gaming robotics, smart home systems and interactions of human-machine, the action recognition of humans is an energetic field of research in the computer vision area. For easy as well as safe living, nuclear families and the aging population increase has led to the technology growth according to the systems that are supportive [1-3]. In literature, for recognition of action in videos namely, optical flow, body skeletons varying modalities and RGB-images have been studied. Using the temporal and spatial, the network of two-stream in human actions are established [4, 5]. Because human skeletons are constant to appearance and illumination, this technique has obtained early victory in the action recognition of humans. This is overwhelmed by 2018Vision' supported action recognition. Nevertheless, cameras need appropriate illumination to perform efficiently and also interrupt the person's routine being noticed [6-8].

A major research field in many applications of computer vision is the prediction and detection of activity of the human. It has vast application for the individual's security and safety as it supports finding abnormal and normal human behavior [9, 10]. The videos of human action comprise many frames consequently, 3D-Convolutional Neural Network is extra powerful to design both the temporal and spatial information in minimal duration. Although, 3D Convolutional Neural Network indicates enormous success for classification of action of human since the video sequences however insufficient to a constant input frame [11-13]. In real-time, throughout the whole video duration, the action of humans may be carried out. A novel combination of techniques to identify the action of a man taking into consideration of deep features in our proposed model is introduced. By this combination, with low computational cost, both low and high-level Spatio-temporal information can be maintained from the complete video sequences [14, 15].

The pre-trained DL is presented in the paper, namely ResNet. Deep Convolutional Neural Network models can perform straight on the raw inputs. The Residual Network (ResNet) was imagined to investigate the depth of NN (Neural Network). Due to the difficulty of the network on learning identity functions, it goals to manage the disappearing gradient problem that deteriorates as stated in the number of layers increases [16]. The number 50 represents the depth of the network, i.e., (the number of layers). Shortly, the ResNet targets to manage the problem of gradient descent. Such that, our aim is to model an effective framework of CNN particularized for detection of thermal images and predicting their night-time actions that obtain the best performance nevertheless of changes of season or low-resolution.

Majority of the architectures of traditional deep neural networks use 3D-CNN or 2D-CNN as a basis. Utilization of 2D CNN on a multi-stream and on a single steam model employing spatial and temporal extraction of features and its combination is implemented to obtain the required output [17,18]. For the classification of large-scale video, Google

researchers have employed 3D-CNN as a model of multi-stream. It uses multi-resolution, architecture which can train a bigger dataset in minimum time, in which 3D-CNN single steam models combine the frame's useful information having data of RGB or other forms of frames with the optical flow area to obtain the required output results. There are several methods according to the model of CNN like, frequency-based, motion-based, optical flow-based, color-based created by the researcher [19,20]. To overcome this a hybrid combination of ResNet50 and 3D-CNN is used for the detection and human detection classification in various activities like drinking, digging, boxing, robbing etc. The hybrid model consists of five sections are as follows: Image acquisition, Pre-processing, segmentation, detecting humans and classification of activities during the night in thermal images. In preprocessing the histogram equalization and the image enhancement is done. The segmentation is carried out using the k-means clustering algorithm technique.
Our major contribution towards this work is:

- We propose the K-means clustering for human segmentation in video sequences.
- Proposed a ResNet50-3D CNN-based method for recognizing various from videos.
- To detect and classify human activities like boxing, striking, chasing, robbing, etc. by the proposed hybrid deep learning techniques.
- Calculating the prominent performance metrics (i.e., precision, recall, accuracy as well as f-score) for the proposed hybrid techniques.

The remainder of the paper is ordered as follows. Section 2 presents the related works in the. In section 3, the problem statement is mentioned. The proposed methodology is shown in section 4. In section 5, the result section is shown. And finally, in Section 6, the conclusion is presented.

II. Literature Review

The method for this research is using image processing. The paper is mainly about thermal imaging of human action recognition at night time. Few challenges confronted and ongoing research to overcome the same and future advancements scope.

Zhou et al. [21] demonstrated that IR images can display the object's state during the night-time, that is to say, an essential path to retrieve the information of the objects during the night. In order to solve the extraction of pedestrians entirely from unique IR images, they analyzed the features of infrared images and presented an algorithm of pedestrian extraction according to a single IR image (infrared image). At first, the models of multi-projection and Neighborhood are designed to situate assumed areas of pedestrians. Finally, the head and global template of the weighted fusion is employed to extract walkers.

Park et al. [22] suggested an efficient and accurate technique for the detection of people in IR CCTV images at night. Therefore, 3 various IR images datasets were designed; two were obtained from infrared CCTV images initiated on a public beach and one more acquired from a FLIR (forward-looking infrared) camera. Besides, for the detection of a fine-grained person, a CNN-supported pixel-wise classifier was applied. The performance of detection of the presented technique was equated to five conventional methods of detection. The results display that the presented Convolutional Neural Network-based human detection technique exceeds conventional techniques detection in every dataset.

Creating three new suggested methods for DIRNV imaging and pedestrian detection track was suggested by Ashiba et al. [23]. The first technique was based on tracking using the GECUGC and gait detection of pedestrians. The GECUGC relies on the improvement technique using ECUG besides preprocessing succeeded by (GE) using CH as well as LAF. The second technique is on the basis of tracking by GHNTC and gait pedestrian detection. Succeeded by the GE employing CH and LAF, the GECUGC relies on a technique of improvement using HENT with pre-processing as well. The third algorithm is according to the analysis of space of scale with a lot of feature points of SURF as the basic criteria for gait detection of pedestrian, classification and tracking.

Liu, et al. [24] presented a method of detection of pedestrian and the robust thermal infrared vehicles in complicated scenarios. An essential parameter of weight β was presented first to improve the module of FSAF by rebuilding the FSAF module loss function in the online process of feature selection, which not only maximizes the infrared detection of objects mAP scores but also solves the issues of low detection precision of movement-blurred things. Then, in both inference and training, the improvised anchor-free FSAF component branches are familiar with the YOLOv3 detector of single-shot and function together with the anchor-supportive branches of the detector of YOLOv3. This technique linked both anchor-free branches and anchor-based branches efficiently and enhances the detection precision of small and thick objects.

Yu, et al. [25] investigated recognition of human action in images still and employed deep ensemble learning (DEL) to disintegrate the pose of the body automatically and recognize its information of the background. At first, they build an end-to-end technique of NCNN-based by connecting the non-sequential CNN module to the pretrained technique top. Nonsequential NCNN network topology can individually study the spatial and features of channel-wise along with branches that are parallel, which helps to enhance the performance of the model. Consequently, to abuse the non-sequential topology profit further, they presented an (end-to-end DEL) according to the weight optimization (DELWO) model. Finally, to obtain the finest prediction they modelled the model of deep ensemble learning according to the strategy

of voting (DELVS) to combine collectively several deep models with loaded coefficients.

Kiran et al. [26] suggested a recent approach for the (DL) utilization for HAR. Using a method of global contrast the video frames were pre-processed initially and then employed to train a model of DL employing transfer learning (TL) domain in the presented method. The model of Resnet-50 pre-trained is utilized as a technique of DL. From two layers the features were extracted: They are Fully Connected (FC) and Global Average Pool (GAP). By the Canonical Correlation Analysis, both the feature layers were combined. By using the function of Shanon Entropy-based threshold, the features were chosen. For final classification, the chosen features were passed to several classifiers finally. The experiments were carried out on five datasets that are available publicly such as, YouTube, UT-Interaction, KTH, UCF Sports as well as IXMAS.

Tu et al. [27] presented a structure that combined techniques of many hard actions as well as units of motion at various levels of hierarchy. They obtained this by presenting a productive model of the topic known as ML-HDP (Multi-label Hierarchical Dirichlet Process). The (ML-HDP) technique designs the motion units and actions and allows a highly accurate recognition co-occurrence relationship. Specifically, the topic model carries the representation of three-level in understanding of action. To perform three recognition tasks, they presented the straightforward methods including: joint classification, classification of action and constant actions segmentation and spatiotemporal action localization. In the paper, they investigated three various features usage and proven the effectiveness of their presented technique on four datasets that are publicly: Hollywood2, UCF101, MSR-II and KTH.

Tas et al. [28] presented a recent representation that encodes sequences of (3D) body skeleton joints in texture-like illustrations originated from rigorous kernel mathematical methods. Similar portrayal becomes the standard Convolutional Neural Network networks primary layer (e.g., ResNet-50), which is then utilized in the pipeline of adaptation of supervised domain to shift information from the source to the target dataset. Particularly, in this paper they utilized the overlying classes among datasets. They demonstrated the results of state-of-the-art on three available benchmarks available in public.

Tuncer et al. [29] demonstrated a method of novel ResNet-based recognition of signal. In this paper, ResNet101, ResNet50 and ResNet18 were used as feature extractor and so every network extracts thousands of features. Then the features extracted were combined and three thousands of features were achieved. At the phase of feature selection, 1000 best distinguishing features are chosen using Relief and such for the polynomial of third degree activation-depending SVM, the chosen features were employed as input. For activity and gender recognitions, the presented method acquired 99.61% and 99.96% accuracy for classification. For sensors signals, these results indicates that the presented method of pre-trained ensemble ResNet-based technique obtained high rate of success.

Khare et al. [30] presented automatic classification and extraction of features by the utilization of several (CNNs). Firstly, using a representation of time–frequency, the presented method transfers the filtered EEG signals into a samples of image. To convert time-domain EEG signals into images, smoothed distribution of pseudo Wigner–Ville is employed. Such images are nourished into pretrained techniques of ResNet-50, VGG16 and AlexNet together with CNN that is configurable. By measuring the precision, accuracy, F1-score, Mathew's correlation coefficient and false-positive rate, the four CNNs performance is calculated. The results achieved by calculating four CNNs indicates that adaptive CNN needs parameters of minimal learning with accuracy that is superior.

III. Problem Statement

In smart videos observation, activity detection is an important issue. It is a basic issue in computer vision, i.e. to detect human activity in supervision videos. The common work of activity of human recognition is extremely wide. This is the reason that the area has been theoretically separated and relying on the complexion of activity, we usually discuss actions, activities, primitive groups as well as interactions. Actions are made of primitives of the action and generally involve in the body movement parts. Examples comprise walking, jumping, running etc. The activities may contain some of the actions executed in a specific order, although group activities and the movement contain more than one person. The interaction and activities very often necessitate a critical visual script comprising extra objects. A different technique to the problem is required by every category. For example, for the reason of acknowledging dealings namely "fighting of two people", the process of high level has to obtain information about what the incident on the lower level namely action "punching" or even lower namely "extending an arm".

IV. Methodology

In the current work, proposed an efficient system for human detection and its attributes recognition like boxing, carrying, digging, robbing etc by deep learning model. The hybrid model consists of five sections are as follows: Image acquisition, Pre-processing, segmentation, detecting humans and classification of activities during the night in thermal images. Firstly, the image is acquired from the videos. Next is image processing and this is to remove the excess noise from the images. After this step, we have to prepare our region of interest (ROI). For this K-means clustering is used for images segmentation. Furthermore, to detect the humans and classify their activities from the sequential videos proposed is a hybrid

ResNet50 and 3D-CNN model. The proposed methodology architecture is shown in figure 1.

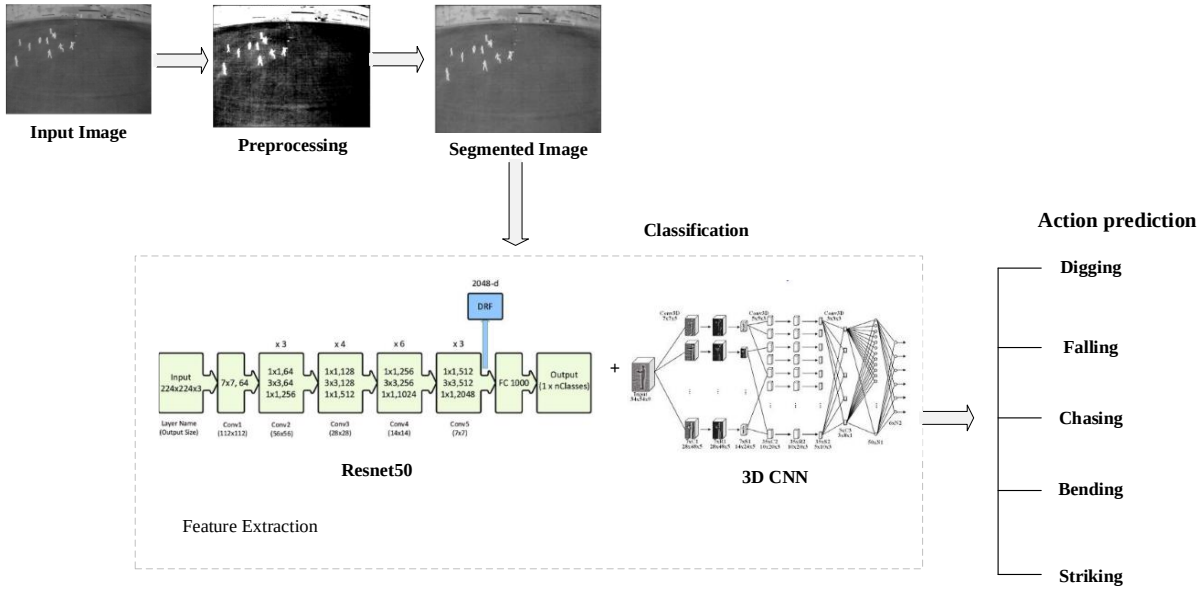


Figure 1. Overall architecture for proposed approach

A. Image Acquisition

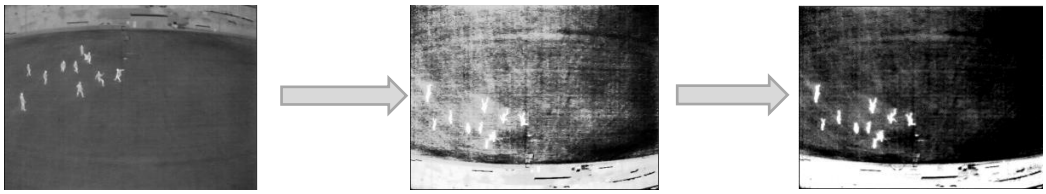
In this paper, since the research is almost about human action, we choose the actions such as digging, boxing, carrying, robbing, and drinking. Therefore, we can extract typical features from these 5 human actions. Due to the amplitude signals and waveform information ruling, several lists of muscles are often very different, and an incorporated mode of sampling was approved in this experiment. To complete five different movements independently, for rejecting the unique differences of influence, two healthy subjects were employed. Simultaneously, in order to reduce the fatigue effect of signal stability of the researcher after finishing many movements, a rest of one minute was taken once gathering two sets of data.

B. Preprocessing

Infrared night time image preprocessing operation mainly comprises of two steps. To improve the quality of the image for few badly determined images we employ image enhancement and histogram equalization. Furthermore, based on the maximum intercept size of the box to guarantee that humans at the edge can be detected well as well as abbreviated, while zero padding is used at the edges of the human image. The consequence is to face the requirement of human detection work in the future.

The human action patterns consisting of normal behavior are to be studied so that they can predict actions once unknown data is fed into them. The Reference Model focuses on the use of the AAU-PD-T dataset of 5 actions (drinking, boxing, robbing, digging, and carrying) for detection of activity. It uses for 4 scenes such as outdoors, outdoors with scale variation, indoors and outdoors with different clothes. Once this preprocessing is done the region of interest should be done. The preprocessing steps involve:

- Splitting the video
- Frame extraction from videos
- Resizing and transforming
- Image normalization



(a) **(b)** **(c)**

Figure 2: Image Preprocessing (A) Original Image (B) After Histogram Equalization (C) After Image Enhancement

C. Resnet-50

DL displayed great victory in the past few periods for many applications of video supervision as well as medical imaging. Video surveillance is prominent in the field of research, yet the researchers face many problems called large datasets as well as imbalanced datasets. We employed a pre-trained model of DL named ResNet-50 in this research article. The general techniques of ResNet generally carry out the bouncing of double or triple layers, in which it consists of batch normalization and (ReLu). From the former layer to the upcoming layer, the motivation for the sake of skipping across layers is to connect the output. This helps decrease the issue of vanishing gradient. Skipping supports network simplification because it utilizes a small number of layers in the starting stages of training. It accelerates the learning process as a result. Since the network will study the features space, it gradually recovers such skipped layers. A (NN) neural network in the portion of residual absence examines the extra space of feature that makes it insecure. ResNet-50 is a (Convolutional Neural Network), which is trained on the AAU-PD-T dataset exactly, comprising of nearly a hundred million images of thousand classes. The ResNet-50 network has 50 layers of intensity. The real layers number and chosen layers are described in Table1:

Image input layer (1)
Addition layer (16)
Convolution 2-D layer (53)
Average pooling 2-D layer (1)
Max pooling 2-D layer (1)
Fully connected layer (1)
Batch normalization layer (53)
Fully connected layer (1)
ReLU layer (49)
output layer Classification (1)

Table 1. ResNet50 CNN technique layers

According to the number of classes, this ResNet50 architecture is modified. This is the reason, we deleted the final layer and also included a new layer that comprises a lot of performance action classes. Then the presented technique is trained on 70% data, where 30% are used for testing. In the process of training, we determine 64 mini-batch sizes, 100 epochs and 0.0001 learning rate. The size of the input is similar to the real deep model. From the final two layers of feature, we extract the features namely Fully Connected and Average Pool layer. For further processing in the last stage, we combined both vectors into one vector feature. In figure 3 the architecture of the ResNet-50 is represented:

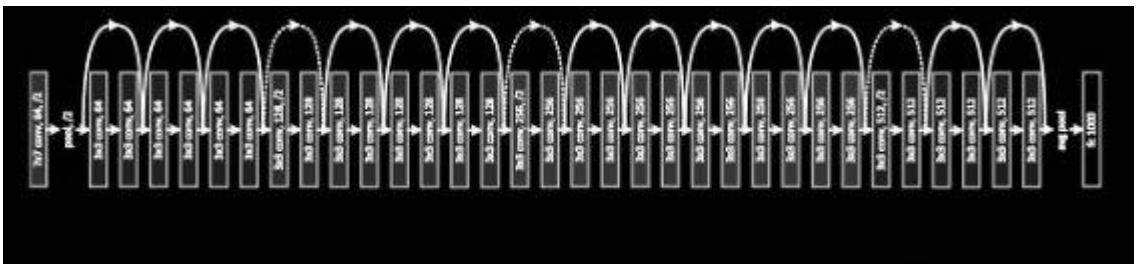


Figure 3. Architecture diagram of ResNet-50

D. Segmentation Using K-Means Clustering

The thermal images are generally with low contrast and low resolution. The edge detection algorithm of K-means is the major traditional partition established method of clustering and it is either the ten data mining classical algorithms. By clustering the K points in the space, the primary notion of the K-means clustering algorithm is to group the materials nearest to them. Repeatedly, the clusters centroid values are updated individually till the best clustering results are achieved. Based on the function of the prototype, the algorithm of K-means clustering is a characteristic representation of the clustering method. As the optimization objective function, it grabs the distance to the prototype from the data point. By the method of finding functions' extreme values, the rules of adjustment of repetitive action are achieved. As the similarity measure, the algorithm of K-means takes the Euclidean distance, which is to discover the ideal classification of the first vector of cluster center, such that the index of evaluation is low. The sum of the function of the error square measure is utilized as a function of the clustering measure. Although the K-means algorithm is effective, the K value should be provided in advance as well as the K value selection is very hard to calculate.

The algorithm efficiency is higher, but according to the requirement to decide the clusters K number, it fetches specific hardships for self-regulating computations. To define the values of K, the method of segmentation conveniently, this technique hybrid the connected domain algorithm to the maximum. After comprehensive experiments, the K value is generally between 2 and 10. We utilize the linked algorithm of the domain to the maximum to rebuild the image including only the object of the target, register the number and relate it with the K value to get an exact value of K. Segmentation using K-means clustering can be depicted in figure 2.

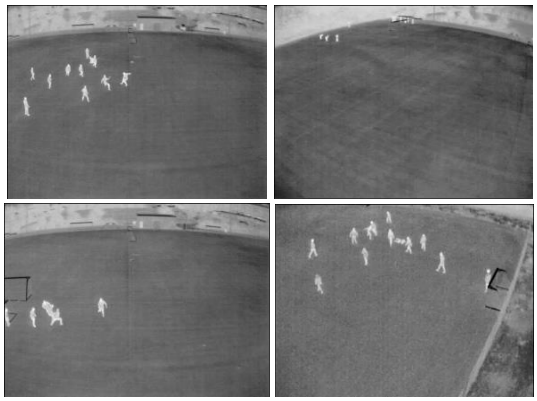


Figure 4. Segmented image for various human action recognition

E. Feature Selection

This is the procedure of choosing the finest features for correct classification through a minimal time of execution is known as Feature selection. In this article, according to the Shannon Entropy, a new function of threshold is proposed to choose the features that which is the best. At first, we compute the entropy value of the combined vector. The equation is denoted as follows.

$$H = -\sum_{a=1}^{N} Pa \log(P_a)$$

In which, 'N' denotes the features total number, 'Pa' is the every features probability, and 'a' denotes each feature index in the combined vector. Next, to select the best features, we executed a threshold function. We compute 20 steps and after every looping, through the fitness function, the chosen features are evaluated. In the end, for the final classification, the feature set that is of higher accuracy is chosen. And for final classification, the chosen features are transferred in the classifiers of supervised learning.

F. Action Recognition and Classification

All these models are calculated with AAU-PD-T, which consists of 5 actions functioned by diverse subjects. Following multiple layers of subsampling and convolution, we obtain a flattened fully connected feature vector finally. The final layer consists of 6 units. A similar setting is utilized for a united Convolutional Neural Network with additional features ResNet50 along with CNN. For training, 85% of random subjects are chosen and 15% are used for testing. Similar training, as well as testing split, are utilized and found that our hybridized CNN gets very good accuracy.

The network classifies a person who requires a more complex structure. It is previously admitted that the system of human vision is a process of multiple scaling. Therefore, for robust human action classification, it will be desirable to combine various levels of movement features. The union of changeless and more features that are delicate creates better motion features representation. We have extracted a dual feature vector (By utilizing ResNet50 and CNN) which explains the temporal and spatial information of the input video. For convolution operation, the above architecture of CNN, we have employed only 7 input frames, which is not sufficient to encrypt information of the higher level of temporal motion from the entire video of the action. Over, if we maximize the frame of the input, managing all the network's trainable parameters will be very difficult. To tackle these issues, we have used (CNN) for capturing the information of motion of every frame present in an action video. There is a global average pooling prior to the layer that is connected fully. After linking of two feature vectors, for final classification one additional fully connected layer followed by a dropout of 0.4 and one denser layer with the function of sigmoid activation is included.

V. Simulation Results

The simulation is done in Python. F-score, precision, recall, accuracy have been used to evaluate the classifier's efficiency. The comparison was made with the existing approaches.

Techniques	F-Score	Precision	Recall	Accuracy
Fourier descriptor-based	92.5	92.04	91.24	87.46
GEI-based	68.47	75.60	63.28	95.2
Fuzzy system-based	90.85	89.52	93.03	98.8
CycleGAN + CNN + LSTM-CNN	84.49	84.02	85.98	99.6
Proposed (ResNet50+ 3D-CNN)	94.28	94.06	95.06	99.9

Table 2. The evaluation of various metrics by the comparison of existing with proposed techniques

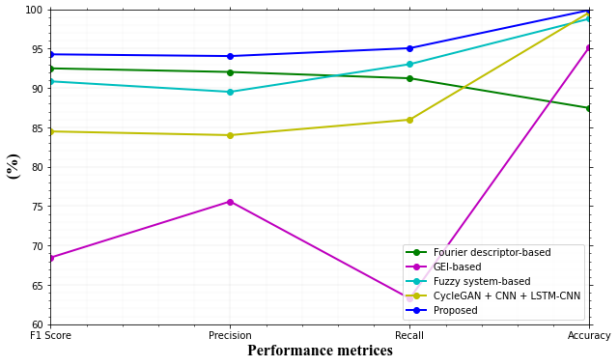


Figure 5. Comparison of overall performance metrics

The evaluation of various metrics by the comparison of existing with proposed techniques can be depicted in table 2. Fourier descriptor-based can achieve 92.5% of F-score, 92.04% of precision, recall 91.24%, accuracy 87.46%. GEI-based technique yield 68.47% off-score, 75.60% of precision, 63.28% of recall and 95.2% of accuracy. The fuzzy system-based approach gains a 90.85% of F-score, 89.52% of precision, recall 93.03% and accuracy of 98.8%. Then CycleGAN + CNN + LSTM-CNN technique yield 84.49% of F-score, 84.02% of precision, recall 85.98% and accuracy 99.6%. And our proposed approach gains 94.28% of F-score, 94.06% of precision, recall 95.06% and accuracy of 99.9%. Compared with the existing approach our proposed approach yields greater performance. Figure 5 represents the overall comparison.

TECHNIQUES	TRAINING (ms)	TESTING (ms)
Faster RCNN	176.11	105.09
YOLOv3	180.52	70.81
Retina net	166.31	103.50
proposed	140.27	64.12

Table 3. Differentiation of proposed training and testing time with existing approach

Training and testing time can be compared with the existing techniques depicted in table 3. The faster RCNN approach yields 176.11 ms run time and 105.09 ms testing time. YOLOv3 approach gains 180.52 ms run time and 70.81 ms testing time. Retina net gain 166.31 ms run time and 103.50 ms testing time. And our proposed method gain 140.27 ms run time and 64.12 ms testing time. When comparing the training and testing time of the proposed technique with the existing approach. The proposed technique yields an efficient training and testing time. Fig 6 shows the computational timings.

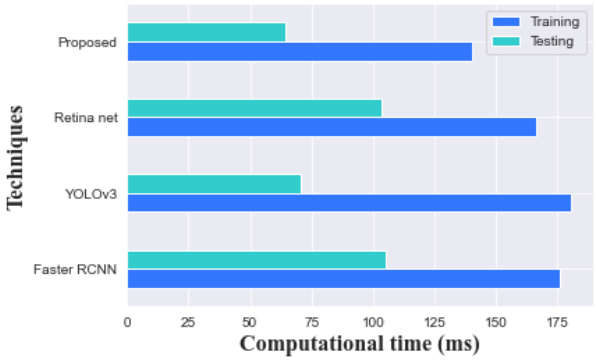


Figure 6. Computational time comparison

HUMAN ACTIONS	ACCURACY (%)
Falling	96.9
Chasing	97.8
Bending	99.5
Ball striking	98.2

Table 4. Comparison of various human actions accuracy

Comparing various human actions accuracy can be depicted in table 4. The various human actions like falling, chasing, bending, ball striking, in which falling can yield 96.9% accuracy, chasing obtain 97.8% of accuracy, bending gain 99.5% of accuracy and ball striking yield 98.2%. Among the four human actions accuracy, the bending action achieves greater accuracy. Figure 5 shows that the graphical representation of human action recognition.

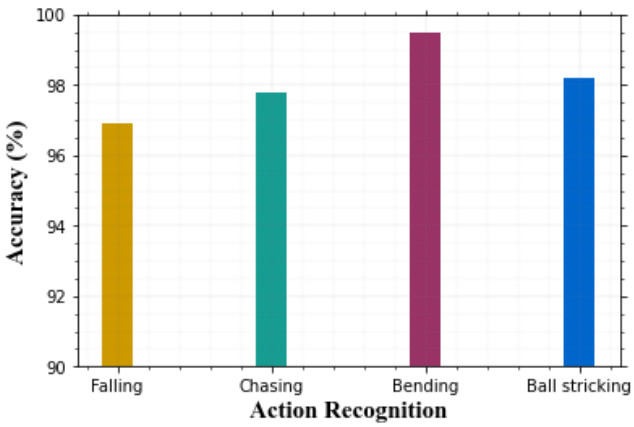


Figure 7. Comparison of training and validation loss among different human action analysis

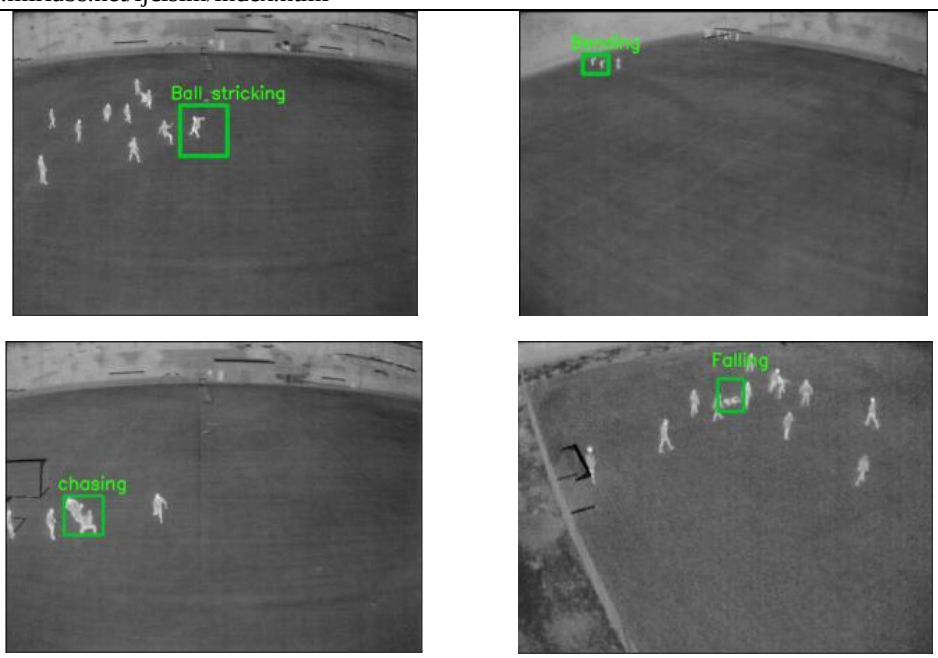


Figure 8. Final detected pose for various actions in human

VI. Conclusion

In this paper, we propose the K-means clustering for human segmentation in video sequences. A combination of ResNet50 and 3D CNN-based method for recognizing from videos. To detect and classify human activities like boxing, carrying, digging, robbing etc by the proposed hybrid deep learning techniques. Calculating the well-known performance measures (i.e., accuracy, precision, recall and f-score) on the proposed hybrid models. Comparing the performance of the proposed models with the existing techniques. We proposed a combined technique of ResNet50 and Convolutional Neural Network model for recognition of human action at night time. This system extract features from both the dimension of spatial and temporal activities like ordinary convolution of 3D. Along with convolution and subsampling, this model also takes extra handcrafted features from the whole input video motion clip. At last, the CNNs last fully connected layer is linked with the additional feature vectors that are handcrafted of ResNet50. We evaluated the ensemble CNN and ResNet50 on the AAU-PD-T dataset and found that the concatenation of handcrafted features with 3D Convolutional Neural Network outperforms the dataset with a rigid number of training as well as testing divisions.

Acknowledgements

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

References

[1] Ashiba, H. I. (2021). Dark infrared night vision imaging proposed work for pedestrian detection and tracking. *Multimedia Tools and Applications*, 1-27.

[2] Baek, J., Hong, S., Kim, J., & Kim, E. (2017). Efficient pedestrian detection at nighttime using a thermal camera. Sensors, 17(8), 1850.

[3] Chen, Y. Y., Li, G. Y., Jhong, S. Y., Chen, P. H., Tsai, C. C., & Chen, P. H. (2020). Nighttime Pedestrian Detection Based on Thermal Imaging and Convolutional Neural Networks. Sensors and Materials, 32(10), 3157-3167.

[4] Chen, Y., & Shin, H. (2020). Pedestrian detection at night in infrared images using an attention-guided encoder-decoder convolutional neural network. *Applied Sciences*, 10(3), 809.

[5] d'Acremont, A., Fablet, R., Baussard, A., & Quin, G. (2019). CNN-based target recognition and identification for infrared imaging in defense systems. Sensors, 19(9), 2040.

[6] Dai, X., Duan, Y., Hu, J., Liu, S., Hu, C., He, Y., & Meng, J. (2019). Near infrared nighttime road pedestrians recognition based on convolutional neural network. *Infrared Physics & Technology*, 97, 25-32.

[7] Dai, X., Hu, J., Zhang, H., Shitu, A., Luo, C., Osman, A., & Duan, Y. (2021). Multi-task faster R-CNN for nighttime pedestrian detection and distance estimation. *Infrared Physics & Technology*, 115, 103694.

[8] Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. IATSS research, 43(4), 244-252.

[9] Hou, Y. L., Song, Y., Hao, X., Shen, Y., Qian, M., & Chen, H. (2018). Multispectral pedestrian detection based on deep convolutional neural networks. Infrared Physics & Technology, 94, 69-77.

[10] Ivasic-Kos, M., Kristo, M., & Pobar, M. (2019, September). Person Detection in thermal videos using YOLO. In Proceedings of SAI Intelligent Systems Conference (pp. 254-267). Springer, Cham.

[11] Khalifa, A. B., Alouani, I., Mahjoub, M. A., & Amara, N. E. B. (2020). Pedestrian detection using a moving camera: A novel framework for foreground detection. Cognitive Systems Research, 60, 77-96.

[12] Khare, S. K., & Bajaj, V. (2020). Time-frequency representation and convolutional neural network-based emotion recognition. IEEE transactions on neural networks and learning systems.

[13] Kiran, S., Khan, M. A., Javed, M. Y., Alhaisoni, M., Tariq, U., Nam, Y., & Sharif, M. (2021). Multi-Layered Deep Learning Features Fusion for Human Action Recognition. CMC-COMPUTERS MATERIALS & CONTINUA, 69(3), 4061-4075.

[14] Krišto, M., Ivasic-Kos, M., & Pobar, M. (2020). Thermal Object Detection in Difficult Weather Conditions Using YOLO. IEEE Access, 8, 125459-125476.

[15] Lahmyed, R., El Ansari, M., & Ellahyani, A. (2019). A new thermal infrared and visible spectrum images-based pedestrian detection system. Multimedia Tools and Applications, 78(12), 15861-15885.

[16] Lahmyed, R., El Ansari, M., & Ellahyani, A. (2019). A new thermal infrared and visible spectrum images-based pedestrian detection system. Multimedia Tools and Applications, 78(12), 15861-15885.

[17] Liu, Y., Su, H., Zeng, C., & Li, X. (2021). A robust thermal infrared vehicle and pedestrian detection method in complex scenes. Sensors, 21(4), 1240.

[18] Ma, M. (2020). Infrared pedestrian detection algorithm based on multimedia image recombination and matrix restoration. Multimedia Tools and Applications, 79(13), 9267-9282.

[19] Park, J., Chen, J., Cho, Y. K., Kang, D. Y., & Son, B. J. (2020). CNN-based person detection using infrared images for night-time intrusion warning systems. Sensors, 20(1), 34.

[20] Song, X., Gao, S., & Chen, C. (2021). A multispectral feature fusion network for robust pedestrian detection. Alexandria Engineering Journal, 60(1), 73-85.

[21] Sun, Y., Shao, Y., Yang, G., & Xie, H. (2021). A Method of Infrared Image Pedestrian Detection with Improved YOLOv3 Algorithm. American Journal of Optics and Photonics, 9(3), 32-38.

[22] TAS, Y., & Koniusz, P. (2018). Cnn-based action recognition and supervised domain adaptation on 3d body skeletons via kernel feature maps. arXiv preprint arXiv:1806.09078.

[23] Tu, N. A., Huynh-The, T., Khan, K. U., & Lee, Y. K. (2018). ML-HDP: A hierarchical bayesian nonparametric model for recognizing human actions in video. IEEE Transactions on Circuits and Systems for Video Technology, 29(3), 800-814.

[24] Tumas, P., Nowosielski, A., & Serackis, A. (2020). Pedestrian detection in severe weather conditions. IEEE Access, 8, 62775-62784.

[25] Tuncer, T., Ertam, F., Dogan, S., Aydemir, E., & Pławiak, P. (2020). Ensemble residual network-based gender and activity recognition method with signals. The Journal of Supercomputing, 76(3), 2119-2138.

[26] Yu, X., Zhang, Z., Wu, L., Pang, W., Chen, H., Yu, Z., & Li, B. (2020). Deep ensemble learning for human action recognition in still images. Complexity, 2020.

[27] Zhang, Y., Yin, Z., Nie, L., & Huang, S. (2020). Attention Based Multi-Layer Fusion of Multispectral Images for Pedestrian Detection. IEEE Access, 8, 165071-165084.

[28] Zhou, D., Qiu, S., Song, Y., & Xia, K. (2020). A pedestrian extraction algorithm based on single infrared image. Infrared Physics & Technology, 105, 103236.

[29] Zhou, D., Qiu, S., Song, Y., & Xia, K. (2020). A pedestrian extraction algorithm based on single infrared image. *Infrared Physics & Technology*, *105*, 103236.

[30] Zhou, D., Qiu, S., Song, Y., & Xia, K. (2020). A pedestrian extraction algorithm based on single infrared image. Infrared Physics & Technology, 105, 103236.

Authors' biography

P.Srihari, Research scholar in school of Computer science and engineering vellore institute of technology VIT-AP. His research interest includes Image Processing,Thermal Imaging,Deep Learning. he can be reached at srihari.19PHD7018@vitap.ac.in, https://orcid.org/0000-0003-2542-4961.

Dr.J.Harikiran, Associate professor in school of Computerscience and engineering vellore institute of technology VIT-AP. His research interest includes Image Processing,Video Processing Microarray Images, Hyperspectral Imaging, Machine learning,Deep Learning. He can be reached at harikiran.j@vitap.ac.in