REFERENCES

1. Agahian, S., Negin, F., & Köse, C. (2020). An efficient human action recognition framework with pose-based spatiotemporal features. Engineering Science and Technology, an International Journal, 23(1): 196-203.
2. Agarwal, P., & Alam, M. (2020). A lightweight deep learning model for human activity recognition on edge devices. Procedia Computer Science, vol. 167, pp. 2364-2373.
3. Ahmad, T., Wu, J., Alwageed, H. S., Khan, F., Khan, J., & Lee, Y. (2023). Human Activity Recognition Based on Deep-Temporal Learning Using Convolution Neural Networks Features and Bidirectional Gated Recurrent Unit With Features Selection. IEEE Access, vol. 11, pp. 33148-33159.
4. Alazrai, R., Hababeh, M., Baha'A, A., Ali, M. Z., & Daoud, M. I. (2020). An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals. IEEE Access, vol. 8, pp. 197695-197710.
5. Alsarhan, T., Alawneh, L., Al-Zinati, M., & Al-Ayyoub, M. (2019, October). Bidirectional gated recurrent units for human activity recognition using accelerometer data. IEEE SENSORS, pp. 1-4.
6. Arifoglu, D., & Bouchachia, A. (2017). Activity recognition and abnormal behaviour detection with recurrent neural networks. Procedia Computer Science, vol. 110, pp. 86-93.
7. Arize, A. C., Bakarezos, P., Kasibhatla, K. M., Malindretos, J., & Panayides, A. (2014). The gini coefficient. decomposition and overlapping. Journal of Advanced Studies in Finance, vol. 5, no. 1 (9), pp. 47.
8. Arshad, M. H., Bilal, M., & Gani, A. (2022). Human activity recognition: Review, taxonomy and open challenges. Sensors, vol. 22, no. 17, pp. 6463.
9. Athota, R. K., & Sumathi, D. (2022). Human activity recognition based on hybrid learning algorithm for wearable sensor data. Measurement: Sensors, vol. 24, pp. 100512.
10. Basavaiah, J., & Patil, C. M. (2020). Robust Feature Extraction and Classification Based Automated Human Action Recognition System for Multiple Datasets. International Journal of Intelligent Engineering & Systems, vol. 13, no.1.

11. Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: a survey. Multimedia Tools and Applications, vol. 79, pp. 30509-30555.
12. Brownlee, J. (2019). A gentle introduction to a standard human activity recognition problem. Machine Learning Mastery. https://machinelearning mastery. com/how-to-load-and-explore-a-standard-human-activity-recognition-problem/(accessed Mar. 1, 2021).
13. Bruce, X. B., Liu, Y., Chan, K. C., Yang, Q., & Wang, X. (2021). Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression. Pattern Recognition, vol. 119, pp. 108095.
14. Cao, C., Zhang, Y., & Lu, H. (2015, October). Spatio-temporal triangular-chain CRF for activity recognition. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1151-1154.
15. Cao, C., Zhang, Y., Zhang, C., & Lu, H. (2016). Action recognition with joints-pooled 3d deep convolutional descriptors. International Journal of Computational Intelligence and Applications, Vol. 1, pp. 3.
16. Cao, C., Zhang, Y., Zhang, C., & Lu, H. (2017). Body joint guided 3-D deep convolutional descriptors for action recognition. IEEE transactions on cybernetics, vol. 48, no. 3, pp. 1095-1108.
17. Casale, P., Pujol, O., & Radeva, P. (2011). Human activity recognition from accelerometer data using a wearable device. In Proceedings of 5th Iberian Conference on Pattern Recognition and Image Analysis, pp. 289-296.
18. Cha, J., Saqlain, M., Kim, D., Lee, S., Lee, S., & Baek, S. (2022). Learning 3D skeletal representation from transformer for action recognition. IEEE Access, vol. 10, pp. 67541-67550.
19. Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. ACM Computing Surveys (CSUR), vol. 54, no. 4, pp. 1-40.
20. Chen, L. H., Liu, K. C., Hsieh, C. Y., & Chan, C. T. (2017, May). Drinking gesture spotting and identification using single wrist-worn inertial sensor. IEEE, in 2017 International Conference on Applied System Innovation (ICASI), pp. 299-302.

21. Chen, Y., Ma, G., Yuan, C., Li, B., Zhang, H., Wang, F., & Hu, W. (2020). Graph convolutional network with structure pooling and joint-wise channel attention for action recognition. Pattern Recognition, vol. 103, pp. 107321.
22. Cho, H., & Yoon, S. M. (2018). Divide and conquer-based 1D CNN human activity recognition using test data sharpening. Sensors, vol. 18, no. 4, pp. 1055.
23. Chua, S. L., Marsland, S., & Guesgen, H. W. (2009). Behaviour recognition from sensory streams in smart environments. In Proceedings on AI 2009: Advances in Artificial Intelligence: 22nd Australasian Joint Conference, Melbourne, Australia, December 1-4, 2009. Vol. 22, pp. 666-675. Springer Berlin Heidelberg.
24. Cumin, J., Lefebvre, G., Ramparany, F., & Crowley, J. L. (2017). Human activity recognition using place-based decision fusion in smart homes. Springer International Publishing in Modeling and Using Context: Proceedings on10th International and Interdisciplinary Conference, CONTEXT 2017, Paris, France, June 20-23, 2017, 10 pp. 137-150.
25. Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, vol. 108, pp. 107561.
26. Dang, Y., Yin, J., & Zhang, S. (2022). Relation-based associative joint location for human pose estimation in videos. IEEE Transactions on Image Processing, Vol. 31, pp. 3973-3986.
27. Deshpnande, A., & Warhade, K. K. (2021, March). An improved model for human activity recognition by integrated feature approach and optimized SVM. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, pp. 571-576.
28. Dhiman, C., Saxena, M., & Vishwakarma, D. K. (2019, September). Skeleton-based view invariant deep features for human activity recognition. In 2019 IEEE fifth international conference on multimedia big data (BigMM), pp. 225-230.
29. Ding, J., & Wang, Y. (2019). WiFi CSI-based human activity recognition using deep recurrent neural network. IEEE Access, vol. 7, pp. 174257-174269.
30. Ding, J., Wang, Y., & Fu, X. (2020). Wihi: WiFi based human identity identification using deep learning. IEEE Access, vol. 8, pp. 129246-129262.

31. Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2019). Image caption generation with high-level image features. Pattern Recognition Letters, vol. 123, pp. 89-95.
32. Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, D., & Marshall, S. (2014). A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiological measurement, vol. 35, no. 11, pp. 2191.
33. Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. the Journal of machine Learning research, vol. 9, pp. 1871-1874.
34. Fatima, I., Fahim, M., Lee, Y. K., & Lee, S. (2013). A unified framework for activity recognition-based behavior analysis and action prediction in smart homes. Sensors, vol. 13, no. 2, pp. 2682-2699.
35. Feng, M., & Meunier, J. (2022). Skeleton graph-neural-network-based human action recognition: a survey. Sensors, vol. 22, no. 6, pp. 1-52.
36. Florence, C. S., Bergen, G., Atherly, A., Burns, E., Stevens, J., & Drake, C. (2018). Medical costs of fatal and nonfatal falls in older adults. Journal of the American Geriatrics Society, vol. 66, no. 4, pp. 693-698.
37. Gumaei, A., Hassan, M. M., Alelaiwi, A., & Alsalman, H. (2019). A hybrid deep learning model for human activity recognition using multimodal body sensing data. IEEE Access, vol. 7, pp. 99152-99160.
38. Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. International Journal of Information Management Data Insights, vol. 1, no. 2, pp. 100046.
39. Han, J., Qian, C., Wang, X., Ma, D., Zhao, J., Xi, W., ... & Wang, Z. (2015). Twins: Device-free object tracking using passive tags. IEEE/ACM Transactions on Networking, vol. 24, no. 3, pp. 1605-1617.
40. Hao, W., & Zhang, Z. (2019). Spatiotemporal distilled dense-connectivity network for video action recognition. Pattern Recognition, vol. 92, no. 13-24.
41. Hao, X., Li, J., Guo, Y., Jiang, T., & Yu, M. (2021). Hypergraph neural network for skeleton-based action recognition. IEEE Transactions on Image Processing, vol. 30, pp. 2263-2275.

42. Helal, S., & Bull, C. N. (2019). From smart homes to smart-ready homes and communities. Dementia and geriatric cognitive disorders, vol. 47, no. 3, pp. 157-163.
43. Hristov, P., Manolova, A., & Boumbarov, O. (2020, October). Deep learning and SVM-based method for human activity recognition with skeleton data. In 2020 28th National Conference with International Participation (TELECOM) proceeding on IEEE, pp. 49-52.
44. https://www.cdc.gov/homeandrecreationalsafety/falls/fallcost.html,2018,online; accessed 03 July 2018.
45. Hur, T., Bang, J., Huynh-The, T., Lee, J., Kim, J. I., & Lee, S. (2018). Iss2Image: A novel signal-encoding technique for CNN-based human activity recognition. Sensors, vol. 18, no. 11, pp. 3910.
46. Hussain, A., Hussain, T., Ullah, W., & Baik, S. W. (2022). Vision transformer and deep sequence learning for human activity recognition in surveillance videos. Computational Intelligence and Neuroscience.
47. Ihianle, I. K., Nwajana, A. O., Ebenuwa, S. H., Otuka, R. I., Owa, K., & Orisatoki, M. O. (2020). A deep learning approach for human activities recognition from multimodal sensing devices. IEEE Access, vol. 8, pp. 179028-179038.
48. Jalal, A., Kamal, S., & Kim, D. (2014). A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. Sensors, vol. 14, no. 7, pp. 11735-11759.
49. Jaouedi, N., Boujnah, N., & Bouhlel, M. S. (2020). A new hybrid deep learning model for human action recognition. Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 4, pp. 447-453.
50. Jayatilaka, A., & Ranasinghe, D. C. (2017). Real-time fluid intake gesture recognition based on batteryless UHF RFID technology. Pervasive and Mobile Computing, vol. 34, pp. 146-156.
51. 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, vol. 35, no. 1, pp. 221-231.

52. Jiang, M., Kong, J., Bebis, G., & Huo, H. (2015). Informative joints based human action recognition using skeleton contexts. Signal Processing: Image Communication, vol. 33, pp. 29-40.
53. Jiang, W., & Yin, Z. (2015, October). Human activity recognition using wearable sensors by deep convolutional neural networks. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1307-1310.
54. Kao, J. Y., Ortega, A., Tian, D., Mansour, H., & Vetro, A. (2019, September). Graph based skeleton modeling for human activity analysis. In 2019 IEEE International Conference on Image Processing (ICIP) IEEE, pp. 2025-2029.
55. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732.
56. Ke, S. R., Thuc, H. L. U., Lee, Y. J., Hwang, J. N., Yoo, J. H., & Choi, K. H. (2013). A review on video-based human activity recognition. Computers, vol. 2, no. 2, pp. 88-131.
57. Khan, M. Z., Taha, A., Farooq, M., Shawky, M. A., Imran, M., & Abbasi, Q. H. (2022, July). Comparative Analysis of Artificial Intelligence on Contactless Human Activity localization. In 2022 International Telecommunications Conference (ITC-Egypt) pp. 1-3.
58. Khan, S., Khan, M. A., Alhaisoni, M., Tariq, U., Yong, H. S., Armghan, A., & Alenezi, F. (2021). Human action recognition: a paradigm of best deep learning features selection and serial based extended fusion. Sensors, vol. 21, no. 23, pp. 7941.
59. Khatun, M. A., Yousuf, M. A., Ahmed, S., Uddin, M. Z., Alyami, S. A., Al-Ashhab, S., & Moni, M. A. (2022). Deep CNN-LSTM with self-attention model for human activity recognition using wearable sensor. IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-16.
60. Kim, H., Lee, S., & Jung, H. (2019). Human activity recognition by using convolutional neural network. International Journal of Electrical and Computer Engineering, vol. 9, no. 6, pp. 5270.
61. Kim, M., Jeong, C. Y., & Shin, H. C. (2018, October). Activity recognition using fully convolutional network from smartphone accelerometer. In 2018

International conference on information and communication technology convergence (ICTC), pp. 1482-1484.

62. Kulsoom, F., Narejo, S., Mehmood, Z., Chaudhry, H. N., Butt, A., & Bashir, A. K. (2022). A review of machine learning-based human activity recognition for diverse applications. Neural Computing and Applications, pp. 1-36.
63. Lan, T., Sigal, L., & Mori, G. (2012, June). Social roles in hierarchical models for human activity recognition. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354-1361.
64. Li, J., Xie, Z., Wang, Z., Lin, Z., Lu, C., Zhao, Z., ... & Ding, W. (2023). A triboelectric gait sensor system for human activity recognition and user identification. Nano Energy, vol. 112, pp. 108473.
65. Li, M., & Sun, Q. (2021). 3D skeletal human action recognition using a CNN fusion model. Mathematical Problems in Engineering, pp. 1-11.
66. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2021). Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3316-3333.
67. Li, X., Zhang, D., Lv, Q., Xiong, J., Li, S., Zhang, Y., & Mei, H. (2017). IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1, no. 3, pp. 1-22.
68. Lillo, I., Soto, A., & Carlos Niebles, J. (2014). Discriminative hierarchical modeling of spatio-temporally composable human activities. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 812-819.
69. Lin, C. S., Huan, C. C., Chan, C. N., Yeh, M. S., & Chiu, C. C. (2004). Design of a computer game using an eye-tracking device for eye's activity rehabilitation. Optics and lasers in engineering, vol. 42, no. 1, pp. 91-108.
70. Liu, J., Akhtar, N., & Mian, A. (2020). Adversarial attack on skeleton-based human action recognition. IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 4, pp. 1609-1622.
71. Liu, W., Fu, S., Zhou, Y., Zha, Z. J., & Nie, L. (2021). Human activity recognition by manifold regularization based dynamic graph convolutional networks. Neurocomputing, vol. 444, pp. 217-225.

72. Liu, X., Li, Y., & Wang, Q. (2018). Multi-view hierarchical bidirectional recurrent neural network for depth video sequence based action recognition. International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 10, pp. 1850033.
73. Majd, M., & Safabakhsh, R. (2020). Correlational convolutional LSTM for human action recognition. Neurocomputing, vol. 396, pp. 224-229.
74. Malki, Z., Atlam, E., Dagnew, G., Alzighaibi, A. R., Ghada, E., & Gad, I. (2020). Bidirectional residual LSTM-based human activity recognition. Computer and Information Science, vol. 13, no. 3, pp. 40.
75. Maragathavalli, P., Vivitha, M. R., Saranya, M. M., & Dinesh, M. U. (2021, February). Incorporating LSTM Method on Modified Deep Learning Technique for Sensor Based Human Activity Recognition System. In IOP Conference Series: Materials Science and Engineering, Vol. 1065, No. 1, p. 012045.
76. Munoz-Organero, M. (2019). Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs. IEEE Access, vol. 7, pp. 74422-74436.
77. Murad, A., & Pyun, J. Y. (2017). Deep recurrent neural networks for human activity recognition. Sensors, 17(11), 2556.
78. Nikolov, P., Boumbarov, O., Manolova, A., Tonchev, K., & Poulkov, V. (2018, July). Skeleton-based human activity recognition by spatio-temporal representation and convolutional neural networks with application to cyber physical systems with human in the loop. In 2018 41st International conference on telecommunications and signal processing (TSP), pp. 1-5.
79. Nouriani, A., McGovern, R. A., & Rajamani, R. (2022). Deep-learning-based human activity recognition using wearable sensors. IFAC-PapersOnLine, vol. 55, no. 37, pp. 1-6.
80. Park, S. U., Park, J. H., Al-Masni, M. A., Al-Antari, M. A., Uddin, M. Z., & Kim, T. S. (2016). A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. Procedia Computer Science, vol. 100, pp. 78-84.
81. Patron-Perez, A., Marszalek, M., Reid, I., & Zisserman, A. (2012). Structured learning of human interactions in TV shows. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 12, pp. 2441-2453.

82. Paydarfar, A. J., Prado, A., & Agrawal, S. K. (2020, November). Human activity recognition using recurrent neural network classifiers on raw signals from insole piezoresistors. In 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob), pp. 916-921.
83. Pei, L., Guinness, R., Chen, R., Liu, J., Kuusniemi, H., Chen, Y., ... & Kaistinen, J. (2013). Human behavior cognition using smartphone sensors. Sensors, vol. 13, no. 2, pp. 1402-1424.
84. Peng, W., Shi, J., Varanka, T., & Zhao, G. (2021). Rethinking the ST-GCNs for 3D skeleton-based human action recognition. Neurocomputing, vol. 454, pp. 45-53.
85. Pham, H. H., Salmane, H., Khoudour, L., Crouzil, A., Velastin, S. A., & Zegers, P. (2020). A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera. Sensors, vol. 20, no. 7, pp. 1-15.
86. Putra, P. U., Shima, K., & Shimatani, K. (2022). A deep neural network model for multi-view human activity recognition. PloS one, vol. 17, no. 1, pp. e0262181.
87. Queirós, A., Dias, A., Silva, A. G., & Rocha, N. P. (2017, July). Ambient assisted living and health-related outcomes—a systematic literature review. In Informatics, MDPI, Vol. 4, No. 3, pp. 19.
88. Raad, M. W., Sheltami, T., Soliman, M. A., & Alrashed, M. (2018). An RFID based activity of daily living for elderly with Alzheimer's. In Proceedings on October 24-25, 2017 Internet of Things (IoT) Technologies for HealthCare: 4th International Conference, HealthyIoT 2017, Angers, France, vol. 4, pp. 54-61.
89. Ramirez, H., Velastin, S. A., Meza, I., Fabregas, E., Makris, D., & Farias, G. (2021). Fall detection and activity recognition using human skeleton features. IEEE Access, vol. 9, pp. 33532-33542.
90. Ranasinghe, S., Al Machot, F., & Mayr, H. C. (2016). A review on applications of activity recognition systems with regard to performance and evaluation. International Journal of Distributed Sensor Networks, vol. 12, no. 8, pp. 1550147716665520.
91. Ravi, D., Wong, C., Lo, B., & Yang, G. Z. (2016). A deep learning approach to on-node sensor data analytics for mobile or wearable devices. IEEE journal of biomedical and health informatics, vol. 21, no. 1, pp. 56-64.

92. Raziani, S., & Azimbagirad, M. (2022). Deep CNN hyperparameter optimization algorithms for sensor-based human activity recognition. Neuroscience Informatics, vol. 2, no. 3, pp. 100078.
93. Ronao, C. A., & Cho, S. B. (2015). Deep convolutional neural networks for human activity recognition with smartphone sensors. In Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Springer International Publishing, Part IV vol. 22, pp. 46-53.
94. Ryu, J., Patil, A. K., Chakravarthi, B., Balasubramanyam, A., Park, S., & Chai, Y. (2022). Angular features-based human action recognition system for a real application with subtle unit actions. IEEE Access, vol. 10, pp. 9645-9657.
95. Saleem, R., Ahmad, T., Aslam, M., & Martinez-Enriquez, A. M. (2022, October). An Intelligent Human Activity Recognizer for Visually Impaired People Using VGG-SVM Model. In Mexican International Conference on Artificial Intelligence, Cham: Springer Nature Switzerland, pp. 356-368.
96. Sánchez-Caballero, A., Fuentes-Jiménez, D., & Losada-Gutiérrez, C. (2023). Real-time human action recognition using raw depth video-based recurrent neural networks. Multimedia Tools and Applications, vol. 82, no. 11, pp. 16213-16235.
97. Shi, W., Fang, X., Yang, G., & Huang, J. (2022). Human Activity Recognition Based on Multichannel Convolutional Neural Network With Data Augmentation. IEEE Access, vol. 10, pp. 76596-76606.
98. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., ... & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, vol. 56, no. 1, pp. 116-124.
99. Siddiqui, N., & Chan, R. H. (2017, July). A wearable hand gesture recognition device based on acoustic measurements at wrist. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) IEEE, pp. 4443-4446.
100. Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., & Holzinger, A. (2017). Human activity recognition using recurrent neural networks. In Machine Learning and Knowledge Extraction: First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, August 29–September 1, 2017, Proceedings 1, Springer International Publishing, pp. 267-274.

101. Son, H. H. (2017). Toward a proposed framework for mood recognition using LSTM Recurrent Neuron Network. Procedia computer science, vol. 109, pp. 1028-1034.
102. Sonia, Singh, M., Baruah, R. D., & Nair, S. B. (2017). A voting-based sensor fusion approach for human presence detection. In Intelligent Human Computer Interaction: 8th International Conference, IHCI 2016, Pilani, India, December 12-13, 2016, Proceedings on Springer International Publishing, vol. 8, pp. 195-206.
103. Srilakshmi, N., & Radha, N. (2019). Body joints and trajectory guided 3D deep convolutional descriptors for human activity identification. Int. J. Innov. Technol. Explor. Eng., vol. 8, pp. 1016-1021.
104. Srilakshmi, N., & Radha, N. (2021). Deep positional attention-based bidirectional RNN with 3D Convolutional video descriptors for human action recognition. In IOP Conference Series: Materials Science and Engineering, IOP Publishing, Vol. 1022, No. 1, pp. 012017.
105. Stork, J. A., Spinello, L., Silva, J., & Arras, K. O. (2012, September). Audio-based human activity recognition using non-markovian ensemble voting. In 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, IEEE, pp. 509-514.
106. Tapia, E. M., Intille, S. S., & Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. In Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004, Proceedings on Springer Berlin Heidelberg, vol. 2, pp. 158-175.
107. Thongrak, A., Sitjongsataporn, S., Khunkhao, S., & Moungnoul, P. (2019). A practical implementation of memristor emulator circuit based on operational transconductance amplifiers. Int. j. intell. eng. syst, vol. 12, no. 6, pp. 37-46.
108. Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In Advances in neural information processing systems, pp. 1799-1807.
109. Torres-Huitzil, C., & Alvarez-Landero, A. (2015). Accelerometer-based human activity recognition in smartphones for healthcare services. Mobile Health: A Technology Road Map, pp. 147-169.

110. Tran, K. N., Gala, A., Kakadiaris, I. A., & Shah, S. K. (2014). Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. Pattern Recognition Letters, vol. 44, pp. 49-57.
111. Vahora, S. A., & Chauhan, N. C. (2019). Deep neural network model for group activity recognition using contextual relationship. Engineering Science and Technology, an International Journal, vol. 22, no. 1, pp. 47-54.
112. Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. Frontiers in Robotics and AI, vol. 2, pp. 28.
113. Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. Mobile Networks and Applications, vol. 25, pp. 743-755.
114. Wang, H., Yu, B., Xia, K., Li, J., & Zuo, X. (2021). Skeleton edge motion networks for human action recognition. Neurocomputing, vol. 423, pp. 1-12.
115. Wang, Z., Ren, J., Zhang, D., Sun, M., & Jiang, J. (2018). A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing, vol. 287, pp. 68-83.
116. Weiyao, X., Muqing, W., Min, Z., & Ting, X. (2021). Fusion of skeleton and RGB features for RGB-D human action recognition. IEEE Sensors Journal, vol. 21, no. 17, pp. 19157-19164.
117. Xu, C., Chai, D., He, J., Zhang, X., & Duan, S. (2019). InnoHAR: A deep neural network for complex human activity recognition. Ieee Access, vol. 7, pp. 9893-9902.
118. Xu, H., Li, J., Yuan, H., Liu, Q., Fan, S., Li, T., & Sun, X. (2020). Human activity recognition based on Gramian angular field and deep convolutional neural network. IEEE Access, vol. 8, pp. 199393-199405.
119. Yadav, S. K., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2022). Skeleton-based human activity recognition using ConvLSTM and guided feature learning. Soft Computing, vol. 26, no. 2, pp. 877-890.
120. Yan, Y., Ricci, E., Liu, G., & Sebe, N. (2015). Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing, vol. 24, no. 10, pp. 2984-2995.
121. Yang, S. (2013). Shah, 2013 Yang Y., Saleemi I., Shah M. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions,

gestures, and expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 1635-1648.

122. Ye, S., Zeng, H., Fan, J., & Wang, X. (2014, June). Lsi-rec: A link state indicator based gesture recognition scheme in a rfid system. In 2014 9th IEEE Conference on Industrial Electronics and Applications, IEEE, pp. 406-411.
123. Yuan, Y., Yu, B., Wang, W., & Yu, B. (2021). Multi-filter dynamic graph convolutional networks for skeleton-based action recognition. Procedia Computer Science, vol. 183, pp. 572-578.
124. Zebin, T., Scully, P. J., Peek, N., Casson, A. J., & Ozanyan, K. B. (2019). Design and implementation of a convolutional neural network on an edge computing smartphone for human activity recognition. IEEE Access, vol. 7, pp. 133509-133520.
125. Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., & Zhang, J. (2014, November). Convolutional neural networks for human activity recognition using mobile sensors. In 6th international conference on mobile computing, applications and services, IEEE, pp. 197-205.
126. Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., & Li, Z. (2017). A review on human activity recognition using vision-based method. Journal of healthcare engineering.
127. Zhao, K., Qian, C., Xi, W., Han, J., Liu, X., Jiang, Z., & Zhao, J. (2015, November). EMoD: Efficient motion detection of device-free objects using passive RFID tags. In 2015 IEEE 23rd International Conference on Network Protocols (ICNP), IEEE, pp. 291-301.
128. Zhao, Y., Yang, R., Chevalier, G., Xu, X., & Zhang, Z. (2018). Deep residual bidir-LSTM for human activity recognition using wearable sensors. Mathematical Problems in Engineering, 2018, vol. 1-13.
129. Zhu, J., Chen, H., & Ye, W. (2020). A hybrid CNN–LSTM network for the classification of human activities based on micro-Doppler radar. IEEE Access, vol. 8, pp. 24713-24720.

Body Joints and Trajectory Guided 3D Deep Convolutional Descriptors for Human Activity Identification

N. Srilakshmi, N. Radha

Abstract*: *Human Activity Identification (HAI) in videos is one of the trendiest research fields in the computer visualization. Among various HAI techniques, Joints-pooled 3D-Deep convolutional Descriptors (JDD) have achieved effective performance by learning the body joint and capturing the spatiotemporal characteristics concurrently. However, the time consumption for estimating the locale of body joints by using large-scale dataset and computational cost of skeleton estimation algorithm were high. The recognition accuracy using traditional approaches need to be improved by considering both body joints and trajectory points together. Therefore, the key goal of this work is to improve the recognition accuracy using an optical flow integrated with a two-stream bilinear model, namely Joints and Trajectory-pooled 3D-Deep convolutional Descriptors (JTDD). In this model, an optical flow/trajectory point between video frames is also extracted at the body joint positions as input to the proposed JTDD. For this reason, two-streams of Convolutional 3D network (C3D) multiplied with the bilinear product is used for extracting the features, generating the joint descriptors for video sequences and capturing the spatiotemporal features. Then, the whole network is trained end-to-end based on the two-stream bilinear C3D model to obtain the video descriptors. Further, these video descriptors are classified by linear Support Vector Machine (SVM) to recognize human activities. Based on both body joints and trajectory points, action recognition is achieved efficiently. Finally, the recognition accuracy of the JTDD model and JDD model are compared.

Index Terms*: *HAI, Body joints, Optical flow, JDD, JTDD, C3D, SVM

I. INTRODUCTION

HAI is the process of recognizing the actions of a person by using the video sequences which contain a complete action execution and retrieving the videos of interest. It can be used in different applications like video surveillance, human-machine interface, smart home [1], healthcare systems [2], etc. In day-by-day, unrealistic numbers of videos are created because of the surveillance systems, movies, YouTube, etc. As a result, HAI becomes an important research area in recent days. Typically, automatic recognition of human abnormal activities in surveillance systems may support the people to aware the related authority of possible illegal or uncertain characteristics. Similarly, the motion recognition in gaming applications can recover the human-machine interface. In healthcare applications, it can support the patient's rehabilitation like automatic recognition of patient's actions can be utilized to facilitate the rehabilitation processes [3-5].

In the earlier period, a number of researches have been projected for different kind of applications based on HAI. In contrast, perfect identification of activities is still a vastly demanding process owing to noisy environments, occlusions, perspective dissimilarities and so on. Most of the recent techniques may provide specific assumptions about the conditions under which the video was captured. But those assumptions were not often employed in real-time applications. Additionally, the two-step approach has been developed in which the attributes from raw video frames were computed and then obtained features were learned by different classifiers. In real-time applications, it was infrequently recognized what features were considerable. In particular, several activity labels may emerge significantly for HAI in terms of their appearances and action models. As a result, different deep learning models have been proposed to train a hierarchy of attributes by constructing high-level attributes from low-level attributes. Those models can be trained to achieve a reasonable performance in HAI systems.

Cao et al. [6] proposed action recognition with JDD to aggregate convolutional activations of a 3D-CNN into discriminative descriptors according to the joint locales. In this method, the video was split into fixed-length clips. For each clip, 3D convolutional feature maps were computed. The annotated or estimated joints of the video were used for localizing points in the 3D feature maps of a convolution layer. Then, the activations at each related locale were pooled and the pooled activations in a similar clip were concatenated together. After that, average pooling and l_2 normalization were utilized for aggregating snip features into video features. Finally, linear SVM was used for the classification process. Moreover, this process was further extended by obtaining the body joint positions [7]. A two-stream bilinear C3D framework was proposed to train the body joints and extract the spatiotemporal features concurrently. After that, the body joint guided feature pooling was achieved by sampling in which the pooling method was devised as a bilinear product function. However, the time consumption for estimating the locales of body joints by using vast data and computational cost of skeleton estimation algorithm were high. Also, the recognition accuracy was needed to be further improvement in an efficient manner.

Revised Manuscript Received on October 10, 2019

N. Srilakshmi, Ph.D Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India.

Dr. N. Radha, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India.

Hence, an optical flow extraction is integrated into the two-stream bilinear model efficiently to improve the recognition accuracy. According to this model, optical flows i.e., trajectory points between two video sequences at each body joint positions are extracted automatically. To achieve this, two C3D streams multiplied with the bilinear product is used for extracting the features, generating the pooled descriptors for video sequences and capturing the spatiotemporal features. After that, the whole network is trained for obtaining the video descriptors based on the two-stream bilinear C3D framework that uses the class label. Finally, the linear SVM is used to sort the obtained video descriptors for recognizing human actions. Thus, action recognition performance is improved by considering the optical flow field with body joints efficiently.

II. LITERATURE SURVEY

Ji et al. [8] proposed a 3D-CNN framework in which the spatiotemporal features were extracted by 3D convolutions. Also, the feature from all channels was combined to represent the absolute feature. Then, the framework was regularized by high-level features. Moreover, the outputs of different frameworks were combined for boosting recognition performance. This was evaluated on KTH dataset and the obtained recognition accuracy was 90.2%. Conversely, a number of labelled features were required since it uses a supervised algorithm i.e., gradient-based learning to train this model.

Karpathy et al. [9] proposed a large-scale video classification with CNN to recognize the YouTube videos. In this method, the connectivity of a CNN was extended in a time-domain based on different approaches that take advantage of local spatiotemporal information. The training process was speed-up by suggesting a multi-resolution. The performance analysis was done by using UCF-101 dataset and the recognition accuracy achieved was 65.4%. However, there is a need for improvement on the action recognition. The efficiency can be further improved by combining the clip-level predictions into the global video-level predictions.

Lillo et al. [10] proposed a discriminative hierarchical framework. By using this approach, the human activity classifier was built to simultaneously model which body parts were related to the action of interest with their appearance and composition. Also, when useful annotations were provided at the intermediate semantic level, powerful multiclass discrimination was achieved by learning in a max-margin model. For performance evaluation, two datasets, namely MSR Action3D and CAD120 datasets were used. The recognition accuracy of MSR Action3D and CAD120 dataset was 89.46% and 33.59%, respectively. But, the accuracy of this model was less.

Tompson et al. [11] proposed new hybrid architecture by using a Deep Convolutional Network (ConvNet) and a Markov Random Field (MRF). In this model, a multi-resolution feature representation was used with overlapping fields. Also, this model can approximate MRF loopy belief propagation which was subsequently back-propagated through and learned by using the same learning method as the part-Detector. The recognition accuracy of this model was evaluated on two different datasets such as FLIC and extended-LSP datasets for elbow and wrist joints. For elbow joints, the accuracy of FLIC and extended-LCP datasets was 95% and 66%, respectively. The accuracy for wrist joints using FLIC and extended-LCP datasets were 91% and 62%, correspondingly. However, further improvement of its performance was required.

Cao et al. [12] proposed a spatiotemporal Triangular-chain Conditional Random Field (TriCRF) model for activity recognition. Initially, the difficulty of complex motion identification with an integrated hierarchical framework was addressed. Then, the TriCRF model was expanded to the spatial dimension. In this model, the labels of behaviour were modeled together and their complex dependencies were developed. The accuracy of this model was evaluated on composable activity dataset which was equal to 79%. However, it requires further improvement by incorporating the other layer for learning pose representations jointly with actions and activity.

Wang et al. [13] proposed an action recognition model with the help of Trajectory-pooled Deep-convolutional Descriptor (TDD). In this model, discriminative convolutional feature maps were learned by deep architectures and aggregated into valuable descriptors by trajectory-constrained pooling. As well, two normalization methods such as spatiotemporal normalization and channel normalization were used that transforms convolutional feature maps and enhance the robustness of TDD. The evaluated accuracies for this model using SVM classifier on HMDB51 and UCF101 datasets were 65.9% and 91.5%, respectively. However, body joints were not considered that can help to increase the accuracy efficiently.

Liu et al. [14] proposed an automatic learning of spatiotemporal representation using Genetic Programming (GP) for action recognition. In this model, the spatiotemporal motion features were automatically learned by the motion feature descriptor. The data-adaptive descriptors were learned for various databases with multiple layers and the GP searching space was simultaneously reduced for effectively accelerating the convergence of optimal solutions. The average cross-validation classification error computed by SVM classifier on the training dataset was adopted as the validation measure for GP fitness function. Then, the best-so-far result chosen by GP was obtained as the optimal action descriptor. The accuracy for this model on KTH, YouTube, Hollywood2 and HMDB51 datasets were 95%, 82.3%, 46.8% and 48.4%, respectively. Nevertheless, the processing speed was less.

III. PROPOSED METHODOLOGY

In this part, the JTDD methodology is explained in detail. The two-stream bilinear C3D network framework is applied for automatically predict the spatiotemporal key points in 3D convolutional feature maps with the guidance of body joints with optical flow i.e., trajectory points in the video sequence.

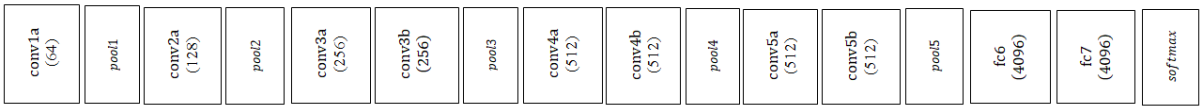

Fig. 1: Architecture of C3D Network

A. Joints and Trajectory-Pooled 3D Deep Convolutional Descriptors

In this process, the C3D network [4] shown in Fig. 1 is used which has architecture as:

$$conv1a(64) - pool1 - conv2a(128) - pool2 - conv3a(256) - conv3b(256) - pool3 - conv4a(512) - conv4b(512) - pool4 - conv5a(512) - conv5b(512) - pool5 - fc6(4096) - fc7(4096) - softmax$$

Here, the number in parenthesis indicates the number of convolutional filters. The number of filters is increased since the combination of the features in each layer is richer than the previous layers. Therefore, increasing number of filters can able to correctly encode the increasingly richer representations of the features. There is a ReLu layer after each convolutional $(conv)$ layer.

Body Joints and Optical Flow Mapping Schemes:

For JTDD, two methods of mapping body joints and optical flow to points in 3D convolutional feature maps, namely fraction scaling and coordinate mapping are compared. Fraction scaling is defined as the fraction of the network's outcome to its input in spatiotemporal-domain for scaling the body joint and optical flow coordinates from the actual video frame into feature maps as follows:

$$(x_c^i, y_c^i, t_c^i) = \left(\overline{(r_x^i \cdot x_v)}, \overline{(r_y^i \cdot y_v)}, \overline{(r_t^i \cdot t_v)}\right) \quad (1)$$

$$(l_c^i, m_c^i, n_c^i) = \left(\overline{(r_l^i \cdot l_v)}, \overline{(r_m^i \cdot m_v)}, \overline{(r_n^i \cdot n_v)}\right) \quad (2)$$

In (1) & (2), $\overline{(\cdot)}$ denotes the rounding operator and (x_c^i, y_c^i, t_c^i) represents the point coordinate in i^{th} 3D convolutional feature maps corresponding to (x_v, y_v, t_v) which is the body joint coordinate in the actual video sequence and (r_x^i, r_y^i, r_t^i) represents the size ratio of i^{th}convolutional feature maps to the video clip in spatial and temporal dimensions. Similarly, (l_c^i, m_c^i, n_c^i) represents the point coordinate in i^{th} 3D convolutional feature maps corresponding to (l_v, m_v, n_v) which is the trajectory point coordinate in the actual video sequence and (r_l^i, r_m^i, r_n^i) represents the fraction of i^{th} convolutional feature maps to the video snip in spatiotemporal-domain.

Coordinate mapping is computing an exact coordinate of the point at the convolutional feature map corresponding to body joint and trajectory point based on the kernel size, stride and padding of each layer. Consider p_i is a point in i^{th} layer, (x_i, y_i, t_i) and (l_i, m_i, n_i) are the coordinate ofp_i. For a givenp_i, the corresponding point p_{i+1} is determined by mapping p_i to the $(i+1)^{th}$ layer. For the convolutional layers and pooling layers, the coordinate mapping from i^{th} layer to $(i+1)^{th}$ layer is developed as follows:

$$x_{i+1} = \frac{1}{s_i^x}\left(x_i + padding_i^x - \frac{k_i^x - 1}{2}\right) \quad (3)$$

$$y_{i+1} = \frac{1}{s_i^y}\left(y_i + padding_i^y - \frac{k_i^y - 1}{2}\right) \quad (4)$$

$$z_{i+1} = \frac{1}{s_i^z}\left(z_i + padding_i^z - \frac{k_i^z - 1}{2}\right) \quad (5)$$

$$l_{i+1} = \frac{1}{s_i^l}\left(l_i + padding_i^l - \frac{k_i^l - 1}{2}\right) \quad (6)$$

$$m_{i+1} = \frac{1}{s_i^m}\left(m_i + padding_i^m - \frac{k_i^m - 1}{2}\right) \quad (7)$$

$$n_{i+1} = \frac{1}{s_i^n}\left(n_i + padding_i^n - \frac{k_i^n - 1}{2}\right) \quad (8)$$

In the above equations, s_i^x, k_i^x and $padding_i^x$ are the $x-$axis element of stride, kernel size and padding of i^{th} layer, correspondingly. Likewise, this is also applied for other dimensions such as y, z, l, m and n, correspondingly.

For ReLU layers, the coordinate mapping correlation is formulated as:

$$(x_{i+1}, y_{i+1}, z_{i+1}) = (x_i, y_i, t_i) \quad (9)$$

$$(l_{i+1}, m_{i+1}, n_{i+1}) = (l_i, m_i, n_i) \quad (10)$$

Once the values of C3D kernel sizes, strides and paddings are applied into (3)-(5) and (9) frequently, the correlation between point coordinates in i^{th} convolutional feature maps and body joint locales in the input video sequence is devised as follows:

$$(x_c^i, y_c^i) = \frac{1}{2^{i-1}} \cdot \left(x_v - \frac{2^{i-1}-1}{2}, y_v - \frac{2^{i-1}-1}{2}\right) \quad (11)$$

$$t_c^i = \frac{1}{2^{i-2}} \cdot \left(t_v - \frac{2^{i-2}-1}{2}\right) \quad (12)$$

Similarly, by applying the values of kernel size, strides and paddings into (6)-(8) and (10) repeatedly, the correlation between point coordinates in i^{th} convolutional feature maps and trajectory points in the input video sequence is devised as follows:

$$(l_c^i, m_c^i) = \frac{1}{2^{i-1}} \cdot \left(l_v - \frac{2^{i-1}-1}{2}, l_v - \frac{2^{i-1}-1}{2}\right) \quad (13)$$

$$n_c^i = \frac{1}{2^{i-2}} \cdot \left(n_v - \frac{2^{i-2}-1}{2}\right) \quad (14)$$

Aggregation of Body Joint Points and Optical Flow:

For classification, the extracted features of frames over time are required to aggregate for obtaining the video descriptor. The positions to pool can be determined by employing body joints and trajectory points in video frames to localize points in 3D feature maps. The pooled representation corresponding to each body joint and trajectory point in a frame of a video sequence is a F dimensional feature vector where F denotes the number of feature map channels. The F dimensional feature vector pooled with the guidance of i^{th} body joint at the t^{th} frame of k^{th} clip is denoted by$f_k^{i,t}$. Similarly, the F dimensional feature vector pooled with the guidance of i^{th} trajectory point at the t^{th} frame of k^{th} clip is represented by$g_k^{i,t}$.

Two methods are used for aggregating the pooled feature vectors in all the frames within a video to a video descriptor. One is fusing all the pooled feature vectors belonging to one frame i.e., a $\mathrm{F} \times N \times O \times L$ dimensional feature as:

$$f_k g_k =$$

$$\begin{bmatrix} f_k^{1,1} g_k^{1,1}, f_k^{2,1} g_k^{2,1}, \dots, f_k^{N,1} g_k^{O,1}, f_k^{1,2} g_k^{1,2}, \\ f_k^{2,2} g_k^{2,2}, \dots, f_k^{N,2} g_k^{O,2}, \dots, f_k^{N,L} g_k^{O,L} \end{bmatrix} \quad (15)$$

In (15), N and O represent the number of body joints and trajectory points in each frame, correspondingly and L denotes the length of the video sequence. After that, average pooling and $L2$ norm are used to fuse k frame representations $\{f_1 g_1, f_2 g_2, \dots, f_k g_k\}$ into a video descriptor where k denotes the number of frames within the video sequence.

Another method of aggregation is fusing the pooled feature vectors corresponding to the body joints and trajectory points in one frame i.e., a $\mathrm{F} \times N \times O$ dimensional feature as:

$$f_k^t g_k^t = \left[f_k^{1,t} g_k^{1,t}, f_k^{2,t} g_k^{2,t}, \dots, f_k^{N,t} g_k^{O,t} \right] \quad (16)$$

After that, one frame is characterized by L representations $\{f_k^1 g_k^1, f_k^2 g_k^2, \dots, f_k^L g_k^L\}$ within the same frame. Max + min pooling is used for aggregating these representations into a frame descriptor.

B. Two-Stream Bilinear C3D Model using Body Joints and Optical Flow

The original body joint and optical flow guided feature pooling in JTDD are realizing by choosing the activations at the corresponding points of body joints and trajectory points on convolutional feature maps. For a given video sequence, a M channel of heat maps $(M = N \times O \times L)$ is generated with the similar spatiotemporal size of the convolutional feature maps to be pooled for each body joint and trajectory point at each frame. In the heat map, the value at the corresponding point of the body joint position and trajectory point is coded as 1, while the others are coded as 0. After that, the process of pooling on one feature map guided by the heat map of one body joint and trajectory point can be formulated as a pixel-wise product between the 3D feature map and the 3D heat map followed by a summation over all the pixels. After that, a two-stream bilinear C3D model is applied to learn the guidance from the body joint positions including trajectory points and capture the spatiotemporal features automatically [4].

Thus, by integrating trajectory points with body joint positions in the two-stream bilinear framework, video descriptors are obtained. Finally, linear SVM [15] is applied for classifying the video descriptors and so the human actions from video sequences are recognized.

Normally, the SVM is built as a hyperplane in an infinite-dimensional space. A perfect HAI is achieved by the hyperplane which has the leading space to the adjacent training sequences of any class i.e., functional margin. The training dataset is represented as a set of instance-label pairs$(x_i, y_i), i = 1, \dots, n, x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$ where x_i denotes the video descriptors (instances) and y_i denotes the labels. The optimal hyperplane with the maximal margin is achieved by resolving the below unconstrained optimization problem for different classes:

$$\min_w \frac{1}{2} w^T w + \mathrm{F} \sum_{i=1}^{n} \xi(w; x_i, y_i) \quad (17)$$

In (17), $\mathrm{F} > 0$ denotes the penalty parameter and w denotes the weight of training sequencesx_i. By solving this optimization problem, human activities are recognized.

IV. RESULTS AND DISCUSSIONS

In this part, the efficiency of JTDD model is analyzed with the JDD model in terms of recognition accuracy by using Matlab2017b. To evaluate the performance, Penn Action dataset is considered which contains 2326 video sequences of 15 action classes. Here, 50% dataset is taken as the training and the rest 50% is considered as testing dataset. For training an attention model, the dataset is splitting into 1163 training and 1163 testing, randomly. The length of videos is from 50 to 100 frames. The body joint coordinates, trajectory points and C3D features are acted as baselines. Therefore, JTDD with these features is evaluated and compared with different pooling settings. The recognition accuracy is the portion of True Positive (TP) and True Negative (TN) rates among the total number of cases. It is computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

In (18),FP is False Positive and FN is the False Negative. The results of body joint and trajectory point extraction are shown in Fig. 2.

Fig. 2(a): Input Video Sequence 1

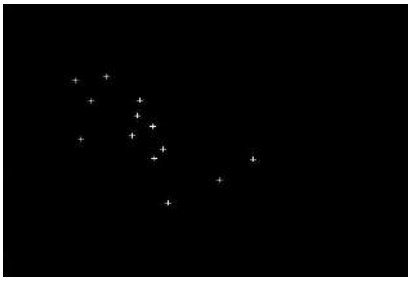

Fig. 2(b): Results for Body Joints Extraction of Input Video Sequence 1

Fig. 2(c): Results for Trajectory Points Extraction of Input Video Sequence 1

Fig. 2(d): Input Video Sequence 2

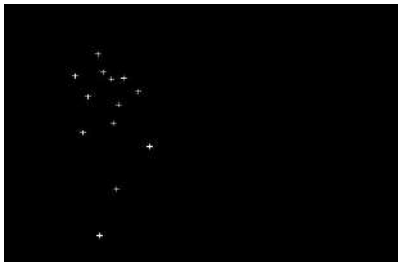

Fig. 2(e): Results for Body Joints Extraction of Input Video Sequence 2

Fig. 2(f): Results for Trajectory Points Extraction of Input Video Sequence 2

In below table, the outcomes on Penn Action dataset are given.

Table 1: Recognition Accuracy of Baselines and JTDD with Various Configurations

	Fuse all the activations	JTDD Fraction Scaling (1×1×1)	JTDD Coordinate Mapping (1×1×1)	JTDD Fraction Scaling (3×3×3)	JTDD Coordinate Mapping (3×3×3)
Joint coordinates + trajectory coordinates	0.6120	-	-	-	-
$fc7$	0.7211	-	-	-	-
$fc6$	0.7368	-	-	-	-
$conv5b$	0.7052	0.8014	0.8599	0.8086	0.8367
$conv5a$	0.6305	0.7583	0.7834	0.7533	0.7628
$conv4b$	0.5324	0.7697	0.7601	0.7847	0.7993
$conv3b$	0.4297	0.7136	0.6845	0.7021	0.7014

From this analysis, it is observed that the accuracy of using the coordinates of body joints and optical flow i.e., trajectory points as a feature is not effective. By using the C3D features which are fusing all the activations of a particular layer as a long vector are highly discriminative because they attain high outcomes. The recognition accuracy of $fc7$ is slightly poorer to that of $fc6$. It is perhaps since the original C3D on Penn Action dataset do not fine-tune that the second fc layer is more fit for the classification of the pre-trained database. For JTDD, the testing on pooling at different 3D $conv$ layers with different body joint and trajectory point mapping formats are publicized.

From Table 1, it is noticed that the JTDD have superior performance compared with C3D features that express the efficiency of body joint and trajectory point guided pooling. The outcomes of various fusing combinations with the scores of SVM on Penn Action dataset is shown in Table 2 and Fig. 3.

Table 2: Recognition Accuracy of Fusing JTDD from Multiple Layers Together with the Scores of SVM

	Fusion Layers					
	$conv5b + fc6$		$conv5b + conv4b$		$conv5b + conv3b$	
	JDD	JTDD	JDD	JTDD	JDD	JTDD
Accuracy	0.855	0.867	0.981	0.987	0.860	0.873

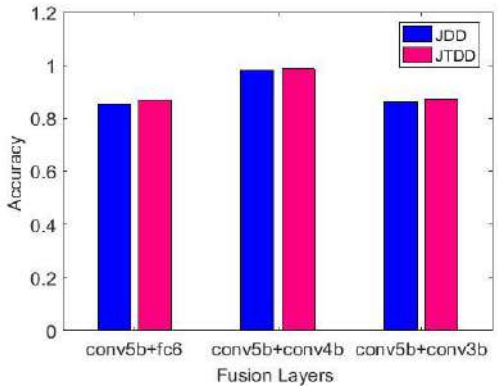

Fig.3: Recognition Accuracy of Fusing JTDD from Multiple Layers

In Fig. 3, it is indicated that fusing JTDDs of different layers certainly increases the recognition outcomes. The combination of JTDDs from $conv5b$ and $conv4b$ increases the recognition performance mostly. High accuracy is achieved by fusing more complementary features.

The results of the impact of estimated body joints + trajectory points versus ground-truth body joints + trajectory points for JDD and JTDD is shown in Table 3 and Fig. 4.

Table 3: Impact of Estimated Body Joints + Trajectories versus Ground-Truth (GT) Body Joints + Trajectories for JDD and JTDD

Method	GT	Estimated	Difference
JDD ($conv5b$)	0.819	0.777	0.042
JTDD ($conv5b$)	0.835	0.810	0.025

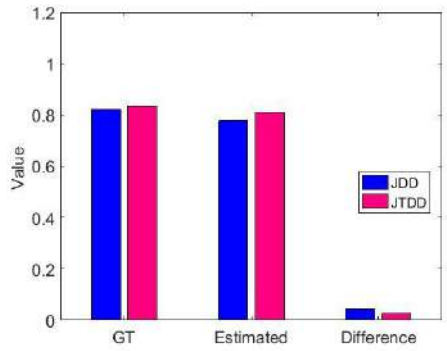

Fig.4: Impact of Estimated Body Joints + Trajectories versus GT Body Joints + Trajectories for JDD and JTDD on Penn Action Dataset

Through Fig. 4, it is noticed that JTDD outperforms competing methods significantly on Penn Action Dataset. JTDD achieves better accuracy not only with GT body joints and trajectory points, but also with estimated body joints and trajectory points, greater than JDD in the order of 10%.

V. CONCLUSION

In this article, JTDD is proposed to extract the optical flow at each body joint positions as the inputs of a C3D model by using two streams of C3D networks which are multiplied with the bilinear product. Based on this, the pooled descriptors for video sequences are generated together and the spatiotemporal features are captured. After that, the entire network is trained end-to-end by using the class label of the two-stream bilinear C3D model to obtain the video descriptors. Moreover, the linear SVM is used to classify the video descriptors for HAR. Finally, the experimental results prove that the recognition accuracy of the proposed JTDD model using Penn Action dataset is increased to 0.987 while fusing JTDDs from $conv5b$ and $conv4b$ with GT body joints and trajectory points. This framework can be applicable for real-time applications such as surveillance, theft identification, motion identification, etc.

REFERENCES

1. A. Sukor, A. Syafiq, A. Zakaria, N. A. Rahim, L. M. Kamarudin, R. Setchi and H. Nishizaki, "A hybrid approach of knowledge-driven and data-driven reasoning for activity recognition in smart homes," J. Intell. Fuzzy Syst., vol. 36, pp. 4177-4188, 2019.
2. J. X. Qiu, H. J. Yoon, P. A. Fearn and G. D. Tourassi, "Deep learning for automated extraction of primary sites from cancer pathology reports," IEEE J. Biomed. Health Inform., vol. 22, no. 1, pp. 244-251, 2018.
3. Y. Kong and Y. Fu, "Human action recognition and prediction: a survey," J. LATEX Cl. Files, vol. 13, no. 9, pp. 1-20, 2018.
4. H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du and D. S. Chen, "A comprehensive survey of vision-based human action recognition methods," Sens., vol. 19, no. 5, p. 1005, 2019.
5. S. R. Ke, H. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo and K. H. Choi, "A review on video-based human activity recognition," Comput., vol. 2, no. 2, pp. 88-131, 2013.
6. C. Cao, Y. Zhang, C. Zhang and H. Lu, "Action recognition with joints-pooled 3D deep convolutional descriptors," in Proc.25th Int. Jt. Conf. Artif. Intell., vol. 1, p. 3, 2016.
7. C. Cao, Y. Zhang, C. Zhang and H. Lu, "Body joint guided 3-D deep convolutional descriptors for action recognition," IEEE Trans. Cybern., vol. 48, no. 3, pp. 1095-1108, 2018.
8. S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221-231, 2013.
9. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1725-1732, 2014.
10. I. Lillo, A. Soto and J. Carlos Niebles, "Discriminative hierarchical modeling of spatio-temporally composable human activities," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 812-819, 2014.
11. J. J. Tompson, A. Jain, Y. LeCun and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in Adv. Neural Inf. Process. Syst., pp. 1799-1807, 2014.
12. C. Cao, Y. Zhang and H. Lu, "Spatio-temporal triangular-chain CRF for activity recognition," in Proc. 23rd ACM Int. Conf. Multimed., pp. 1151-1154, 2015.
13. L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4305-4314, 2015.
14. L. Liu, L. Shao, X. Li and K. Lu, "Learning spatio-temporal representations for action recognition: a genetic programming approach," IEEE Trans. Cybern., vol. 46, no. 1, pp. 158-170, 2016.
15. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang and C. J. Lin, "LIBLINEAR: A library for large linear classification," J. Mach. Learn. Res., vol. 9, no. Aug, pp. 1871-1874, 2008.

AUTHORS PROFILE

N. Srilakshmi completed MCA and M.Phil degree in Computer Science from Bharathiyar University in the year 2005 and 2008, respectively. Currently, she is working as a guest lecturer of Computer Science in Government Arts College, Udhagamandalam, affiliated by Bharathiyar University. She has 11 years of teaching experience. Her area of interests is data mining and pattern recognition.

Dr. N. Radha, working as an Associative Professor in the department of Computer Science (PG) at PSGR Krishnammal College for Women, Coimbatore. She has 22 years of teaching experience. She has more than 35 publications in both national/international journals. She has guided more than 25 M.Phil scholars. Her research area includes data mining, biometric and information security.

Deep Positional Attention-based Bidirectional RNN with 3D Convolutional Video Descriptors for Human Action Recognition

To cite this article: N Srilakshmi and N Radha 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1022** 012017

Deep Positional Attention-based Bidirectional RNN with 3D Convolutional Video Descriptors for Human Action Recognition

N Srilakshmi[1] and N Radha[2]

[1]Ph.D. Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India
[2]Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India

srilakshmi.mphil@gmail.com

Abstract. This article presents the Joints and Trajectory-pooled 3D-Deep Positional Attention-based Bidirectional Recurrent convolutional Descriptors (JTPADBRD) for recognizing the human activities from video sequences. At first, the video is partitioned into clips and these clips are given as input of a two-stream Convolutional 3D (C3D) network in which the attention stream is used for extracting the body joints locations and the feature stream is used for extracting the trajectory points including spatiotemporal features. Then, the extracted features of each clip is needed to aggregate for creating the video descriptor. Therefore, the pooled feature vectors in all the clips within the video sequence are aggregated to a video descriptor. This aggregation is performed by using the PABRNN that concatenates all the pooled feature vectors related to the body joints and trajectory points in a single frame. Thus, the convolutional feature vector representations of all the clips belonging to one video sequence are aggregated to be a descriptor of the video using Recurrent Neural Network (RNN)-based pooling. Besides, these two streams are multiplied with the bilinear product and end-to-end trainable via class labels. Further, the activations of fully connected layers and their spatiotemporal variances are aggregated to create the final video descriptor. Then, these video descriptors are given to the Support Vector Machine (SVM) for recognizing the human behaviors in videos. At last, the experimental outcomes exhibit the considerable improvement in Recognition Accuracy (RA) of the JTDPABRD is approximately 99.4% achieved on the Penn Action dataset as compared to the existing methods.

1. Introduction

Human Activity Recognition (HAR) is the method of using the videos that include a specific activity and recover videos of interest to identify an individual's conduct. This has been applied in potential fields including video processing, human-computer interface design, medical services and so on. By the day, an incredible number of videos are generated due to monitoring devices, media, YouTube and others. Correspondingly, HAR is significant in the field of machine learning in the current era. Many deep learning models were suggested based on either supervised or unsupervised learning algorithms that can support HAR systems [1-3].

Among many deep learning methods, Joints-pooled 3D-Deep convolutional Descriptors (JDD) [4-5] has better efficiency by aggregating the convolutional activations of the 3D-deep Convolutional Neural

Network (3DCNN) into the discriminative descriptors based on the joint locations. On the contrary, the estimation of joints locations takes more time for a huge dataset and also the estimation of skeletons has a high complex. As a result, Joints and Trajectory-pooled 3D-Deep convolutional Descriptors (JTDD) is suggested [6] that extracts both body joints and trajectory points between two video sequences by multiplying two C3D streams: feature and attention with the bilinear product function. Also, the pooled descriptors are generated to extracting the spatiotemporal features together. Then, the video descriptors are obtained by training the whole network in an end-to-end manner according to the class labels. Moreover, these video descriptors are classified via the SVM to recognize individual behaviors. Nonetheless, the max-min poling was applied as the feature aggregation method that has high flexibility to spatially smooth over the adjacent kernels. This eliminates the necessary spatiotemporal variances between class labels.

Therefore in this article, JTDPABRD is proposed that integrates the PABRNN model into a two-stream C3D network to extract significant spatiotemporal features and increase the accuracy of recognizing individual activities. Initially, the video is split into many clips and these clips are fed to the two-stream C3D network as input. In a two-stream C3D network, the attention stream is used to extract the guidance of body joints locations and the feature stream is used to extract the trajectory points along with significant spatiotemporal features. After, each convolutional feature vector representations of each clip belonging to the single video are aggregated using the PABRNN to create the clip descriptor. Also, these two streams are multiplied by the bilinear product and end-to-end trained via class labels. Moreover, the activations of fully connected layers and their spatiotemporal variances are also aggregated to generate the final video descriptor. This video descriptor is applied to the SVM to recognize the individual activities in video sequences. Thus, the accuracy of recognizing human activities is increased efficiently.

2. Literature Survey

Rahman et al. [7] investigated the HAR system using textural features with classical shape and motion features from low-quality videos. But, it needs to learn the richer features from video sequences for enhancing the performance. Li et al. [8] proposed a novel two-layer framework for HAR via defining the video with low-level local and mid-level motion features. However, several groups in the video were not represented the activity part and it cannot determine the number of groups in various datasets which affects the efficiency of mid-level encoding.

Jin et al. [9] proposed a multilevel action descriptor that provides absolute information on human activities. But, it was not able to learn deep motion flow from the video sequences. Shou et al. [10] suggested a lightweight generator network to get more Discriminative Motion Cue (DMC) for HAR. Conversely, it has less accuracy. Huo et al. [11] proposed the new mobile HAR system, but it needs to consider the attention scheme for further improving the accuracy.

Nida et al. [12] proposed a feedforward learning method for recognizing the instructor's action in the classroom. However, an overfitting problem occurred while increasing the hidden layers. Sudhakaran et al. [13] proposed a Long Short-Term Attention (LSTA) to extract the features from relevant spatial parts and recognize the egocentric activity. But, the accuracy was less.

3. Proposed Methodology

This section explains the JTDPABRD method in brief. The block diagram of the JTDPABRD method is depicted in Figure 1.

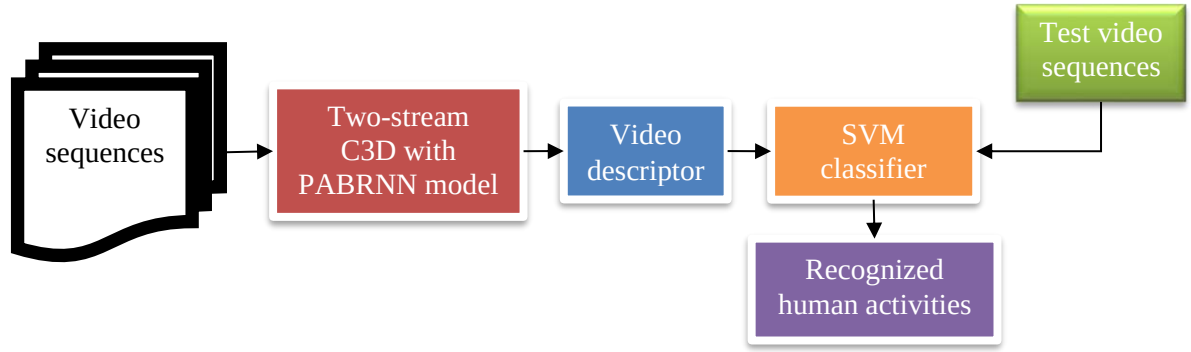

Figure 1. Block diagram of JTDPABRD based HAR system.

Originally, each video sequence is split into many clips or frames and given as input to the two-stream C3D network. In this network, the input is given to the attention stream and feature stream, accordingly. The attention stream is used for extracting the guidance of body joint locations and the feature stream is used for extracting the trajectory points or optical flow between each clip including spatiotemporal features. The activations of each corresponding body joint location and trajectory point are pooled from each channel. To obtain the pooled feature vectors belonging to one clip, RNN, namely JTDRD method is applied instead of max-min pooling. But, the standard RNN has a problem of how to aggregate network outputs in an optimized manner as various networks trained on similar data can no longer be regarded as independent. Therefore, JTD-Bidirectional RNN-Descriptor (JTDBRD) is applied to solve the problems in the standard RNN and trained using all available input information in the past and future of particular time frames i.e., video clips. The concept is splitting the state neurons of a standard RNN in a part for both forward and backward states. The results from forward-states are not linked to inputs of backward-states and vice versa. Using both states, input information in the previous and the future of the currently estimated frames can be directly used for reducing the objective function without the requirement for delays to include future information.

This BRNN can be trained with similar algorithms as a standard RNN since there are no interactions between two kinds of state neurons and so can be extended into the common feed-forward network, Few specific solutions are required only at the beginning and the end of training samples. The forward-state input at $t = 1$ and the backward-state inputs $t = T$ are not observed. But, they are set randomly to a predetermined value (0.5). Also, the local state derivatives at $t = T$ for the forward-states and at $t = 1$ for the backward-states are not known and are set to 0, considering that the information beyond that point is not significant for the current update. On the other hand, BRNN cannot be used for providing significant feature vectors with the highest likelihood. Also, the problem of BRNN is how to aggregate the hidden vectors for feature representations. As a result, PABRNN is proposed in this JTDPABRD method that assumes if a feature in one video frame occurs in another video frame, it will have guidance on the adjacent context. In other words, the adjacent features should be given more attention that those far away since they may include more body joint and trajectory relevant information. The entire trainable end-to-end two-stream C3B with the PABRNN framework is depicted in Figure 2.

Figure 2. Block diagram of two-stream bilinear C3D with PABRNN-based feature aggregation method.

3.1. PABRNN model

This PABRNN adopts the BRNN for feature vector representation which takes the pre-trained body joints and trajectory points embeddings as the input and creates the hidden vectors by recurrent updates. To aggregate the feature vector representations, the standard attention is used which especially relies on the hidden vectors for the attentive weight generation. For this purpose, a positional attention scheme is proposed and additional steps are performed based on the standard attention as:

- Discover the occurrence feature positions in each clip related to a single video sequence.
- Propagate the guidance of feature vectors to other positions with a position-aware guidance propagation approach.
- Create the position-aware guidance vector for every feature in clips according to the propagated guidance.
- Combine the position-aware guidance vector into the standard attention scheme.

By using the attentive representations of both original and aggregated feature vectors, different similarity functions are used for measuring the relevance between each dimension. The Manhattan distance similarity function (sim) is used with l_1-norm as:

$$sim(F, F_a) = e^{-(\|F - F_a\|_1)} \quad (1)$$

In Eq. (1), F and F_a are the original feature vector and aggregated feature vectors corresponding to in each clip and $\|\cdot\|_1$ is the l_1-norm. The structure of this PABRNN model is illustrated in Figure 3.

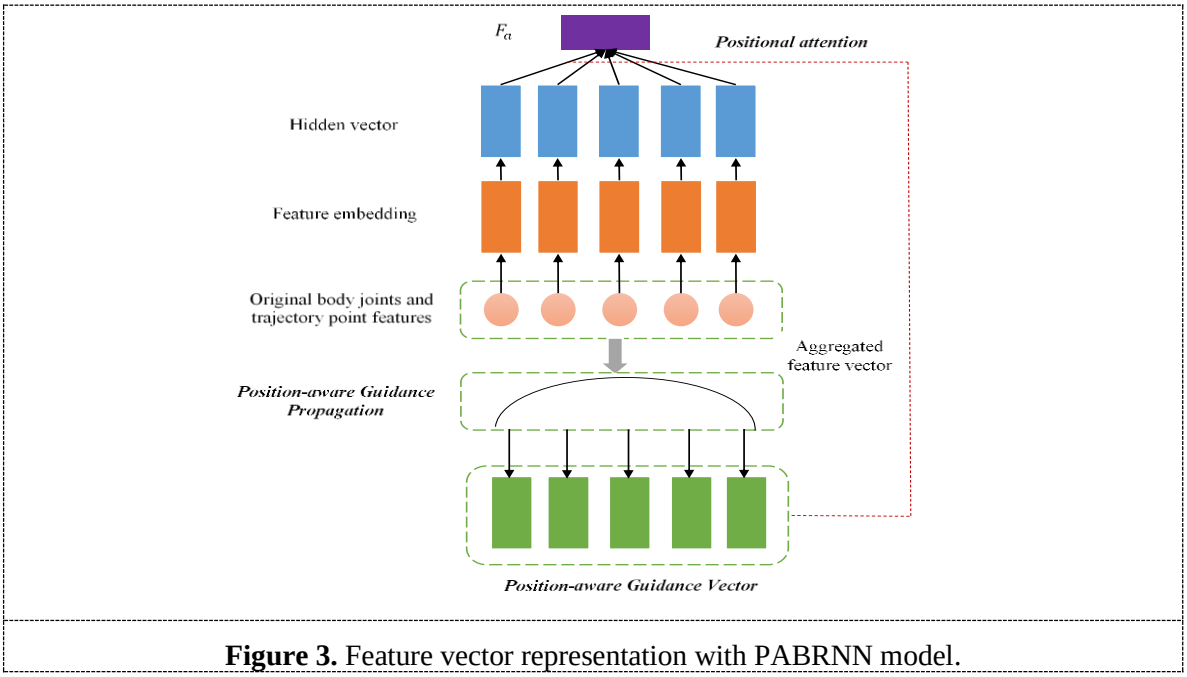

Figure 3. Feature vector representation with PABRNN model.

3.2. Position-aware guidance propagation

Based on the above consideration, the features will have guidance on the adjacent context if it occurs in other clips. Here, the position-aware guidance propagation is modeled with the Gaussian kernel as:

$$Kernel(d) = e^{\left(-d^2/2\sigma^2\right)} \quad (2)$$

In Eq. (2), d is the distance between the original and aggregated features, σ is a parameter that constraints the propagation scope and $Kernel(d)$ is the obtained guidance related to the distance of d based on the kernel. Observe that the position-aware guidance is fading while the distance increases. Particularly, when $d = 0$, the maximum propagated guidance is obtained. Here, a fixed σ value is applied for all feature vectors and focused on combining the positional context into attentions.

3.3. Position-aware guidance vector

The guidance in a high-dimensional space for attention is modeled by obtaining the position-aware guidance vector for each feature vector in the video clips. Initially, consider the guidance for a particular distance follows the Gaussian distributions over the hidden dimensions. After, a guidance base matrix G is defined based on the assumption where each column is the guidance base vector related to the particular distance. Each element of G is described as follows:

$$G(i,d) \sim N(Kernel(d), \sigma') \quad (3)$$

In Eq. (3), $G(i,d)$ is the guidance related to the distance of d in the i^{th} position and N is the normal density with a predicted value of $Kernel(d)$ and standard variance of σ'. Using the guidance base matrix, the guidance vector for a feature at a particular position is obtained by aggregating the guidance of all features occurring in the video clips:

$$A_j = Gc_j \quad (4)$$

In Eq. (4), A_j is the aggregated guidance vector for the feature at position j and c_j is the distance count vector which estimates the count of features with different distances. Particularly, for the feature at position j, the count of body joint and trajectory point features with a distance of d i.e., $c_j(d)$ is computed as follows:

$$c_j(d) = \sum_{f \in F} [(j-d) \in pos(f)] + [(j+d) \in pos(f)] \quad (5)$$

In Eq. (5), F is the 3D feature maps containing multiple features, f is either a body joint location or a trajectory point feature in F, $pos(f)$ is the group of f's occurrence positions in different clips and $[\cdot]$ is an indicator function which equals to 1 if the criteria satisfy, or else equals to 0.

3.4. Positional attention

A positional attention scheme is proposed that incorporates the position-aware guidance of the features into the aggregated feature's attentive representations. In particular, the attentive weight of a feature at position j in the aggregated feature vector is formulated as:

$$\alpha_j = \frac{e^{\left(e\left(h_j, A_j\right)\right)}}{\sum_{k=1}^{l} e^{\left(e\left(h_k, A_k\right)\right)}} \quad (6)$$

In Eq. (6), h_j is the hidden vector at position j based on BRNN, A_j is the aggregated position-aware guidance vector obtained by Eq. (4), l is the video sequence length and $e(\cdot)$ is the score function which estimates the feature significance based on the hidden vector and the position-aware guidance vector. Then, the score function is defined as:

$$e\left(h_j, A_j\right) = v^T tanh\left(W_H h_j + W_A A_j + b\right) \quad (7)$$

In Eq. (7), W_H and W_A are matrices, b is the bias vector, $tanh$ is the hyperbolic tangent function, v is the global vector and v^T is its transpose. By using the obtained attentive weights, the resultant aggregated feature vector is represented by the weighted sum of all the hidden vectors:

$$F_a = \sum_{j=1}^{l} \alpha_j h_j \quad (8)$$

Thus, the aggregated all the pooled feature vectors belonging to one clip is achieved to get the clip descriptors. Then, these clip descriptors obtained from different convolutional layers are fused using the bilinear production for improving its representation ability [6]. By aggregating the clip descriptors, the final video descriptor is generated and the entire network is trained end-to-end with softmax loss supervised by the class label. Once the video descriptor is obtained, these are fed to the SVM for recognizing the human activities in a specific video sequence.

Algorithm:

Input: Video sequences from Penn Action Dataset

Output: Extracted body points, trajectory points (Video descriptor)

Begin
Split video sequences into clips;
for$(each\ clip)$
Initialize CNN parameters for both attention and feature streams;
Compute all the activations in convolutional layers;
Aggregate activations of each convolutional layers using PABRNN;
//PABRNN
Formulate the position-aware guidance propagation via Gaussian kernel;
Calculate the guidance base matrix related to a certain distance;
Aggregate the guidance of all features in convolutional layers;
Obtain the aggregated guidance vector;
Determine the score function and the attentive weight of features;
Find the resultant aggregated feature vector (clip descriptors) belonging to one clip;
Combine attention and feature streams using bilinear product function;
Apply fully connected and softmax layer;
Train the C3D using aggregated guidance feature vector;
Predict the video descriptors for a video sequence;
Perform SVM classifier;
Recognize the individual activities in a particular video sequence;
End

4. Experimental Results

In this section, the JTDPABRD method is implemented in MATLAB 2017b as well as its efficiency is evaluated with the JTDBRD, JTDRD and JTDD based on the RA. In this experiment, the Penn Action dataset is taken into consideration which includes 2326 video sequences of 15 activity classes. The videos are captured from various online video repositories. The length of each video is ranging between 50-100 frames. For every frame, 13 body joints are annotated.

To validate the efficiency, 80% of the data is taken from the entire dataset for training and 20% of data is taken for testing. The body joint coordinates, trajectory points and C3D features are used as baselines. As a result, JTDPABRD with these features is evaluated with various pooling i.e., feature aggregation configurations.

The RA is the percentage of True Positive (TP) and True Negative (TN) rates among the overall amount of trails performed.

$$RA = \frac{TP+TN}{TP+FP+FN+TN} \quad (9)$$

In Eq. (9), FP and FN stand for the false positive and false negative. TP is the amount of correctly recognized legal activities and they are legal. TN is the amount of correctly recognized illegal activities and they are illegal. FP is the amount of wrongly recognized legal activities but they are illegal. FN is the amount of incorrectly recognized illegal activities but they are legal. The results of body joints and trajectory points extraction are portrayed in Figure 4.

Figure 4(a). Sample input video sequence.

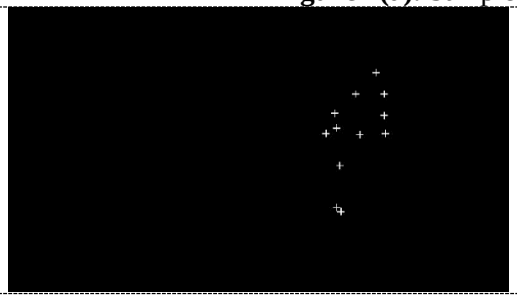

Figure 4(b). Results for body joints extraction of input video sequence.

Figure 4(c). Results for trajectory points extraction of input video sequence.

The RA results on the Penn Action dataset are provided in Table 1.

Table 1. RA of baselines and JTDPABRD with various configurations on Penn action dataset.

	Concatenate all the activations	JTDPABRD Ratio Scaling (1×1×1)	JTDPABRD Coordinate Mapping (1×1×1)	JTDPABRD Ratio Scaling (3×3×3)	JTDPABRD Coordinate Mapping (3×3×3)
Joint coordinates+ trajectory coordinates	0.6452	-	-	-	-

$fc7$	0.7638	-	-	-	-
$fc6$	0.7811	-	-	-	-
$conv5b$	0.7345	0.8358	0.8829	0.8385	0.8683
$conv5a$	0.6675	0.7768	0.8047	0.7722	0.7831
$conv4b$	0.5684	0.7965	0.7873	0.8135	0.8258
$conv3b$	0.4602	0.7268	0.7059	0.7336	0.7315

In Table 1, the 1st column is the RAs of directly utilizing body joint coordinates with trajectory point coordinates as C3D features. The other columns are the RAs achieved by the aggregation of all the features in the particular layer. This scrutiny observes that the RA of $fc7$ is marginally lower to that of $fc6$ since the actual C3D on the Penn Action dataset is not able to fine-tune the $fc7$ layer which is highly preferable to construct the video descriptor for the pre-learned dataset. Also, it observes the results of PABRNN-based feature aggregation at various 3D $conv$ layers. To end, it concludes the JTDPABRD has better efficiency as compared to the JTDBRD, JTDRD and JTDD for aggregating the guided feature vectors of body joint and trajectory points in the video sequence.

The results of various combinations of layers using the scores of SVM with late fusion on the Penn Action dataset are given in Table 2.

Table 2. RA of fusing JTDPABRD from multiple layers together on Penn action dataset.

Fusion layers	JTDD	JTDRD	JTDBRD	JTDPABRD
			RA	
$conv5b + fc6$	0.867	0.871	0.875	0.883
$conv5b + conv4b$	0.987	0.989	0.991	0.994
$conv5b + conv3b$	0.873	0.875	0.879	0.883

Figure 5 indicates that the fusion of JTDPABRD of various layers particularly improves the feature extraction and recognition results. The mixture of JTDPABRD from $conv5b + conv4b$ can maximize the accuracy of recognizing individual activities efficiently. This is because aggregating more significant features in the $conv$ layers.

The results of the impact of estimated body joints + trajectory points versus Ground-Truth (GT) body joints + trajectory points for different HAR methods on the Penn Action dataset is given in Table 3.

Table 3. Impact of estimated body joints + trajectories versus GT body joints + trajectories for different methods on Penn action dataset.

Methods	GT	Estimated	Difference
JTDD ($conv5b$)	0.835	0.810	0.025
JTDRD ($conv5b$)	0.838	0.815	0.023
JTDBRD ($conv5b$)	0.843	0.821	0.022
JTDPABRD ($conv5b$)	0.847	0.828	0.019

From Figure 6, it is observed that the JTDPABRD gives more efficiency than compared with the other methods on Penn Action Dataset. The JTDPABRD attains the maximum efficiency not only with GT body joints and trajectory points, however also with the estimated body joints and trajectory points, beyond the other methods.

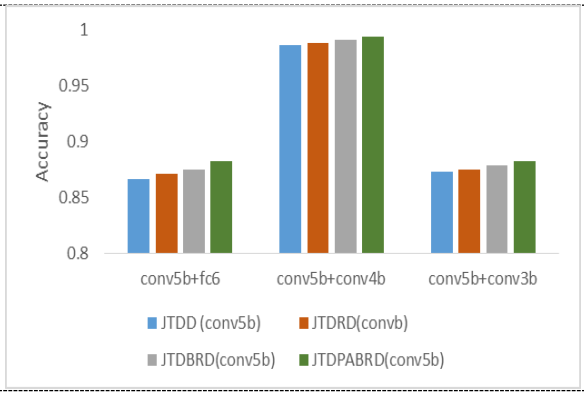

Figure 5. RA of fusing JTDPABRD from multiple layers together on Penn action dataset.

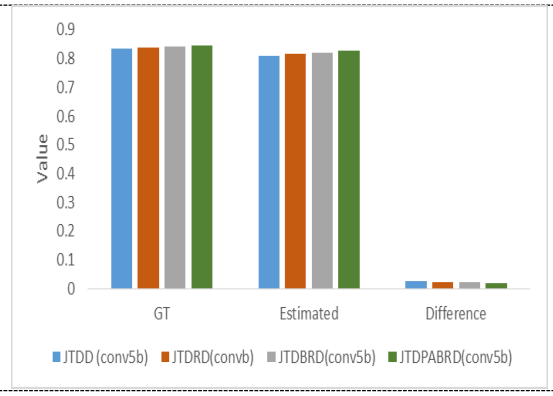

Figure 6. Impact of estimated body joints + trajectories versus GT body joints + trajectories for different methods in Penn action dataset.

5. Conclusion

In this article, JTDPABRD is suggested to combine the PABRNN and two-stream C3D network for extracting the necessary spatiotemporal features and increasing the accuracy of recognizing individual activities. First, the video is divided into several clips and these clips are fed to the two-stream C3D network as input. In a two-stream C3D network, the attention stream is used to extract the guidance of body joints locations and the feature stream is used to extract the trajectory points along with significant spatiotemporal features. After, every convolutional feature vector representation of each clip belonging to the single video is aggregated via the PABRNN to create the clip descriptor. Also, these two streams are multiplied by the bilinear product and end-to-end trained via class labels. Moreover, the activations of fully connected layers and their spatiotemporal variances are also aggregated to generate the final video descriptor. This video descriptor is fed to the SVM to identify the individual activities in videos. To end, the experimental outcomes proved that the RA of JTDPABRD is improved by fusion of $conv5b$ and $conv4b$ with GT feature vectors as compared to the other methods for HAR systems.

References

[1] Wan S, Qi L, Xu X, Tong C and Gu Z 2019 Deep learning models for real-time human activity recognition with smartphones *Mob. Netw. Appl.* 1-13
[2] Ding S, Qu S, Xi Y, Sangaiah A K and Wan S 2019 Image caption generation with high-level image features *Pattern Recognit. Lett.* **123** 89-95
[3] Nweke H F, Teh Y W, Al-Garadi M A and Alo U R 2018 Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges *Expert Syst. Appl.* **105** 233-261
[4] Cao C, Zhang Y, Zhang C and Lu H 2017 Body joint guided 3-D deep convolutional descriptors for action recognition *IEEE Trans. Cybern.* **48** 1095-1108
[5] Ji S, Xu W, Yang M and Yu K 2012 3D convolutional neural networks for human action recognition *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 221-231
[6] Srilakshmi N and Radha N 2019 Body joints and trajectory guided 3D deep convolutional descriptors for human activity identification *Int. J. Innov. Technol. Explor. Eng.* **8** 1016-1021
[7] Rahman S, See J and Ho C C 2017 Exploiting textures for better action recognition in low-quality videos *EURASIP J. Image Video Process.* **2017** 74
[8] Li X, Wang D and Zhang Y 2017 Representation for action recognition using trajectory-based low-level local feature and mid-level motion feature *Appl. Comput. Intell. Soft Comput.* **2017** 1-7

[9] Jin C B, Do T D, Liu M and Kim H 2018 Real-time action recognition using multi-level action descriptor and DNN *Intell. Video Surveill.* IntechOpen.
[10] Shou Z, Lin X, Kalantidis Y, Sevilla-Lara L, Rohrbach M, Chang S F and Yan Z 2019 Dmc-net: generating discriminative motion cues for fast compressed video action recognition *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* 1268-1277
[11] Huo Y, Xu X, Lu Y, Niu Y, Lu Z and Wen J R 2019 Mobile video action recognition *arXiv preprint arXiv:1908.10155*.
[12] Nida N, Yousaf M H, Irtaza A and Velastin S A 2019 Instructor activity recognition through deep spatiotemporal features and feedforward extreme learning machines *Math. Probl. Eng.* **2019** 1-13
[13] Sudhakaran S, Escalera S and Lanz O 2019 LSTA: long short-term attention for egocentric action recognition *Proc. IEEE Conf. Comput. Vis. Pattern Recog.* 9954-9963

Received: January 24, 2022. Revised: March 18, 2022. 406

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Deep Positional Attention-Based Hierarchical Bidirectional RNN with CNN-Based Video Descriptors for Human Action Recognition

Srilakshmi Nagarathinam[1]* **Radha Narayanan[1]**

[1]*Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India*
* Corresponding author's Email: srilakshmiphd123@gmail.com

Abstract: Human Action Recognition (HAR) is a highly notable area of study in contemporary computer vision. Many investigations focused on recognizing a person's actions from video streams based on extracting features regarding orientation and motion. This article presents a Joints and Trajectory-pooled 3D-Deep Positional Attention (PA)-based Hierarchical Bidirectional Recurrent Convolutional Descriptors (JTDPAHBRD) approach which uses a PA-based Hierarchical Bidirectional Recurrent Neural Network (PAHBRNN) for enhancing the feature aggregation process. First, the entire video is segregated into multiple blocks and they are provided to the 2-stream bilinear Convolutional 3D (C3D) model which applies the PAHBRNN as feature aggregation. In PAHBRNN, the feature vectors related to the different parts of a human skeleton in a certain clip are hierarchically aggregated using the position-aware guidance vector. Then, 2 different streams in the C3D network are fused and trained end-to-end using the softmax loss to get the final video descriptor for a particular video sequence. Further, the Support Vector Machine (SVM) classifier is applied to classify the resultant video descriptor to recognize the person's actions. At last, the investigational outcomes demonstrate the JTDPAHBRD achieves 99.6% better recognition accuracy than the classical state-of-the-art approaches.

Keywords: Human action recognition, Deep learning, JTDPABRD, Feature aggregation, Hierarchical BRNN, SVM.

1. Introduction

HAR is a technique used to identify videos that contain a particular task and retrieve relevant videos to distinguish the behavior of a person. It is often used in major domains such as object tracking, the development of human-computer interfaces, and hospital assistance. Thanks to surveillance systems, the internet, Livestream, etc., a vast quantity of videos is recorded every day. In computer vision, HAR is also extremely crucial in modern scenarios [1-3]. Automated detection of specific suspect activities in surveillance systems can also assist in understanding improper or irrelevant actions e.g., automated identification of a loitering person at places like aerodromes, subways, etc.

The motion recognition may allow different functionalities such as the automatic recognition of many gamers' movements. In the healthcare sector, patient rehabilitation can be supported by automated recognition of patients' actions [4-5]. Commonly, HAR is categorized into 3 levels: low, mid, and high. In low-level recognition, identification of edges, extraction of features, and recognition of actions are conducted. In mid-level recognition, identification of human-machine interaction and recognition of abnormal actions are conducted. Similarly, high-level recognitions are useful in different sophisticated applications.

Various findings have been reported in the previous decades to design different forms of HAR systems [6-8]. Alternatively, effective recognition of actions is also quite challenging to different situations, disparities in perception, and so on. In several latest approaches, video is captured under some conditions. However, those ideas have still not been applied in real-world applications.

Besides, a two-stage approach is implemented to learn and identify the features of original video streams by different classification models. The

features that are important in many applications are hardly recognized since feature selection is highly problematic. Specifically, the orientation and trajectories of several scenes in the HAR can be completely distinct [9].

Thus, different deep learning approaches have been applied for training hierarchical characteristics by extracting low- and high-level features [10-13]. Such approaches are guided by either supervised or unsupervised classifiers to ensure an acceptable HAR performance. Unlike other deep learning approaches, Cao et al. [14, 15] designed Joints-pooled 3D-Deep convolutional Descriptors (JDD) to pool the convolutional activations of the 3D-deep Convolutional Neural Network (3DCNN) into the discriminated descriptors depending on the joint coordinates. Originally, a complete video sequence was partitioned into many blocks of fixed dimension and for every block, 3D convolutional attribute maps have been determined. Then, the stable joint coordinates were situated in the 3D attribute maps of a convolutional unit. Also, the activations of every joint coordinate in certain blocks were aggregated and resampled. Further, the mean pooling and l_2-norm were performed to pool these features into the video descriptors which were classified by the linear SVM.

Moreover, this approach was extended by the 2-stream C3D model to simultaneously train the reference from the joints and extract the spatiotemporal characteristics. In C3D, the joint coordinates were extracted using either preprocessing or skeleton extraction [16]. A max-min pooling was performed to pool the body joint-guided feature vector descriptions. Then, the feature and attention streams were multiplied with the bilinear product and given to the Fully Connected (FC) layers to form the resultant video descriptor. But, the time consumption for extracting the joint coordinates was high for complex datasets, and also extracting skeletons was a complicated process.

Thus, Joints and Trajectory-pooled 3D Descriptors (JTDD) have been designed to extract and concatenate the trajectory coordinates or optical flow between any video streams along with the joint coordinates in the C3D approach [17]. Then, the pooled feature descriptors are trained to get the resultant video descriptor which was applied to the SVM for categorizing the human actions. In contrast, the max-min pooling was performed to fuse the features which have more versatility to spatially perfect over the nearby filters. So, the required spatiotemporal disparities among classes were removed.

To tackle this issue, JTDPABRD has been developed which exploits the PA-Bidirectional RNN (PABRNN) model rather than the max-min pooling-based feature aggregation in the two-stream bilinear C3D network [18]. By using PABRNN, the body joint and trajectory point coordinates extracted from two different streams were concatenated to get the final video descriptor for HAR. In contrast, the vanishing gradient problem was occurred due to the use of more parameters. Also, it needs to consider prior input sequences for extracting the long-term spatiotemporal features from long-range video sequences.

Therefore in this paper, a JTDPAHBRD approach is proposed which uses PAHBRNN for enhancing the feature aggregation process. First, the entire video pattern is segregated into multiple blocks and they are provided to the 2-stream C3D model. After extracting the joint and trajectory coordinates at the convolutional layer, the obtained feature vectors are passed to the PAHBRNN to perform feature aggregation rather than max-min pooling. In PAHBRNN, the feature vectors related to the different parts of a human skeleton in a certain clip are hierarchically aggregated using the position-aware guidance vector. Then, 2 different streams in the C3D network are multiplied by the bilinear product and trained end-to-end using the softmax loss to get the final video descriptor for a particular video sequence. Further, the created video descriptor is given to the SVM to recognize the person's actions. Thus, it can extract long-term spatiotemporal features and preserves sequence information over time. Also, it does not tend to vanish with back-propagation through time. As a result, the accuracy of recognizing human actions from video sequences is improved effectively.

The remaining sections of the article are prepared as follows: Section 2 studies the recent HAR systems using deep learning. Section 3 describes the methodology of JTDPAHBRD and Section 4 displays its performance. Section 5 concludes the research work and suggests future enhancements.

2. Literature survey

Spatio-Temporal Distilled Dense-Connectivity Network (STDDCN) [19] was designed to recognize the human actions in the video sequences. In this model, knowledge distillation and dense-connectivity were applied. The main goal of this blockwork was to find interaction policies among shape and movement points with various structures. The spatiotemporal interaction was enabled at the

feature representation layer by using the block-level dense interactions among shape and movement pathways. Further, both points were interacted at the high-level layers based on the knowledge distillation between two streams and their final merging. Also, effective hierarchical spatiotemporal features were obtained. But, its accuracy was not effective for more complex datasets.

Cross-covariance [20] has been introduced to create Symmetric Positive Definite (SPD) matrix-based interpretations for recognizing 3D actions. The cross-covariance was created by the correlation data among the interval-elevated appearances to acquire highly informative attributes and interval-order configuration. Besides, a fashion of expression was devised to combine cross-covariance statics and covariance statistics as a greater SPD matrix obtained from the Riemannian geometry. After that, the symmetric cross-covariance was extended into the non-symmetric version in which the interval-order data was included in the associated matrix forms. However, it needs to extract highly complex non-linear correlations between factors, saliency weighting of appearances, spatiotemporal SPD representation, and so on to increase efficiency.

Timed-image-based CNN [21] has been suggested to recognize the actions in video sequences. Initially, intrinsic 3D attribute training was investigated from Hilbert-based meta-image representation of 3D information. After that, 2D+X representation was developed based on the duality among spatial 2D examples and an extra size X which may associate interval, frequency, or intensity. This description was used to acquire a better balance between spatial data and extra data provided by differences in X. But, it needs to combine interval-image characteristics and their respective spatial vs. temporal features.

A multiple-stream deep learning model [22] has been developed for characterizing global and local movement characteristics in a video sequence. At first, global movements were defined effectively using the intensity-based 3-layer movement record images. Then, the local spatiotemporal samples were mined from the skeleton. After that, the results of such levels were combined and the field information was considered for recognizing human actions. But, its efficiency was analyzed only for fixed background scenarios.

A Deep-Wide network (DWnet) [23] has been developed for human action recognition depending on 3D skeleton information. Initially, attributes were mined using the pruned deep model. After, these were synchronized to a large-dimensional attribute space and identified using the superficial configuration. But, its efficiency was less while dealing with more samples.

Correlation Network (CorNet) with a Shannon fusion [24] has been developed to learn a pre-trained CNN. The CorNet was used to capture a spatiotemporal correlation in a block-by-block manner without time correspondence. Also, Shannon fusion was applied to choose features depending on distribution entropy. The final layers of the pre-trained spatial and temporal networks were correlated for creating a 2D correlation tensor. After, this was fed to the FC layers to train the model. Further, predictions were made by combining the output of CorNet with that from the spatial and temporal stream's outcome. But, its performance was not effective when spatial and temporal streams were not balanced.

Integration of a new representation model [25] has been developed in which Multilayer Deep Features (MDF) of a person area and entire image region were integrated into an Extended Region-aware Multiple Kernel Learning (ER-MKL) scheme. First, the off-the-shelf semantic segmentation was applied to utilize the human cues. After, highly effective representations MDF were built via integrating activations at the final convolutional and FC layers. At last, ER-MKL was applied to train a strong classifier for combining person- and entire image-regions MDF. On the other hand, its classification accuracy was not effective and the computational cost was high.

A novel hybrid deep learning blockwork [26] has been designed to recognize human actions depending on the motion tracking and extraction of spatial features in video sequences. In this blockwork, Gaussian Mixture Model (GMM) and Kalman Filter (KF) were used for identifying and mining the traveled people and Gated Recurrent Neural Networks (GRNN) for gathering the attributes in every block which helps to estimate the people activity. But, its accuracy was not effective for complex datasets and the time of video classification was high.

A Correlational Convolutional LSTM (C2LSTM) network [27] has been suggested which perceives motion data, spatial features, and temporal dependencies to recognize human actions. Initially, convolution and correlation functions were leveraged to credit both the spatial and motion data of the video sequence. Then, a deep network was designed by using the suggested units for recognizing human actions. But, it has less accuracy for more complicated datasets.

3. Proposed methodology

This section describes the JTDPAHBRD approach briefly. First, every video pattern is segregated into several clips and they are fed to the 2-stream C3D model. The 2 streams: attention and feature streams use the convolutional layer for mining the joint and trajectory coordinates, respectively along with the spatiotemporal features of different human skeleton parts in each clip. Then, the activations of every joint and trajectory coordinate with their related spatiotemporal features for specified human parts are aggregated from every channel. For this purpose, PAHBRNN is employed rather than the max-min pooling [8]. In PAHBRNN, the feature vectors related to the human skeleton are considered into 5 different categories as Left Arm (LR) and Left Leg (LL), Trunk (TK), Right Arm (RA) and Right Leg (RL). These are initially extracted from the convolutional layer of the C3D model along with the deep features. After the convolution layer, the extracted features are given to the five different PABRNNs as input. To form the motions from the adjacent skeleton features, the interpretation of the trunk feature is aggregated with those of another 4 types of features, accordingly. After, the occurrence of feature positions in all video clips associated with the particular sequence is obtained. The guidance of the extracted feature vectors to every other position is propagated using the position-aware guidance propagation method, which creates the position-aware guidance vector for each feature vector related to the entire human skeleton. Thus, the position-aware guidance vector gives separate feature vectors for different parts of the human skeleton. Further, these separate feature vectors are combined into their corresponding attention weight to get the resultant aggregated feature vectors. Thus, the CNN with PAHBRNN can extract and reduce the feature dimensionality efficiently.

Moreover, a two-stream bilinear C3D network is trained with the help of a position-aware guidance-based feature vector from the entire human skeleton automatically. The two streams are multiplied by the bilinear product. The entire network is trained end-to-end with softmax loss supervised by class labels. As a result, the feature descriptor for a particular video sequence is obtained and classified by the SVM to identify the person's actions. The schematic representation of JTDPAHBRD-based HAR and the 2-stream C3D using PAHBRNN is illustrated in Fig. 1 and 2, respectively.

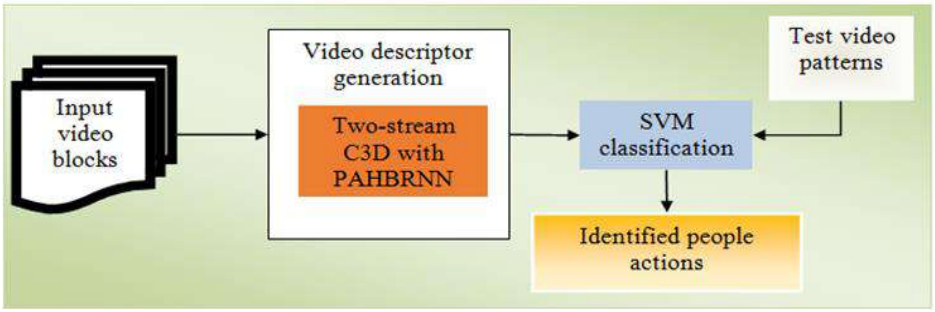

Figure. 1 Schematic representation of JTDPAHBRD-based HAR

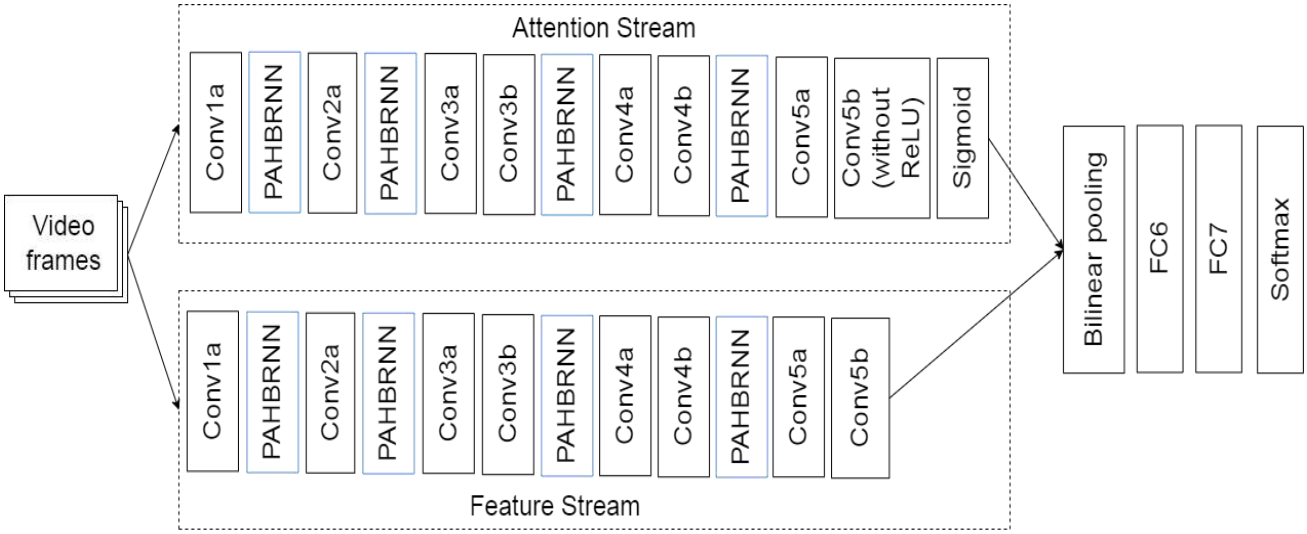

Figure. 2 Architecture of 2-stream bilinear C3D with PAHBRNN-based feature aggregation approach

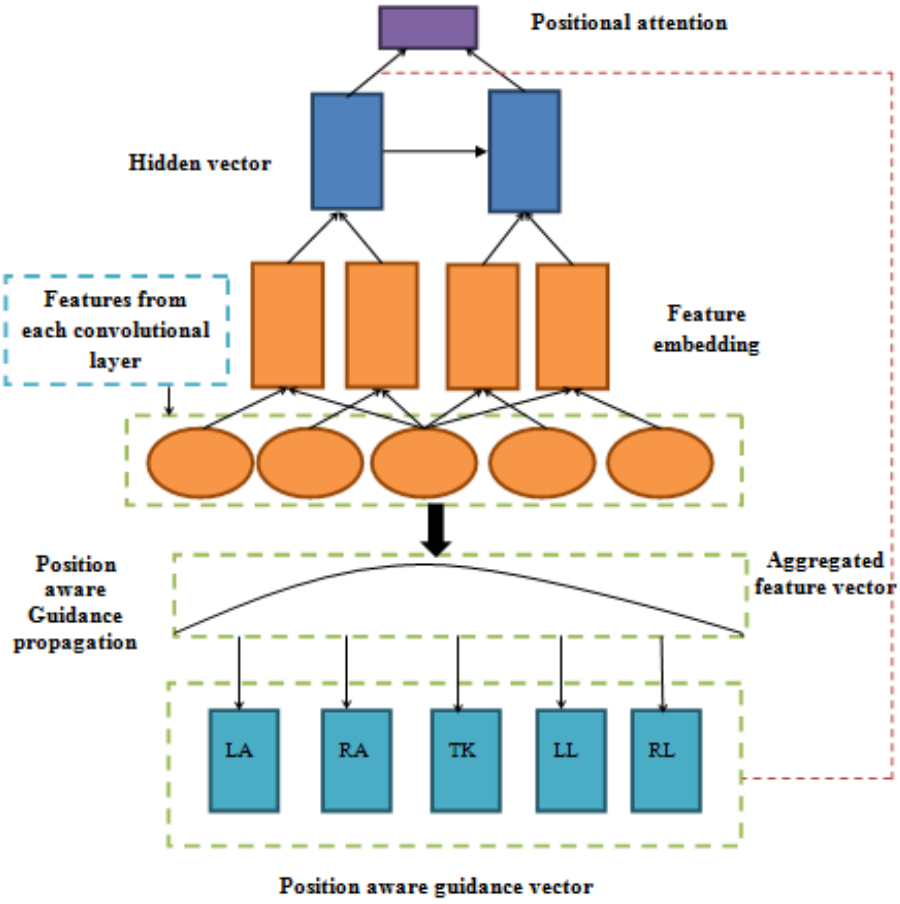

Figure. 3 Aggregated feature vector representation for entire human skeleton using PAHBRNN model

3.1 Positional attention-based hierarchical BRNN

Simple human actions are executed only by a particular segment of them e.g., hitting and kicking forwards are based on tilting the arms and legs, accordingly. Few activities are carried out by shifting the top or bottom body e.g., bowing down is primarily concerned with the top body. Also, highly complicated activities are created by the movements of such 5 segments e.g., jogging and sailing have to cooperate on the movement of the entire body.

To identify different person activities effectively, modeling the motions of such person's segments and their combinations is highly required. For this reason, the PAHBRNN is introduced to mine the long-term contextual data of spatiotemporal patterns. Fig. 3 shows the aggregated feature vector representation of an entire human skeleton using PAHBRNN.

The PAHBRNN adopts the HBRNN with positional attention method to represent the feature vectors by considering the different feature (body joints and trajectory points extracted from different parts of the human skeleton such as LA, RA, TK, LL and RL) embeddings as the input. Consider that the features contain position-aware guidance which is propagated to direct consecutive video clips based on the Gaussian kernel as:

$$Kernel(d) = e^{\left(-d^2/2\sigma^2\right)} \quad (1)$$

Where, d refers to the gap between the actual and aggregated feature vectors and σ represents the variable that limits the propagation scope. Then, the guidance base matrix G associated with the certain distance d and position i is defined as:

$$G(i,d) \sim N(Kernel(d), \sigma') \quad (2)$$

Where N defines the mean density with an estimated $Kernel(d)$ and standard deviation σ'. Moreover, the guidance vector for a feature at a particular position (i.e., LA, RA, TK, LL and RL) is obtained by aggregating the guidance of every feature extracted from the video clips:

$$A_j = Gc_j \quad (3)$$

Where

$$c_j(d) = \sum_{f \in F}[(j-d) \in pos(f)] + [(j+d) \in pos(f)] \quad (4)$$

In Eqs. (3) and (4), A_j is the aggregated guidance vector for the feature at the position j and c_j is the distance count vector which estimates the count of features with different distances. Also, F is the 3D feature maps containing multiple features, f is either a body joint location or a trajectory point feature in F, $pos(f)$ is the group of f's occurrence positions in different clips and $[\cdot]$ is an indicator function which equals to 1 if the criteria satisfy; or else, equals to 0.

Moreover, the position-aware guidance vector for a certain feature is combined into the aggregated feature's attentive weight (α_j) at the position j as:

$$F_a = \sum_{j=1}^{l} \alpha_j h_j \quad (5)$$

Where

$$\alpha_j = \frac{e^{\left(e\left(h_j, A_j\right)\right)}}{\sum_{k=1}^{l} e^{\left(e\left(h_k, A_k\right)\right)}} \quad (6)$$

$$e\left(h_j, A_j\right) = v^T tanh\left(W_H h_j + W_A A_j + b\right) \quad (7)$$

In Eqs. (5) and (6), F_a is the final aggregated feature vector for an entire human skeleton in a certain clip, h_j is the hidden vector at position j, A_j is the aggregated position-aware guidance vector obtained by Eq. (3), l is the video sequence length. In Eq. (7), $e(\cdot)$ is the score function which estimates the feature significance based on the hidden vector and the position-aware guidance vector, W_H and W_A are matrices, b is the bias vector, $tanh$ is the hyperbolic tangent function, v is the global vector and v^T is its transpose.

Thus, this PAHBRNN can generate the feature vectors based on the different parts of the entire human skeleton using five different PABRNNs efficiently. Moreover, the 2 streams in C3D are multiplied using the bilinear product and the aggregated feature vectors for all blocks are concatenated to get the final video descriptor [6] by training the C3D network end-to-end using the softmax loss. After obtaining the video descriptors, SVM is applied to learn these video descriptors and recognize human actions.

Algorithm:
Input: Training video patterns
Output: Human actions

Begin
Split video sequences into blocks;
for$(each\ frame)$
Set CNN variables for attention and feature streams;
Extract the features from different parts of the entire human skeleton such as LA, RA, TK, LL and RL at convolutional layers;
Concatenate the features extracted from each convolutional layer using PAHBRNN;
//PAHBRNN
Create the position-aware guidance propagation through Gaussian filter using Eq. (1);
Compute $G(i,d)$ by Eq. (2);
Aggregate the guidance of (i) RA and LA, (ii) RL and LL with TK features;
Aggregate the guidance of the upper and lower body to get the resultant aggregated position-aware guidance vector using Eqns. (3) (4);
Get the final combined feature vector belonging to a human skeleton in a single clip by calculating α_j and $e\left(h_j, A_j\right)$ using Eqns. (5), (6) & (7);
Fuse attention and feature streams in C3D network with the aid of bilinear product;
Train the two-stream C3D network end-to-end using softmax loss for a whole video sequence;
Obtain the final video descriptors;
Apply the SVM classification;
Identify the human actions from a specified video;
end for
End

4. Experimental results

This part analyzes the efficiency of the JTDPAHBRD approach on the Penn Action dataset by implementing it in MATLAB 2017b. This dataset comprises 2326 video sequences of 15 action labels. Each video is collected from different online video repositories and has 50-100 blocks including 13 body joints are annotated for every block. In this experiment, 1861 video sequences are used for training, and the remaining 465 video sequences are used for testing. The joint and trajectory coordinates, as well as C3D features, are considered as sources. So, the recognition accuracy of JTDPAHBRD with these features is analyzed by using different aggregation configurations.

The ratio of human activities which are correctly identified is called accuracy.

$$Accuracy = \frac{No.of\ recognized\ actions}{Total\ no.of\ actions\ tested} \times 100\% \quad (8)$$

(a) (b)

Figure. 4 (a) Sample input video block and (b) Outcomes of joint and trajectory coordinate extraction

Table 1. Recognition accuracy (%) of sources and JTDPAHBRD with different settings on Penn action dataset

	Aggregate all the activations	**JTDPAHBRD Ratio Scaling (1×1×1)**	**JTDPAHBRD Coordinate Mapping (1×1×1)**	**JTDPAHBRD Ratio Scaling (3×3×3)**	**JTDPAHBRD Coordinate Mapping (3×3×3)**
Joint + trajectory coordinates	0.6621	-	-	-	-
$FC7$	0.7758	-	-	-	-
$FC6$	0.7983	-	-	-	-
$conv5b$	0.7605	0.8533	0.9064	0.8542	0.8885
$conv5a$	0.6834	0.7956	0.8257	0.7961	0.8032
$conv4b$	0.5817	0.8134	0.8015	0.8385	0.8471
$conv3b$	0.4826	0.7517	0.7293	0.7554	0.7566

Table 2. Recognition accuracy (%) of aggregating JTDPAHBRDs from different units on penn action dataset

Aggregation Layers	**JDD [14]**	**STDDCN [19]**	**Dwnet [23]**	**CorNet [24]**	**JTDD [17]**	**JTDPABRD [18]**	**JTDPAHBRD**
$conv5b + FC6$	85.5	85.8	86.1	86.3	86.7	88.3	88.9
$conv5b + conv4b$	98.1	98.2	98.4	98.5	98.7	99.4	99.6
$conv5b + conv3b$	86.0	86.2	86.5	86.8	87.3	88.3	88.6

The experimental outcomes of extracting the joints and trajectory coordinates are illustrated in Fig. 4.

The recognition accuracy results of JTDPAHBRD on the Penn Action dataset are given in Table 1.

In Table 1, the 1st column indicates the accuracies of recognizing joint and trajectory coordinates including C3D features. It exhibits the accuracy of recognizing joint and trajectory coordinates directly as a feature is not sufficiently high. So, all the features in a specific layer should aggregate to achieve higher efficiency. The accuracy of $FC7$ is slightly less than the $FC6$. It is achievable since the real C3D cannot adjust $FC7$ that is well suitable to produce an effective video descriptor. For this reason, the results on PAHBRNN-based pooling at different 3D $conv$ units in JTDPAHBRD using more joints and trajectory coordinates are analyzed. The JTDPAHBRD attains greater performance than the JTDPABRD, JTDD, and JDD to concatenate the guided feature vectors of joint and trajectory coordinates in a video pattern based on 5 different segments.

Also, JTDPAHBRDs from different $conv$ units are concatenated to know if they can balance every other. Table 2 presents the outcomes of various mixtures using late fusion with the SVM scores on the Penn Action dataset. It compares the accuracy of JTDPAHBRD approach with the existing approaches such as JDD [14], STDDCN [19], DWnet [23], CorNet [24], JTDD [17] and JTDPABRD [18].

Fig. 5 demonstrates that the accuracy of fusing JTDPAHBRD from $conv5b + conv4b$ is higher than other combinations and it indicates that the features are interrelated. Thus, the accuracy of the JTDPAHBRD approach for identifying the human actions from the video sequences is effectively increased than all other existing approaches.

Similarly, Table 3 presents the outcomes of the effect of extracted joints + trajectory coordinates vs. Ground-Truth (GT) joints + trajectory coordinates for proposed and existing HAR approaches pooled from of $conv5b$ on the Penn Action dataset.

Fig. 6 proves that the JTDPAHBRD attains a very minimum variance between GT joints+ trajectory coordinates and extracted joints + trajectory coordinate. Hence, it achieves a high performance than the JDD, STDDCN, DWnet, CorNet, JTDD, and JTDPABRD approach on the Penn Action dataset.

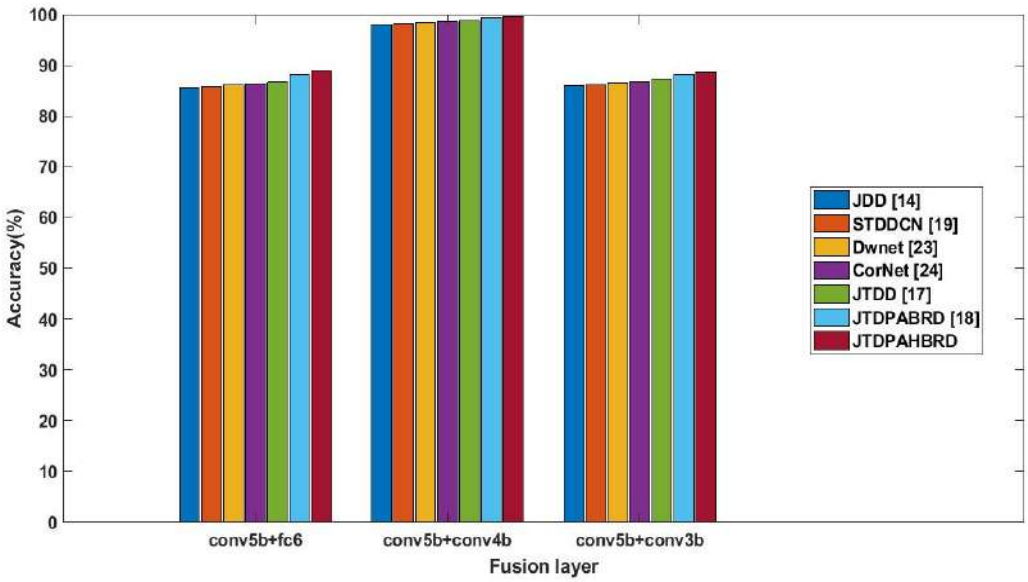

Figure. 5 Accuracy of aggregating JTDPAHBRD from different units on Penn action dataset

Table 3. Effect of extracted joints + trajectories vs. GT joints + trajectories for proposed and existing approaches on Penn action dataset

Approaches Pooled from $conv5b$	**GT**	**Extracted**	**Variance**
JDD [14]	0.819	0.777	0.042
STDDCN [19]	0.823	0.782	0.041
DWnet [23]	0.826	0.786	0.040
CorNet [24]	0.829	0.791	0.038
JTDD [17]	0.835	0.810	0.025
JTDPABRD [18]	0.847	0.828	0.019
JTDPAHBRD	0.860	0.849	0.011

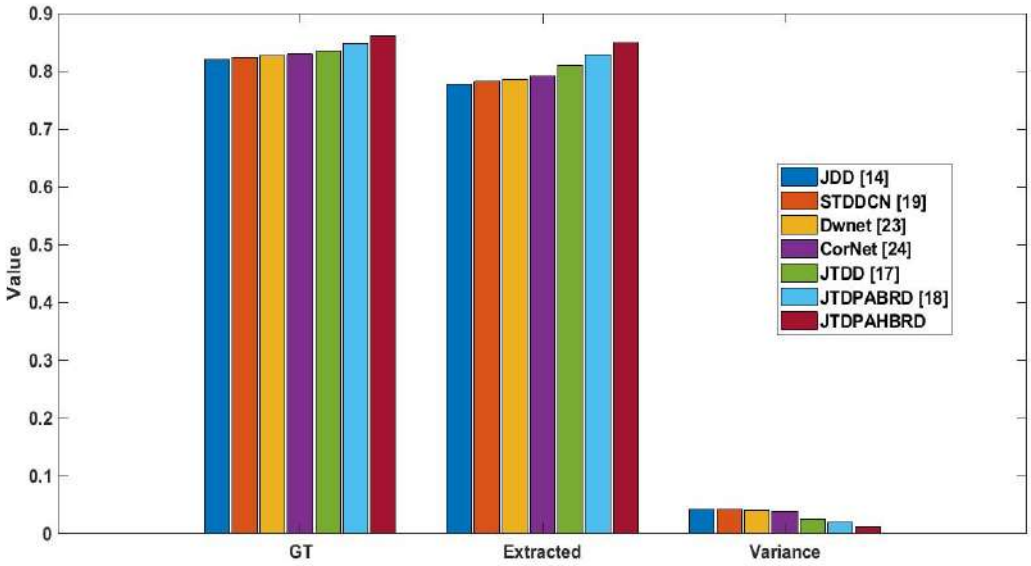

Figure. 6 Influence of identified joints + trajectories vs. GT joints + trajectories for various approaches on Penn action dataset

5. Conclusion

This study proposes the JTDPAHBRD approach in which PAHBRNN is employed to develop the aggregation of features from each video sequence. Initially, each block is fed to the 2-stream C3D model to capture the different features from the different parts of a human skeleton. After that, these characteristics are given to the PAHBRNN which

aggregates them hierarchically into a single feature vector. Also, two streams in the C3D network are combined and trained end-to-end using the softmax loss to acquire the resulting video descriptor. Further, the SVM is trained on the obtained video descriptor and used to recognize the person's actions. To conclude, the investigational outcomes proved that JTDPAHBRD on Penn Action dataset has an accuracy of 99.6% when concatenating it from $conv5b$ and $conv4b$, which is 1.07% greater than all other approaches. While concatenating $conv5b + FC6$ layers, the JTDPAHBRD on Penn Action dataset has an accuracy of 88.9%, which is 2.83% greater than all other approaches. Similarly, concatenating $conv5b + conv3b$ layers, the JTDPAHBRD on Penn Action dataset has an accuracy of 88.6%, which is 2.01% greater than all other approaches.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y. Zhang, W. Qu, and D. Wang, "Action-scene model for human action recognition from videos", *AASRI Procedia*, Vol. 6, pp. 111-117, 2014.

[2] J. Brownlee, "A gentle introduction to a standard human activity recognition problem", *Deep Learning for Time Series*, Vol. 2019, 2019.

[3] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: a survey", *Procedia Computer Science*, Vol. 155, pp. 698-703, 2019.

[4] S. Ranasinghe, F. A. Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation", *International Journal of Distributed Sensor Networks*, Vol. 12, No. 8, pp. 1-22, 2016.

[5] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method", *Journal of Healthcare Engineering*, Vol. 2017, pp. 1-31, 2017.

[6] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods", *Frontiers in Robotics and AI*, Vol. 2, pp. 1-28, 2015.

[7] A. L. H. Ps and U. Eranna, "A simplified machine learning approach for recognizing human activity", *International Journal of Electrical & Computer Engineering*, Vol. 9, No. 5, pp. 3465-3473, 2019.

[8] K. P. Reddy, G. A. Naidu, and B. V. Vardhan, "View-invariant feature representation for action recognition under multiple views", *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 6, pp. 1-13, 2019, doi: 10.22266/ijies2019.1231.01.

[9] J. Basavaiah, C. Patil, and C. Patil, "Robust feature extraction and classification based automated human action recognition system for multiple datasets", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 1, pp. 13-24, 2020, doi: 10.22266/ijies2020.0229.02.

[10] H. Kim, S. Lee, and H. Jung, "Human activity recognition by using convolutional neural network", *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 6, pp. 5270-5276, 2019.

[11] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones", *Mobile Networks and Applications*, pp. 1-13, 2019.

[12] I. R. Moreno, J. M. M. Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition state-of-the-art", *Sensors*, Vol. 19, No. 14, pp. 1-25, 2019.

[13] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features", *Pattern Recognition Letters*, Vol. 123, pp. 89-95, 2019.

[14] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3D deep convolutional descriptors", In: *Proc. of Twenty-Fifth International Joint Conf. on Artificial Intelligence*, pp. 3324-3330, 2016.

[15] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Body joint guided 3-d deep convolutional descriptors for action recognition", *IEEE Transactions on Cybernetics*, Vol. 48, No. 3, pp. 1095-1108, 2017.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221-231, 2012.

[17] N. Srilakshmi and N. Radha, "Body joints and trajectory guided 3D deep convolutional descriptors for human activity identification", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 12, pp. 1016-1021, 2019.

[18] N. Srilakshmi and N. Radha, "Deep positional attention-based bidirectional RNN with 3D

Convolutional video descriptors for human action recognition", *IOP Conf. Series: Materials Science and Engineering*, pp. 1-10, 2021.

[19] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition", *Pattern Recognition*, Vol. 92, pp. 13-24, 2019.

[20] J. Zhang, Z. Feng, Y. Su, and M. Xing, "Cross-covariance matrix: time-shifted correlations for 3D action recognition", *Signal Processing*, Vol. 171, pp. 1-13, 2020.

[21] A. M. Atto, A. Benoit, and P. Lambert, "Timed-image based deep learning for action recognition in video sequences", *Pattern Recognition*, Vol. 104, pp. 1-13, 2020.

[22] Y. Gu, X. Ye, W. Sheng, Y. Ou, and Y. Li, "Multiple stream deep learning model for human action recognition", *Image and Vision Computing*, Vol. 93, pp. 1-10, 2020.

[23] Y. Dang, F. Yang, and J. Yin, "DWnet: deep-wide network for 3D action recognition", *Robotics and Autonomous Systems*, Vol. 126, pp. 1-8, 2020.

[24] N. Yudistira and T. Kurita, "Correlation net: spatiotemporal multimodal deep learning for action recognition", *Signal Processing: Image Communication*, Vol. 82, pp. 1-9, 2020.

[25] B. Sheng, J. Li, F. Xiao, and W. Yang, "Multilayer deep features with multiple kernel learning for action recognition", *Neurocomputing*, Vol. 399, pp. 65-74, 2020.

[26] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 4, pp. 447-453, 2020.

[27] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition", *Neurocomputing*, Vol. 396, pp. 224-229, 2020.

INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features

N Srilakshmi[1]*, **N Radha**[2]

1 Ph.D. Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Tamilnadu, Coimbatore, India
2 Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Tamilnadu, Coimbatore, India

OPEN ACCESS

Received: 20-02-2023
Accepted: 20-06-2023
Published: 03-08-2023

Citation: Srilakshmi N, Radha N (2023) An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features. Indian Journal of Science and Technology 16(29): 2190-2197. https://doi.org/10.17485/IJST/v16i29.379

* **Corresponding author**.

*srilakshmiphd123@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To learn different geometric features of body joints from video frames, as well as trajectory point coordinates, for Human Activity Recognition (HAR). **Methods**: Joints and Trajectory-pooled 3D-Deep Geometric Positional Attention-based Hierarchical Bidirectional Recurrent convolutional Descriptors (JTDGPAHBRD)-based HAR framework is proposed. This framework considers the skeleton graph to extract geometric features such as joints, edges, and surfaces, along with the trajectory point coordinates. A new 3D-deep convolutional network with View Conversion (VC) and Temporal Dropout (TD) layers is designed that uses a Positional Attention-based Hierarchical Bidirectional Recurrent Neural Network (PAHBRNN) to learn more discriminatory high-level features. Then, a Fully Connected Layer (FCL) is applied to get the Video Descriptor (VD) of a particular frame. Moreover, the obtained VD is classified by the Support Vector Machine (SVM) classifier to recognize various kinds of human activities. **Findings**: The test findings show that the JTDGPAHBRD framework using the Penn Action database achieves a recognition rate of 99.7% compared to the existing HAR frameworks. **Novelty**: This framework has significantly improved the recognition of human activities. Thus, it represents a promising framework for the HAR.

Keywords: Human activity recognition; JTDPAHBRD; Geometric features; View conversion; Temporal dropout; SVM

1 Introduction

An efficient HAR can be difficult due to a variety of circumstances, views, and other factors. Over the past years, numerous HAR frameworks have been developed using deep learning algorithms. Deshpnande and Warhade [(1)] presented an improved model for HAR by integrated feature approaches such as Histogram of Gradient (HOG) local feature descriptor and Principal Component Analysis (PCA) as global features, as well as an optimized SVM classifier. But it cannot learn the local relationship among the image pixels and it needs a large number of input parameters. Weiyao et al. [(2)] developed

a multi-modal HAR framework based on Bilinear Pooling and Attention Network (BPAN). The RGB and skeleton information was pre-processed and a multimodal fusion network was devised to obtain fused characteristics. The FC 3-unit perceptron was used to make the final classification decision, but the training database was limited and the total accuracy was influenced by the weight value in the loss function. Muhammad et al.[3] designed a Bidirectional Long Short-Term Memory (BiLSTM)-based attention strategy with a dilated Convolutional Neural Network (CNN) to choose effective features in the input frame and recognize various human actions. Also, the center loss with softmax was used to minimize the loss function in video-based HAR. But it used a single-stream learning strategy, which was not suitable to learn more discriminative features from the video frames and recognize complex actions in large-scale datasets.

Khan et al.[4] developed a deep learning model, which comprises feature mapping, feature fusion, and feature selection. The feature mapping was conducted by DenseNet201 and InceptionV3. Then, deep features were extracted and fused by the serial-based extended model. The best features were chosen by the Kurtosis-controlled weighted K-Nearest Neighbor (KNN). Finally, those features were classified by many supervised learning algorithms. But it has a high computational time during the original deep feature extraction. Wang et al.[5] designed a novel HAR technique called Skeleton Edge Motion Networks (SEMN) to extract gesture data. The SEMN was designed by combining many spatiotemporal segments to obtain a deep interpretation of skeleton structures. A novel advanced rank error was applied to preserve sequential imperative data, but it was difficult to differentiate individual activities from granular skeleton images.

Saleem et al.[6] utilized pre-trained VGG-19 for extracting the body joints from the 2D body skeleton and applied SVM classifier for classifying the human actions. But, its accuracy was less since it did not learn spatiotemporal relationships among different pixels.Yadav et al.[7] designed a Convolutional LSTM (ConvLSTM) network for skeletal-based HAR. Human identification and pose estimation were used to determine skeleton coordinates, which were combined with geometric and kinematic traits to create reference traits. A categorizer head was employed, but it did not consider edges and surface-related geometric traits to enhance HAR efficiency. Putra et al.[8] developed a Deep Neural Network (DNN) using transfer learning and shared-weight schemes for classifying human actions. This model consisted of pre-trained CNNs, attention layers, LSTM with residual learning, and softmax layers. But it did not satisfy outcomes analyzed for online cases, which need to classify sequences of ambiguous actions. Li et al.[9] developed a triboelectric gait sensor system for HAR. They applied LSTM and residual units to extract deep features from multichannel time-series gait data for improving HAR performance. But it needs more geometric features for effective HAR. The above-studied frameworks used only a single-stream learning strategy, whereas a two-stream learning strategy has emerged recently to learn more discriminative features from the video sequences and recognize complex actions accurately. From this perspective, the JTDPAHBRD framework has been developed, which employs PAHBRNN to improve the attribute concatenation task[10]. In this PAHBRNN-based pooling, the attribute vectors associated with the human skeleton in all clips were split into multiple parts according to the body structure. Such parts were fed to the multiple PABRNNs to hierarchically capture and concatenate the long-term spatiotemporal traits. Also, the FCL was utilized to provide the absolute VD that was classified by the SVM for HAR.

On the contrary, these frameworks merely fuse the joint and trajectory coordinates at every interval, while the geometric correlation among joints is ignored in the feature extraction and concatenation process. A typical activity is the formation of a fossil skeleton linked by joints. Therefore, a meaningful description of activities is provided by the relative geometries among joints. The trajectory of a specific joint only conveys gesture data and lacks contour or geometrical data.

Hence, the purpose of this research is to consider the relative geometries in the human body to improve HAR. The JTDGPAHBRD-based HAR framework is proposed, which considers the skeleton graph to extract geometric features such as joints, edges, and surfaces along with the trajectory point coordinates. The joints are separate points of the body. The edges are bones that link 2 nearby joints and are represented via the related joint's locations. The surfaces are the planes made through 2 nearby articulated bones. A new 3D-deep convolutional network with VC and TD layers is designed that uses the PAHBRNN to learn more discriminatory high-level features. Then, the FCL is applied to get the VD of a particular frame. Moreover, the obtained VD is classified by the SVM classifier to recognize various kinds of human activities. Thus, this framework can increase the recognition rate of HAR systems.

2 Methodology

This section briefly explains the JTDGPAHBRD framework for HAR. A general schematic representation of the JTDGPAHBRD framework for HAR is depicted in Figure 1. The major goal is to predict an activity label for an unknown video sequence. Initially, an entire video sequence is split into frames. For each frame, basic geometries like joints, edges, and surfaces are defined with the help of a skeleton graph structure. Also, the trajectory coordinates at each joint location are retrieved. Then, those geometry and trajectory coordinates are passed to the 2-stream C3D network, which comprises PAHBRNN for the pooling process rather than the max-min pooling strategy. Afterward, the features from both streams, such as feature and attention, are concatenated

by the bilinear product, and an absolute VD is obtained by the FCL. The obtained VD is further learned by the SVM classifier for predicting the action labels of test video sequences.

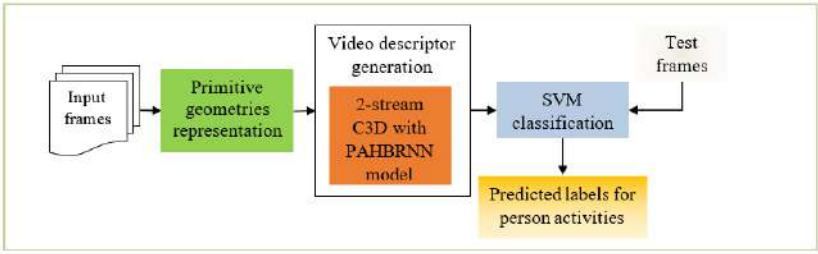

Fig 1. Schematic representation of JTDGPAHBRD-based HAR

2.1 Representation of Primitive Geometries from Skeleton Information

The skeleton information is an arrangement of 3D coordinates of points that create the distorted pattern of the body. Different motions of the points exist while the body changes deliberately. Such points are linked according to the physical pattern of body joints. The body's shape is represented as a graph, where joints are called points and bones are called edges. For a particular person, the skeleton information includes 2 geometric restraints: (1) since a bone's size is fixed, the gap between 2 nearby points along a linked fragment is constant, and (2) three points that create 2 overlapping fragments lie on a similar plane.

According to these interpretations, the skeleton information carries 3 kinds of data: the remote joints, the edges that represent the linked fragments, and the surfaces covered by overlapping fragments. These are explained below.

- Joints

Consider M joints for the body pattern, the coordinates of points at an interval create $M \times 3$ matrix. When the video sequence length is T, the skeleton information is represented by a tensor X with dimension $T \times M \times 3$. The joint coordinates that vary over the period reveal the temporal dynamics of activities. The joint coordinates of a given view are converted into the other view by the rotation matrix. Consider p_k is the joint's coordinate vector at a specific period, the new coordinate vector is attained by

$$\widetilde{p}_k = R p_k \tag{1}$$

In Eq. (1), R denotes the revolution matrix with a size of 3×3. For a given video, consider that R is equal for various joints and various intervals. So, for the joints tensor X, the novel $\widetilde{X}$ detected from the other view is defined by

$$\widetilde{X} = X \times_3 R^T \tag{2}$$

In Eq. (2), $\times_3$ is 3-mode tensor multiplication, $\widetilde{X}$ and X take equal magnitudes.

- Edges

In addition to the temporal features of joints, bone movement forms different activities. A graph is utilized to define the physical links of joints. The joints are represented by the nodes and the bones are represented by the edges.

For a graph of M nodes, there are $M - 1$ edges. The edge indicates the bone orientation. All nodes have a coordinate vector and all edges are denoted by subtracting the vector of the beginning point from that of the endpoint. That is,

$$e_k = p_i - p_j \tag{3}$$

In Eq. (3), e_k, p_i, p_j define the coordinate vectors of the edge, endpoint, and beginning point, correspondingly. The skeleton structure edges are defined by a tensor Y with sizes $T \times (M-1) \times 3$. The node is represented using the edge vector that terminates at that node. For a node that doesn't contain end points of edges, it is represented by a 0 vector. In this manner, the size of Y is incremented by 1 and the sizes of X and Y are created as equal.

The coordinate vectors of edges of a certain view are converted into the other view using a revolution matrix. For the coordinate vector of an edge at a specified interval, the conversion is realized depending on Eqns. (1) and (3):

$$\widetilde{e}_k = \widetilde{p}_i - \widetilde{p}_j = Re_k \tag{4}$$

In Eq. (4), $\widetilde{e}_k$ is the converted vector of e_k. Also, it is defined as:

$$\widetilde{Y} = Y \times_3 R^T \tag{5}$$

In Eq. (5), $\widetilde{Y}$ is the converted tensor of edges. Relating Eq. (2) and Eq. (5), it is observed that the revolution matrices of joints and edges are equal.

- Surfaces

The edges reflect the pairwise correlations among the joints. It is unable to represent the scenario wherein 2 joints with nearby edges are situated next to one another. Similarly, HAR also benefits from the relative motions of nearby bones. Because 2 nearby edges create a plane surface, the standard vector is utilized to represent the surface. Consider e_i, e_j are the vectors denoting 2 nearby edges, the standard vector s_k is:

$$s_k = e_i \times e_j \tag{6}$$

In Eq. (6), $\times$ is the cross product in the 3D area. The vector is not regularized because the magnitude represents the overlapping angle of the respective edges. To maintain the dimension of the standard vector similar to the coordinate vector, it is multiplied by 100. For the body with M joints, there are $(M+2)$ planes. To create a proper relation with joints and edges, 2 surfaces with duplicate data (the standard vector is defined by another standard vector) are omitted. This provides M surfaces so the standard vectors of a sequence are defined using a tensor Z with sizes $T \times M \times 3$.

The standard vector of a particular view is detected from the other views. According to the Eq. (6) and Eq.(4), the novel standard vector of a plane at a certain interval is defined by

$$\widetilde{s}_k = (Re_i) \times (Re_j) = Co(R)\, s_k \tag{7}$$

In Eq. (7), $Co(R)$ denotes the cofactor matrix of R, which is the transpose of the adjoint matrix. For an invertible matrix R, get:

$$Co(R) = (det(R)) \left(R^{-1}\right)^T \tag{8}$$

In Eq. (8), $\left(R^{-1}\right)^T$ denotes the transpose of the inverse R. It must be observed that the determinant of a revolution matrix is one and R^{-1} is its transpose. Eq. (8) is deduced as:

$$Co(R) = \left(R^{-1}\right)^T = R \tag{9}$$

Thus, for the tensor interpretation of planes, it pursues that:

$$\widetilde{Z} = Z \times_3 R^T \tag{10}$$

In Eq. (10), $\widetilde{Z}$ denotes the converted tensor of surface standard vectors. Relating Eq. (2), Eq.(5), and Eq. (10), it is determined that joints, edges, and planes possess an equal revolution matrix.

2.2 Recognition of Human Activities

For HAR, the 3 categories of skeleton information such as the coordinates of joints, edges, and surfaces along with the trajectory coordinates are fed to the 2-stream C3D network. The entire network structure for HAR is depicted in Figure 2. In this structure, the VC layer and TD layer are added to enhance attribute mining and VD generation.

A. View conversion: Human skeletons can be captured from a random camera perception in a real-time circumstance. To create view-invariant interpretations, this framework intends to utilize the VC layer to convert the skeleton information into 3D space by capturing the joints, edges, and planes.

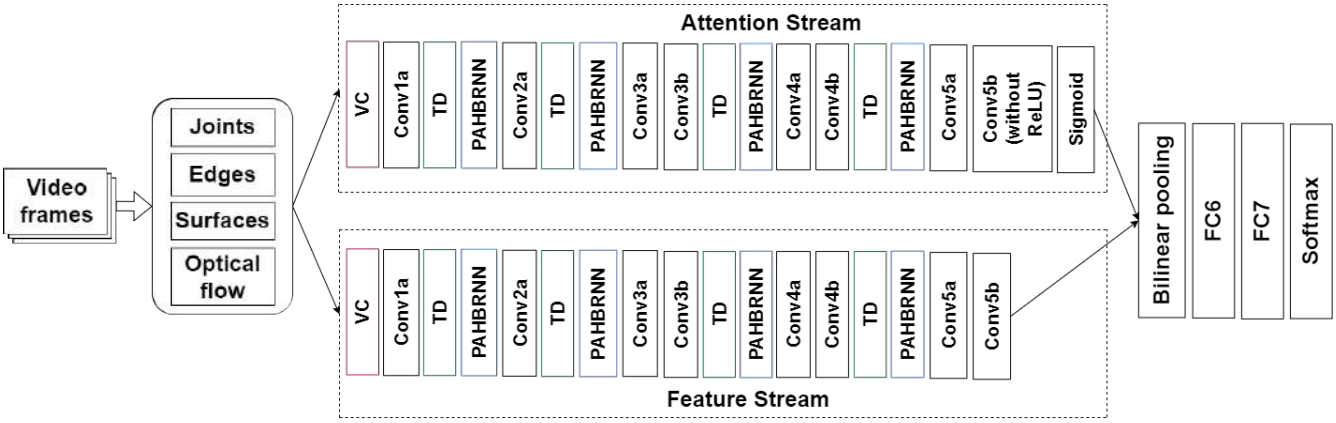

Fig 2. Structure ofproposed 2-stream bilinear C3D network for HAR.

For a given video, X, Y, Z are converted by a similar conversion matrix R. According to Euler's rotation theory, R is denoted as a mixture of revolutions regarding x, y, z axes:

$$R = R_x(\alpha) R_y(\beta) R_z(\gamma) \tag{11}$$

In Eq. (11), α, β, γ denote rotate angles of x, y, z, correspondingly. Given a skeleton structure, R is calculated using 3 separate orientation variables, which are predicted from the skeletons by creating a few significant hypotheses. In the learning task, the orientations in a specific value are arbitrarily chosen and R is computed to convert the inputs. Here, α, β are sampled from $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and γ is set as 0 since the surface is nearly perpendicular to the z-axis. In the test phase, α, β, γ are set as 0, and the actual tensors of joints, edges, and planes are utilized.

B. Temporal dropout: The skeletons gathered might not often be accurate because of noise and pose variations. To solve this issue, a method is adopted depending on dropout, which enhances the framework's robustness. For a typical dropout, all hidden units are arbitrarily neglected from the model with a chance of p_{drop} in the learning. For the test stage, each activation is utilized and $1 - p_{drop}$ is multiplied to consider the rise in the estimated bias. TD is marginally varied from the typical dropout. For $T \times d$ matrix interpretation of a frame, where T denotes the frame size and d denotes the feature size, merely T dropout tests are executed, and the dropout range is extended among the feature size. This method is motivated by the spatial dropout to analyze the convolution feature 4D tensor. In this study, it is altered for 3D tensor and applied for attribute training from frames. As illustrated in Figure 2, the TD is conducted before the PAHBRNN.

Thus, the 2-stream C3D network is trained to create the absolute VD of a given sequence. The obtained VD is provided to the SVM algorithm to categorize the activities of subjects in specified videos.

3 Results and Discussion

The effectiveness of the JTDGPAHBRD framework is assessed by executing it in MATLAB 2017b. The Penn Action Corpus is used in this scrutiny, which comprises 2326 video sequences that each include 15 activity tags. All clips are assembled from several web video libraries and involve 50–100 blocks, each of which has 13 body joints annotated. With this dataset, 1861 video sequences are utilized for learning, while 465 video sequences are utilized for testing. Sources include C3D features, coordinates of primitive geometries, and trajectory coordinates. To assess the recognition accuracy of JTDGPAHBRD using these characteristics, several aggregation setups are used.

The proportion of a person's actions that are correctly recognized is referred to as recognition accuracy. It is computed by using Eq. (9).

$$Accuracy = \frac{Number\ of\ recognized\ actions}{Total\ number\ of\ actions\ tested} \times 100\%$$

The sample input video frame and its corresponding skeleton image for representing the coordinates of primitive geometries are displayed in Figure 3.

Table 1 displays the JTDGPAHBRD recognition rate values for the Penn Action dataset.

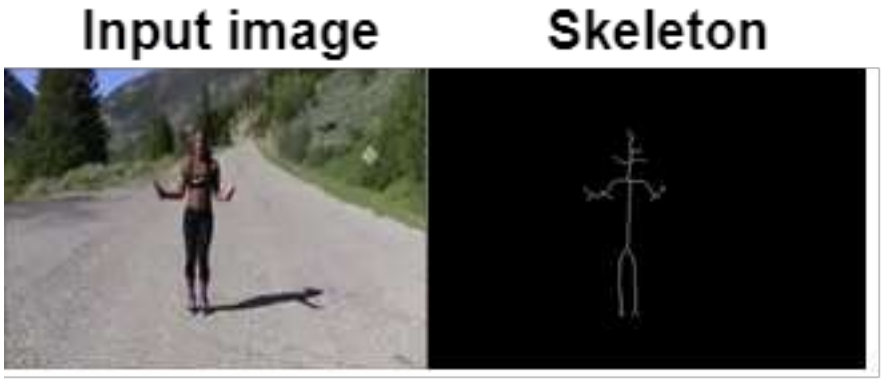

Fig 3. Inputimage and its corresponding skeleton image for primitive geometry coordinates representation.

Table 1. Recognition Rate (%) of Sources and JTDGPAHBRD with Distinct Settingson Penn Action Database

	Cumulative all the activations	**JTDGPAHBRD Ratio Scaling (1×1×1)**	**JTDGPAHBRD Coordinate Mapping (1×1×1)**	**JTDGPAHBRD Ratio Scaling (3×3×3)**	**JTDGPAHBRD Coordinate Mapping (3×3×3)**
Primitive geometries + trajectory coordinates	0.7018	-	-	-	-
$fc7$	0.8045	-	-	-	-
$fc6$	0.8298	-	-	-	-
*conv*5*b*	0.7931	0.8763	0.9231	0.8718	0.9042
*conv*5*a*	0.7084	0.8251	0.8518	0.8146	0.8275
*conv*4*b*	0.6102	0.8406	0.8227	0.8555	0.8633
*conv*3*b*	0.5095	0.7794	0.7504	0.7791	0.7727

The accuracy of identifying coordinates for primitive geometries and trajectories is shown in Table 1's first column. It demonstrates that the direct recognition of primitive geometries and trajectories as a trait is insufficiently accurate. Therefore, to increase accuracy, each trait in a given layer must be concatenated. $fc7$'s accuracy is somewhat worse than $fc6$'s accuracy. It is possible because the genuine C3D can't change $fc7$, which is necessary to create a meaningful VD. Since additional primitive geometries and trajectory coordinates were used, the outcomes of PAHBRNN-based pooling at various 3D *conv* units in JTDGPAHBRD are examined. It is clear that when concatenating the primitive geometries and trajectory coordinates in video patterns according to separate parts of the human body (e.g., right leg, right arm, trunk, left leg, and left arm), the JTDGPAHBRD performs better than the other HAR systems.

Additionally, JTDGPAHBRDs from several *conv* units are combined to determine whether they can balance one another. The outcomes of various configurations applying late merging and the SVM grades on the Penn Action database are shown in Figure 4. It compares the accuracy of JTDPAHBRD framework with the existing frameworks: BPAN [2], SEMN [5], VGG19-SVM [6], ConvLSTM [7], JTDD [11], JTDPABRD [12], and JTDPAHBRD [10].

Figure 4 displays that concatenating *conv*5*b* + *conv*4*b* in the JTDGPAHBRD has a greater recognition rate than other groupings and that the traits are interconnected. Thus, it is concluded that the JTDGPAHBRD framework can accurately recognize human activities in different video sequences compared to the other existing frameworks. For various HAR frameworks on the Penn Action dataset, Table 2 provides the performance outcomes of the obtained coordinates of the primitive geometries and trajectories vs. Ground-Truth (GT) geometries + trajectory coordinates. From these analyses, it is clear that the proposed JTDGPAHBRD framework outperformed the other HAR frameworks applied to the Penn Action database.

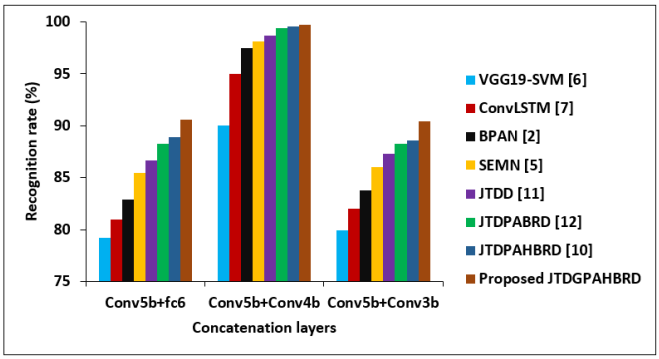

Fig 4. Recognitionrate of JTDGPAHBRD by concatenating different layers for penn action dataset

Table 2. Effect of Obtained Primitive Geometries + Trajectories vs. GT Geometries + Trajectoriesfrom *conv5b* for Various HAR Frameworks on Penn Action Database

Frameworks	GT	Obtained	Variation
VGG19-SVM [6]	0.733	0.671	0.062
ConvLSTM [7]	0.751	0.694	0.057
BPAN [2]	0.784	0.735	0.049
SEMN [5]	0.819	0.777	0.042
JTDD [11]	0.835	0.810	0.025
JTDPABRD [12]	0.847	0.828	0.019
JTDPAHBRD [10]	0.860	0.849	0.011
JTDGPAHBRD	0.893	0.886	0.007

4 Conclusion

The JTDGPAHBRD framework was developed in this study that learned both the coordinates of primitive geometries of body joints and trajectories from multiple video frames for HAR. This framework achieved promising results with a recognition rate of 99.7%. This framework could also be used in video surveillance systems, sports, defense, etc., for recognizing a person's actions precisely. The extraction of different geometries among body joints along with the trajectories enhanced the performance of HAR when using large-scale video sequences. Though it recognizes different human actions, it did not efficiently learn the spatiotemporal relationships among various geometries and a manual extraction of geometries from long-range video sequences was complex. So, future work will focus on introducing a graph-based neural network for automatically learning spatiotemporal relationships among geometric features to create more robust video descriptors.

References

1) Deshpnande A, Warhade KK. An Improved Model for Human Activity Recognition by Integrated feature Approach and Optimized SVM. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE. 2021;p. 571–576. Available from: https://doi.10.1109/ESCI50559.2021.9396914.
2) Weiyao X, Muqing W, Min Z, Ting XZ. Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition. *IEEE Sensors Journal*. 2021;21(17):19157–19164. Available from: https://doi.10.1109/JSEN.2021.3089705.
3) Muhammad K, Mustaqeem, Ullah A, Imran AS, Sajjad M, Kiran MS, et al. Human action recognition using attention based LSTM network with dilated CNN features. *Future Generation Computer Systems*. 2021;125:820–830. Available from: https://doi.org/10.1016/j.future.2021.06.045.
4) Khan S, Khan MA, Alhaisoni M, Tariq U, Yong HSS, Armghan A, et al. Human Action Recognition: A Paradigm of Best Deep Learning Features Selection and Serial Based Extended Fusion. *Sensors*. 2021;21(23):7941–7941. Available from: https://doi.org/10.3390/s21237941.
5) Wang H, Yu B, Xia K, Li J, Zuo X. Skeleton edge motion networks for human action recognition. *Neurocomputing*. 2021;423:1–12. Available from: https://doi.org/10.1016/j.neucom.2020.10.037.
6) Saleem R, Ahmad T, Aslam M, Martinez-Enriquez AM. An Intelligent Human Activity Recognizer for Visually Impaired People Using VGG-SVM Model. In: Advances in Computational Intelligence;vol. 13613. Springer Nature Switzerland. 2022;p. 356–368. Available from: https://doi.org/10.1007/978-3-031-19496-2_28.
7) Yadav SK, Tiwari K, Pandey HM, Akbar SA. Skeleton-based human activity recognition using ConvLSTM and guided feature learning. *Soft Computing*. 2022;26(2):877–890. Available from: https://doi.10.1007/s00500-021-06238-7.

8) Putra PU, Shima K, Shimatani K. A deep neural network model for multi-view human activity recognition. *PLOS ONE*. 2022;17(1):e0262181. Available from: https://doi.org/10.1371/journal.pone.0262181.
9) Li J, Xie Z, Wang Z, Lin Z, Lu C, Zhao Z, et al. A triboelectric gait sensor system for human activity recognition and user identification. *Nano Energy*. 2023;112. Available from: https://doi.org/10.1016/j.nanoen.2023.108473.
10) Nagarathinam S, Narayanan R. Deep Positional Attention-Based Hierarchical Bidirectional RNN with CNN-Based Video Descriptors for Human Action Recognition. *International Journal of Intelligent Engineering & Systems*. 2022;15(3):406–415. Available from: https://doi:10.22266/ijies2022.0630.34.
11) Srilakshmi N, Radha N. Body Joints and Trajectory Guided 3D Deep Convolutional Descriptors for Human Activity Identification. *International Journal of Innovative Technology and Exploring Engineering*. 2019;8(12):1016–1021. Available from: https://doi.10.35940/ijitee.K1985.1081219.
12) Srilakshmi N, Radha N. Deep Positional Attention-based Bidirectional RNN with 3D Convolutional Video Descriptors for Human Action Recognition. *IOP Conference Series: Materials Science and Engineering*. 2021;1022(1):1–10. Available from: https://doi.10.1088/1757-899X/1022/1/012017.