Library Progress International
Vol.44 No. 6, Jul-Dec 2024: P. 1237-1247

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Multi-modal fusion enhances activity recognition by integrating RGBD and skeletal data

Archana Vinod Bansod[1], Shailesh Kumar[2]

[1]Research Scholar, JJTU
[2]Research Guide, JJTU

How to cite this article: Archana Vinod Bansod, Shailesh Kumar (2024) Multi-modal fusion enhances activity recognition by integrating RGBD and skeletal data. *Library Progress International,* 44(6), 1237-1247

Abstract: In this paper, the proposed work is based on a multi-modal fusion framework for human activity recognition (HAR). This approach makes use of three modalities such as RGB, depth maps and 3D-Skeletion joint position to develop robust HAR system. Two 3DCNN models with different network parameters and an LSTM model are used to obtain the features from each modality. Next, the score of each activity is obtained using SVM in each model and optimized using two evolutionally algorithms. The experimental work on the public dataset has also been discussed to validate the proposed approach. The experimental results show that the proposed framework is an improvement over previous work and is capable of accurately recognizing human activities

Keywords: Human Activity Recognition (HAR), Deep Learning, LSTM, 3DCNN

I. INTRODUCTION

The RGB video sequences used in the development of the human activity detection system are insufficient to accurately identify every human activity. This is because the human body's articulated structure is only captured in two dimensions by the RGB video frames; it is not feasible to record every movement of the body simultaneously in three dimensions, which leads to the loss of three dimensions and the inability to achieve human locations and scale variations in three dimensions. On the other hand, depth maps give an indication of 3D joint positions and are less impacted by changes in lighting. The sensitivity of the depth map is lower than that of RGB videos. Moreover, 3D skeleton data uses 3D coordinate coordinates to represent the head, neck, abdomen, and other body parts of a human in 3D space. When compared to other lighting conditions and perspectives, this representation is more potent. When all three modalities are used combined, the robust activity recognition may be more accurate than when utilising any one of them alone. As a result, employing many modalities to precisely identify the activities becomes simple.

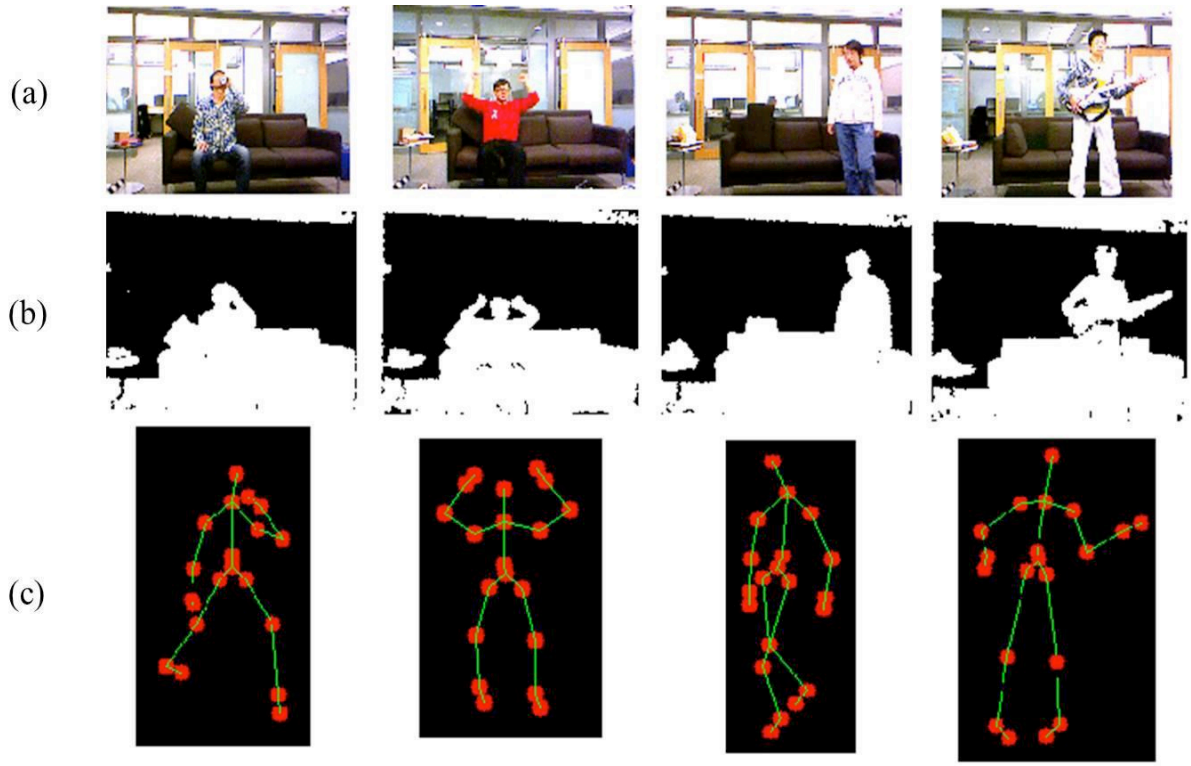

Figure 1.1: Activity recognition in RGB, Depth, and Skeleton information (a) Activities ("drink," "cheer-up," "standup," "playing-guitar") only in RGB frames that lose some of the body movement in 3D space (b) Activities ("drink," "cheer-up," "standup," "playing-guitar") only in depth data that loses appearance and spatial information (c) Activities ("drink," "cheer-up," "standup," "playing-guitar") only in skeleton that loses appearance information but retains 3D movement information and is insensitive to changes in illumination and perspective. [1]

Figure 1.1 (a) [1] depicts human activity in RGB frames with some body movement information lost in 3D space. Comparably, Figure 1.1 (c) [1] depicts activity in 3D skeleton joint positions in 3D space and is insensitive to changes in illumination and view variations, while Figure 1.1 (b) [1] only depicts activity in-depth data and does not contain appearance information.

Even though numerous studies have combined multi-modality (RGB, depth, and skeletal data) to perform activity recognition [2–11], there is still room for improvement in terms of recognition accuracy. Therefore, in order to create a reliable activity detection system, the research project that is suggested in this chapter makes use of all three modalities, including RGB, depth, and 3D-skeleton data, employing evolutionary algorithms.

II. LITERATURE SURVEY

To accomplish activity recognition tasks with robustness and efficiency, multimodality activity recognition integrates many modality features, including RBG, Depth, and Skeleton data. Ijjina et al. [3], for instance, presented a motion sequence-based deep learning method for RGBD data-based activity recognition. With this method, every activity sequence has a crucial pose. The convolutional neural network receives the pattern derived from the RGB and depth video as input to learn discriminative features. Using four datasets, including the SBU Kinect interaction, NATOPS gesture, MIVIA action, and Weizmann datasets, the effectiveness of the proposed strategy is shown. A

multimodal method for activity recognition that makes use of both local and global motion cues was presented by Gu et al. [4]. Depth based 3-Channel motion history pictures (MHIs) are utilised for global feature learning. In a similar vein, skeleton graphs are used to facilitate learning of local spatial and temporal aspects. The scores from each stream are then combined. The effectiveness of the suggested strategy is demonstrated by the experimental findings, which are based on two RGB-D datasets. Zho et al. [5] developed a three-stream 3D space-time convolutional neural network (3DSTCNN) architecture for activity recognition based on depth and skeletal data. For global space-time feature learning, the original depth map, depth motion maps derived from depth data, and 3D skeleton sequence are input into separate streams. Using skeleton and depth data, all streams are optimised to learn space and time features efficiently. Three activity recognition datasets, including UTD-MHAD, MSRAction3D, and UTKinectAction3D, are used to assess the suggested methodology. The outcomes of the experiment demonstrate how successful the suggested system is. In order to recognise human activity, Singh et al. [6] developed a multi-modal system that makes use of RGB, Depth, and 3D joint coordinate data. Using all of the information that is simultaneously available, they proposed the deep bottleneck multimodal feature fusion (D-BMFF) framework. Prior to being coupled with RGB and depth frames, 3D joint coordinates are transformed into RGB skeleton motion history images (RGB-SklMHI), from which features are retrieved for deep network training. The characteristics obtained from the bottleneck layer immediately before the top layer are fused using a multi-set discriminant correlation analysis (M-DCA). The characteristics are then divided into several activity classes using a multiclass SVM. Four activity recognition datasets, including UTKinectAction3D, SBU Interaction, CAD-60, and Florence 3D dataset, are used to assess the proposed methodology. Additionally, Weiyao et al. [7] introduced a multi-modal system for activity recognition based on the Bilinear Pooling and Attention Network (BPAN). The characteristics that are extracted from RGB and skeleton data are compressed using the BPAN model after a data preparation operation for the RGB and skeleton data. Lastly, a fully connected perceptron network is employed for activity classification. The acquired experimental findings show that the proposed strategy is better than the state-of-the-art methods discussed before.

A technique for large-scale video categorization based on RGBD data was presented by Li et al. [8]. This method involves first extracting a sequence of 32 frames from the RGB and depth modalites, which are then fed into the C3D model for the purpose of learning space-time characteristics. Subsequently, the features that were retrieved are merged to minimise superfluous synthetic data and enhance overall performance. Using an SVM classifier, the proposed method validates the Chalearn LAP ISoGD dataset with 49.2% accuracy. The method achieves the highest recognition accuracy of 56.9% with test data, outperforming the baseline and other approaches. Additionally, Zhu et al. [9] proposed a multimodal gesture detection system based on ConvLSTM memory networks and 3D convolution. The approach that is being given uses 3DCNN to learn short-term spatial-temporal characteristics, and ConvLSTM network, which is based on extracted short-term features, is used to learn long-term spatial-time features. The model is fine-tuned to avoid the overfitting. The technique achieves 98.89% on SKIG and 51.02% recognition accuracy on the IsoGD validation set. It is validated using the ChaLearn LAB ISoGD dataset. An unsupervised learning method that quickly picks up motion information from the video was presented by Luo et al. [10]. The proposed framework effectively acquires long-term 3D motion from two movie clip pictures. To get around the framework's

intricacy, the motion is depicted as a three-dimensional flow. Moreover, these 3D flows are predicted using an encode-decoder based on RNNs. The NTU RGB+D and MSRDailyActivity3D multimodality datasets are used to test the efficiency of the suggested method. The suggested framework works with any kind of modality, including RGB, depth, and RGBD. For RGB and skeleton-based activity recognition, Mahasseni et al. [11] introduced a regularisation of LSTM. The method that is being given also makes use of an additional encode LSTM (eLSTM) that is based on the human skeleton. The 3D body joints should aid in the learning of pertinent motion patterns because the skeleton coordinates positions are unaffected by view variation and backdrop clutter.

III. PROPOSED METHODOLOGY

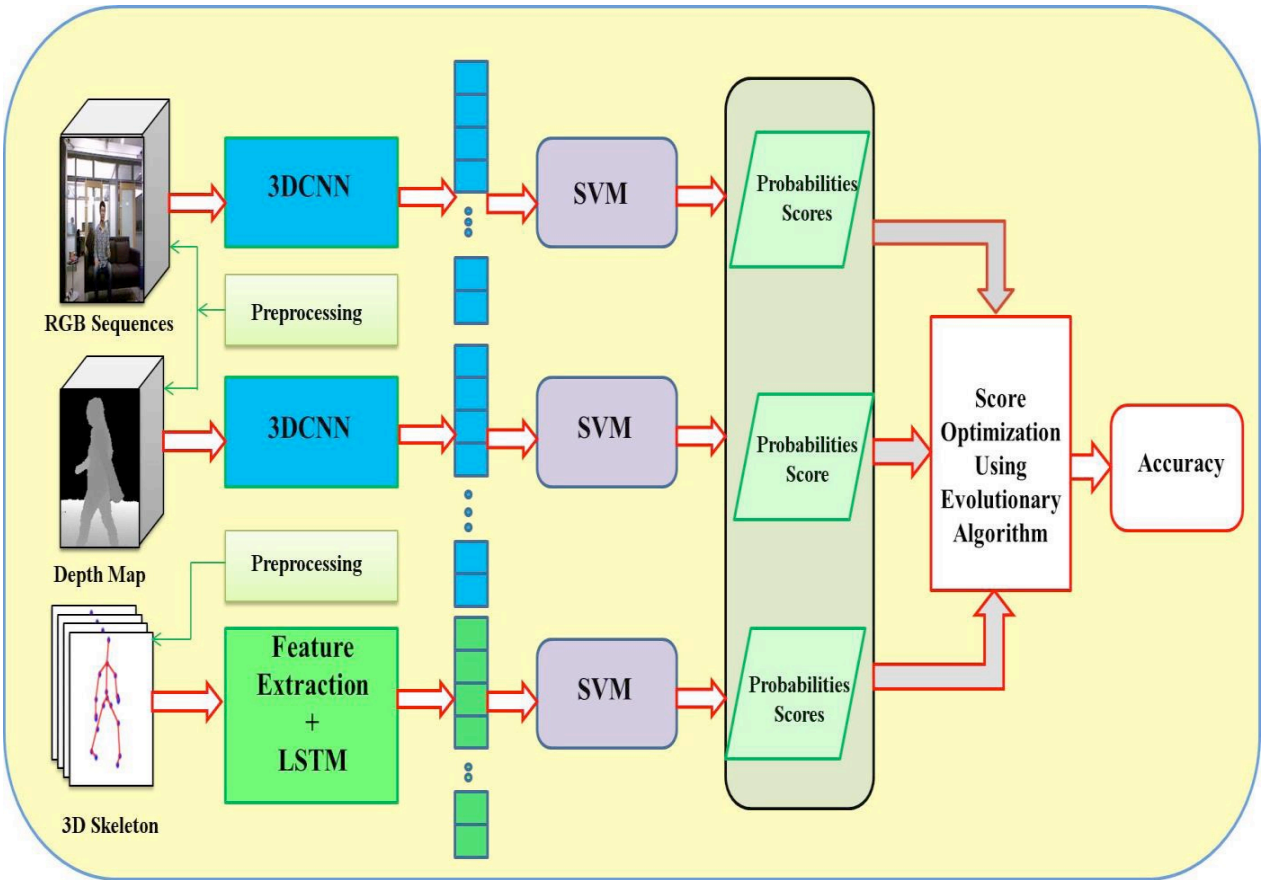

Figure 1.2 Flow diagram of the proposed approach

The suggested method's flow diagram is shown in Figure 1.2. The steps that follow outline the total contribution of our suggested work.

- The first phase involves learning features from skeleton data using an LSTM network and extracting space-time information from RGB and depth video sequences using two 3D-Convolutional Neural Networks.
- To provide the class scores for each test activity in distinct streams, an SVM network is then trained using features that were taken from each of the three models.

- Ultimately, two evolutionary algorithms are employed to fuse and optimise the class score of the test activities.
- The suggested methodology is validated using two publicly accessible human activity datasets, UTKinectAction3D and MSRDailyActivity3D, yielding 96.05% and 85.74% recognition accuracy compared to the aforementioned cutting-edge techniques.
- The outcomes of the experiment show how successful the suggested strategy is.

3.1 Learning spatial-temporal characteristics from RGB data with 3DCNN:

A spatial-temporal 3D-convolutional neural network that may simultaneously acquire information along the time and space dimensions is the 3DCNN, as described in [12]. By convolving the 3D filter across the 3D volume input data, the convolutional 3D features may be derived. A series of frames piled one on top of the other makes up the 3D volume data. The depth dimension and temporal information are both learned, and several feature maps are derived from several neighbouring frames in the preceding layer. Multiple convolutional layers are employed to extract characteristics in both lower and higher dimensions. Consequently, a large number of convolutional layers are used in order to enhance the number of feature mappings in the network. The 3D filter is thus convolved on the 3D input cube to produce a convolutional 3D.

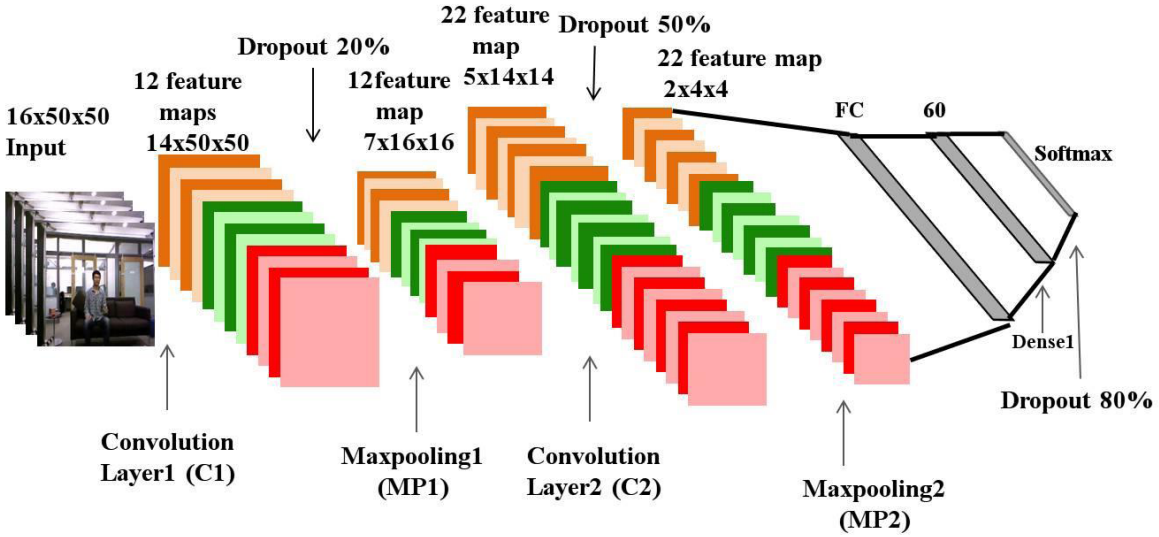

Figure 1.3 illustrates the 3DCNN feature learning architecture using RGB data.

RGB video sequences are utilised to teach a 3DCNN model spatial and temporal features. The three layers of the 3DCNN are two pooling and two convolutional. We stacked a fixed number of frames along with the depth dimension because it is a difficult effort to standardise all films in the dataset to have an equal number of frames because each video in the dataset has a distinct number of frames. In order to conduct the experiment, the depth dimension was captured in 16 frames. As a result, the 3DCNN model receives an input cube with the dimensions [16×50×50]. The input cube is first processed by the first convolution layer (C1), and the output of C1 is subsequently processed by the first max pooling (ML1) layer. In the meantime, there is a 20% dropout rate between ML1 and C1. With 50% dropout between C2 and ML2, the model employs a second set of convolutional layers (C2) and a maxpooling layer (ML2). The model employs a fully connected layer (FC) to discover the feature

vector after the second max pooling layer. Following the completely connected layer in the model is a single dense layer with sixty neurons. Our proposed model ends with an output layer that uses a softmax layer. The final dropout layer, which has an 80% dropout amount, is utilised right before the softmax layer. The model employs 12 and 64 filters in the C1 and C2 layers, respectively, with sizes of (3×1×1) and (3×3×3). In the MP1 and MP2 levels, the network uses the same down sampling size (2×3×3) for each maxpooling layer. For the MSRDailyActivity3D and UTKinectAction3D datasets, RGB video sequences have been processed using the same network settings.

IV. RESULTS & DISCUSSION

In this work, we trained and tested our proposed technique on two datasets: MSRDailyActivity3D [1] and UTKinectAction3D [13]. The dataset is captured by a Kinect sensor. Every action within the dataset is coordinated across all platforms.

The Kinect depth sensor was used to create the MSRDailyActivity3D [1] dataset. It lists sixteen human actions, including calling a cell phone, using a laptop, walking, sitting, lying down on a sofa, eating, playing an instrument, standing up, sitting motionless, drinking, throwing paper, using a hoover, reading a book, playing a game, writing on paper and cheering yourself up. Ten subjects—five of whom were male and five of whom were female—generated the dataset. Every activity was recorded while the subjects were in a room with a sofa. This indicates that there is interaction between the item and the person while they are engaging in an activity during activity capture. Every participant completes each task twice, once while seated and once while standing. Each action is recorded in three distinct modalities: skeleton, depth map, and RGB. A single modality corresponds to 16×10×2 = 320 activity sequences, resulting in 320×3 = 960 total. Sample frames from the "Cheer up" activity of the MSRDailyActivity3D dataset are shown in Figure 1.4 in the RGB, depth, and skeleton modalities at various time intervals.

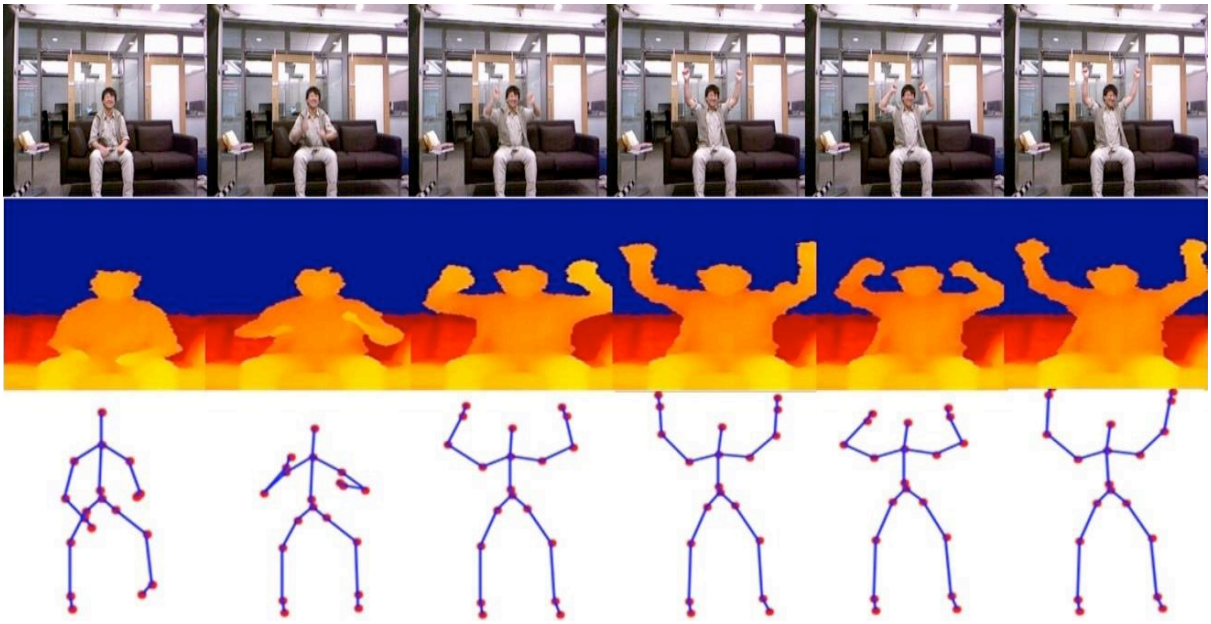

Figure 1.4: A selection of example frames from the MSRDailyActivity3D dataset's "Cheer up" activity in the RGB, depth, and skeleton modalities at various time intervals.

The suggested method is validated using the human activity dataset UTkinectAction3D [13]. There are ten indoor human actions in the dataset. All activity sequences were recorded at a frame rate of 30

frames per second using the Windows SDK kit and the Kinect depth camera. The Kinect camera is capable of capturing all three modalities, RGB Depth, and skeleton, at a range of 4 to 12 feet. Ten human indoor activities that were recorded by ten participants are included in the dataset; of these, nine are performed by men and one by women twice. Except for one person who is left handed, every subject is right handed. This video collection has 10×10×2 = 200 activity sequences altogether for each modality. Consequently, 200 × 3 = 600 activity sequences total.

Standing, carrying, walking, pushing, wave handing, pulling, throwing, sitting, picking up, and clap handing are the ten activities included in the dataset. The actions were recorded in an indoor setting. An additional text file included in the dataset has labels for every activity sequence. A selection of sample frames from the UTKinectAction3D dataset, comprising ten distinct activity sequences in each of the three modalities, are shown in Figure 1.5.

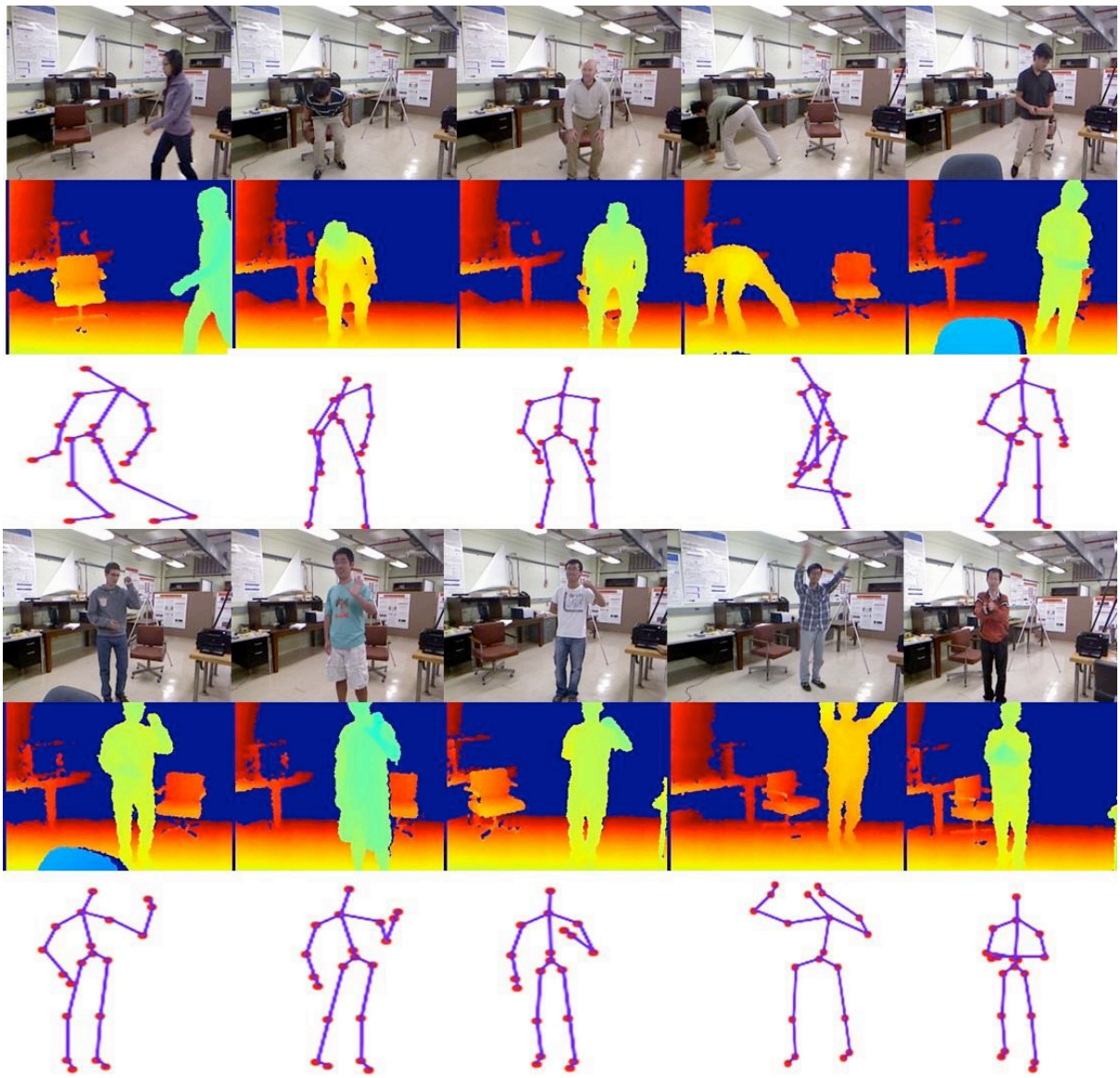

Figure 1.5: Some sample frames of UTKinectAction3D dataset with 10 different activities in all three modalities RGB, Depth and Skeleton sequences.

Experiments include the use of 3DCNN to extract spatiotemporal features from RGB video sequences, another 3DCNN to extract spatiotemporal features from depth maps, an LSTM network to extract

spatiotemporal features from skeleton sequences, and a trained SVM model for classification. The first 3DCNN network was tested with several input cube combinations, such as (13×50×50), (14×50×50), (15×50×50), and (16×50×50), and it produced the best spatial-temporal characteristics at the (16×50×50) input cube size. To extract spatio-temporal characteristics from the depth sequences and achieve the best set of features at an input size of (13 × 32 × 32), several combinations of input cubes, such as (13 × 32 × 32), (14 × 32 × 32), (15 × 32 × 32), and (16 × 32 × 32), are also utilised. The learning rate for both networks was set to 5×10^{-4} for training. During the network training, an Adam optimizer and a categorical-cross entropy loss function were employed.

To validate our suggested strategy, we used the Leave One User Out Cross Validation (LOUOCV) technique. Ten-fold cross validation has been employed in this work, where nine users are used to train the network and the remaining user is used to test the network after each fold. One × 16 × 2 = 32 activity sequences is utilised for testing in each fold during validation, while nine × 16 × 2 = 288 activity sequences are used for training. Each test activity's class score is recorded at the time the network is tested.

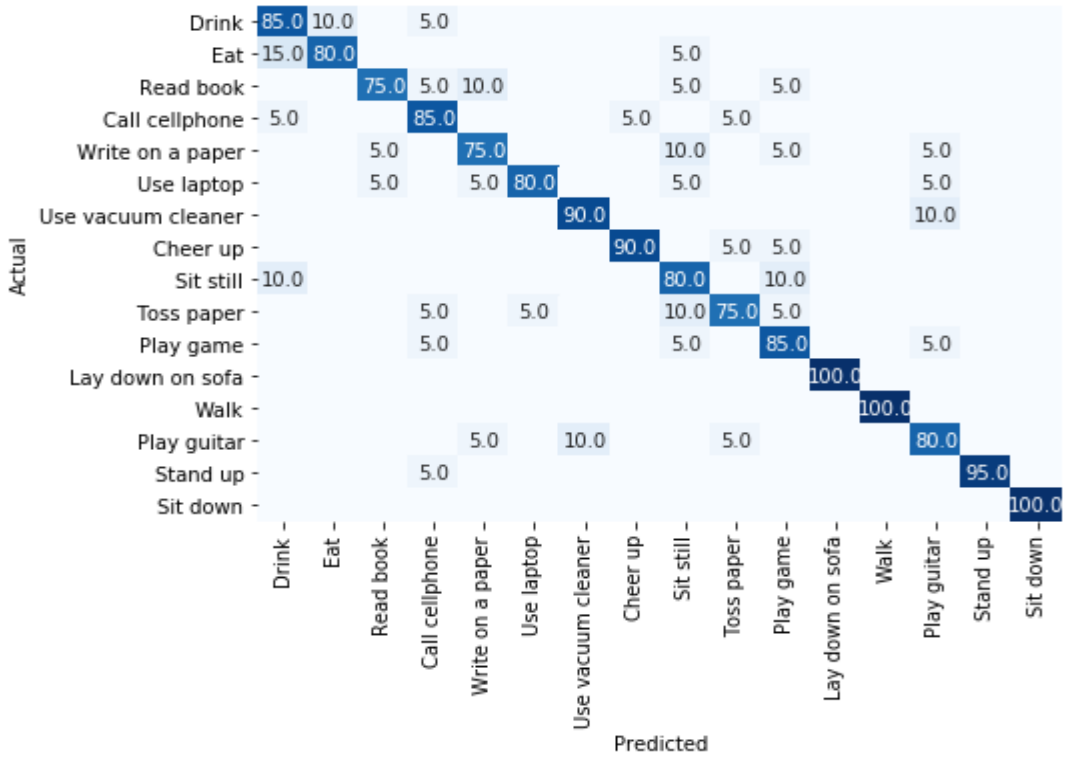

Figure 1.6: The suggested method's confusion matrix for the MSRDailyActivity3D Dataset

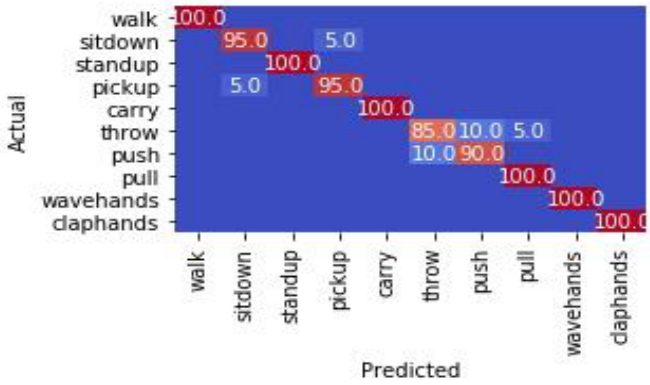

Figure 1.7: The proposed approach's confusion matrix on the UTKinectAction3D dataset

V. CONCLUSION

An approach to deep multi-modal fusion for human activity recognition is presented in this research. The suggested method operates in three phases. First, using two 3DCNN models and an LSTM model, respectively, the spatial and temporal characteristics are extracted from the RGB, depth, and skeleton sequences in all three modalities. Second, three SVM models with the same network parameters are trained in parallel for each modality using the deep features that were recovered in the first phase. These SVM models are then used to provide class scores for each test activity. The final step involves merging and optimising each test activities generated class scores using GA and PSO, two evolutionary algorithms. Based on two 3D Convolutional Neural Networks and an LSTM network, the experimental findings show that the proposed technique learns high level spatial and temporal characteristics from all three modalities relatively rapidly. The method works better than using just one modality at a time because it uses RGB, depth, and skeleton sequences, which are the three modalities. The class label judgements from all three parallel models are completely exploited by providing an optimization-based score fusion approach. For this purpose, a Genetic Algorithm (GA) is used.

References

[1] Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012, June). Mining actionlet ensemble for action recognition with depth cameras. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1290-1297). IEEE.

[2] Duan, J., Zhou, S., Wan, J., Guo, X., & Li, S. Z. (2016). Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition. arXiv preprint arXiv:1611.06689.

[3] Ijjina, E. P., & Chalavadi, K. M. (2017). Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognition, 72, 504-516.

[4] Gu, Y., Ye, X., Sheng, W., Ou, Y., & Li, Y. (2020). Multiple stream deep learning model for human action recognition. Image and Vision Computing, 93, 103818.

[5] Zhao, C., Chen, M., Zhao, J., Wang, Q., & Shen, Y. (2019). 3d behavior recognition based on multi-modal deep space-time learning. Applied Sciences, 9(4), 716.

[6] Singh, T., & Vishwakarma, D. K. (2021). A deep multimodal network based on bottleneck layer features fusion for action recognition. Multimedia Tools and Applications, 1-21.

[7] Weiyao, X., Muqing, W., Min, Z., & Ting, X. (2021). Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition. IEEE Sensors Journal.

[8] Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Li, R., & Song, J. (2016, December). Large-scale gesture

recognition with a fusion of rgb-d data based on the c3d model. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 25-30). IEEE.
[9] Zhu, G., Zhang, L., Shen, P., & Song, J. (2017). Multimodal gesture recognition using 3-D convolution and convolutional LSTM. Ieee Access, 5, 4517-4524.
[10] Luo, Z., Peng, B., Huang, D. A., Alahi, A., & Fei-Fei, L. (2017). Unsupervised learning of long-term motion dynamics for videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2203-2212).
[11] Mahasseni, B., & Todorovic, S. (2016). Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3054-3062).
[12] 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.
[13] Xia, L., Chen, C. C., &Aggarwal, J. K. (2012, June). View invariant human action recognition using histograms of 3d joints. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (pp. 20-27). IEEE. [14] Koohzadi, M., and N. Moghadam Charkari. 2017. Survey on Deep Learning Methods in Human Action Recognition. IET Computer Vision 11 (8):623–32. doi:10.1049/iet-cvi. 2016.0355.
[15] Nweke, H. F., Y. Wah Teh, M. Ali Al-garadi, and U. Rita Alo. 2018. Deep Learning Algorithms for Human Activity Recognition Using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges. Expert Systems with Applications 105:233–61. doi:10.1016/j. eswa.2018.03.056.
[16] Zhang, H.-B., Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, D. Ji-Xiang, and D.-S. Chen. 2019. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. Mpdi. doi:10.3390/s19051005.
[17] Singh, T., and D. Kumar Vishwakarma. 2019. Video Benchmarks of Human Action Datasets: A Review. Artificial Intelligence Review 52 (2):1107–54. doi: 10.1007/s10462-018-9651-1.
[18] Liu, B., H. Cai, J. Zhaojie, and H. Liu. 2019. RGB-D Sensing Based Human Action and Interaction Analysis: A Survey. Pattern Recognition 94:1–12. doi: 10.1016/j. patcog.2019.05.020.
[19] Zawar, H., Q. Z. Sheng, and W. Emma Zhang. 2020. A Review and Categorization of Techniques on Device-Free Human Activity Recognition. Journal of Network and Computer Applications 167:102738. December 2019. doi:10.1016/j.jnca.2020.102738.
[20] Minh Dang, L., K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon. 2020. Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey. Pattern Recognition 108:107561. doi:10.1016/j.patcog.2020.107561.
[21] Lei, W., D. Q. Huynh, and P. Koniusz. 2020. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. IEEE Transactions on Image Processing 29:15–28. doi:10.1109/TIP.2019.2925285.
[22] Jegham, I., A. Ben Khalifa, I. Alouani, and M. Ali Mahjoub. 2020. Vision-Based Human Action Recognition: An Overview and Real World Challenges. Forensic Science International: Digital Investigation 32:200901. doi:10.1016/j.fsidi.2019.200901.
[23] Majumder, S., and N. Kehtarnavaz. 2021. Vision and Inertial Sensing Fusion for Human Action Recognition: A Review. IEEE Sensors Journal 21 (3):2454–67. doi:10.1109/ JSEN.2020.3022326.

[24] Özyer, T., A. Duygu Selin, and R. Alhajj. 2021. Human Action Recognition Approaches with Video Datasets—A Survey. Knowledge-Based Systems 222:106995. doi:10.1016/j. knosys.2021.106995.
[25] Verma, K. K., B. Mohan Singh, and A. Dixit. 2022. A Review of Supervised and Unsupervised Machine Learning Techniques for Suspicious Behavior Recognition in Intelligent Surveillance System. International Journal of Information Technology (Singapore) 14 (1):397–410. doi:10.1007/s41870-019-00364-0.