IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 2, June 2024, pp. 2102~2110
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i2.pp2102-2110

Activity recognition based on spatio-temporal features with transfer learning

Seemanthini Krishne Gowda[1], Shobha Narasimha Murthy[2], Jayaprada S. Hiremath[3], Sowmya Lakshmi Belur Subramanya[4], Shantala S. Hiremath[5], Mrutyunjaya S. Hiremath[6]
[1]Department of Machine Learning, BMS College of Engineering, Bangalore, India
[2]Department of Computer Science and Design, Dayananda Sagar College of Engineering, Bangalore, India
[3]Department of Computer Science, Visvesvaraya Technological University, Belgaum, India
[4]Department of Machine Learning, BMS College of Engineering, Bangalore, India
[5]Department of Image Processing, Tata Elxsi Limited, Bangalore, India
[6]Department of Video Processing, eMath Technology Pvt. Ltd., Bangalore, India

Article Info

Article history:

Received Jan 6, 2023
Revised Nov 3, 2023
Accepted Dec 2, 2023

Keywords:

Convolutional neural network
Deep learning
Human action recognition
Multiclass classification
ResNet50
Support vector machine
Transfer learning

ABSTRACT

Human action recognition has emerged as a significant area of study due to it is diverse applications. This research investigates convolutional neural network (CNN) structures to extract spatio-temporal attributes from 2D images. By harnessing the power of pre-trained residual network 50 (ResNet50) and visual geometric group 16 (VGG16) networks through transfer learning, intricate human actions can be discerned more effectively. These networks aid in isolating and merging spatio-temporal features, which are then trained using a support vector machine (SVM) classifier. The refined approach yielded an accuracy of 89.71% on the UCF-101 dataset. Utilizing the UCF YouTube action dataset, activities such as basketball playing and cycling were successfully identified using ResNet50 and VGG16 models. Despite variations in frame dimensions, 3DCNN models demonstrated notable proficiency in video classification. The training phase achieved a remarkable 95.6% accuracy rate. Such advancements in leveraging pre-trained neural networks offer promising prospects for enhancing human activity recognition, especially in areas like personal security and senior care.

Corresponding Author:

Mrutyunjaya S. Hiremath
Department of Image Processing, eMath Technology
Bangalore-560072, Karnataka, India
Email: mrutyunjaya.publications@gmail.com

1. INTRODUCTION

Research in computer vision encompasses human activity recognition [1]. Systems designed for such recognition can be pivotal for video cataloging, content-driven retrieval, interactive human-computer interfaces, and surveillance [2], [3]. Analyzing video footage to discern human actions is crucial, but factors such as background distractions, obstructions, variable sizes, differing appearances, and inconsistent lighting can complicate the task. Typically, urban areas have denser populations than rural regions, leading to a higher probability of unauthorized assemblies, altercations, and similar events [4]. The identification of human actions in videos necessitates the extraction of feature vectors. Much of the research in this domain emphasizes specific attributes like local space-time elements [5], spatio-temporal dynamics [6], and motion boundary histograms (MBH) [7]. The accuracy of recognition is often contingent on factors such as lighting [8], [9] and the perspective of the video.

Modern discourse on computer vision has been influenced by advancements in neural networks [10]. This piece delves into the 3D convolutional neural networks (CNN) and their role in medical imaging. The 3DCNN architecture is employed for categorizing videos, which is inherently complex due to the sequential data involved. The method focused on visual surveillance frequently monitors the elderly [11].

3DCNNs are adept at isolating spatio-temporal elements, and the efficacy of recognition is tied to these unique features. While handcrafted attributes are intriguing, they don't always scale well with expansive video datasets [12]. Hence, architectures like 3D residual network (ResNet) [13] and visual geometric group 16 (VGG-16) are employed to derive feature descriptors. It is essential that video data comprehensively represents motion on a frame-by-frame basis. Thus, the 3D convolution in CNNs effectively mirrors spatio-temporal nuances [14]. The conceptualized approach is illustrated in Figure 1.

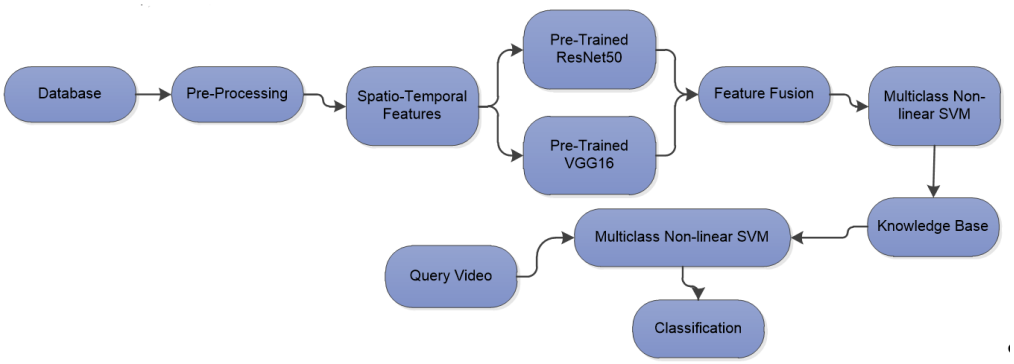

Figure 1. Block diagram of the proposed method

2. LITERATURE SURVEY

Haroon and Eranna [15] contend that existing systems don't prioritize computational efficiency, instead focusing on identifying application-specific anomalies. The emerging technology can pinpoint human movement accurately and affordably, even without leveraging all features. Recognizing human skeletal movements assists in identifying crucial joints. The innovative approach of the proposed system holds promise for enhancing human activity recognition. Systems trained using k-nearest neighbor (KNN) have shown superior performance.

Sakr *et al.* [16] highlight that while primary activity recognition has been explored extensively, intricate activity recognition remains challenging [17]. Recognizing complex activities involves multilabel classification [18], where a single test instance is categorized into multiple concurrent activities. Existing techniques [19] are limited in recognizing just two concurrent activities, underscoring the need for a multilabel activity training dataset. Notably, a limited training dataset has been effective in discerning intricate multilabel activities using emerging patterns and fuzzy sets. Trials that assume linear separability of each activity for individual residents demonstrate the method's robustness and superiority. Our models cater to nonlinear, distinctive, and multi-resident activities.

Haroon and Eranna [20] posit that real-time event streams can present variability, especially when digital image processing is employed for activity identification. Many contemporary techniques fall short when dealing with real-world events, rendering them suboptimal. The proposed model for activity detection harmoniously integrates accuracy and efficiency, making extracting event features based on spatial and temporal cues easier. Machine learning could be harnessed in scenarios with significant subject movements to amplify these features. The recommended approach underscores improvements in both precision and operational performance.

Han *et al.* [21] introduced networks influenced by autoencoders. A principal coefficient encoder model (PCEM) was instrumental. PCEM, relying on training data, determines network weights and employs two hidden layers with core coefficient encoding for profound architecture. Evaluating PCEM's efficacy independent of the subject involves leveraging training and testing datasets from varied participants, with the PCEM successfully identifying 97% of activities.

Vrskova *et al.* [22] championed a 3DCNN for discerning human motions from videos. 3DCNNs surpass other neural network architectures when optimized, especially in tests involving motion, static, and combined attributes. The foremost 3DCNN model boasts an impressive 87.4% recognition accuracy, compared to 65.4%, 63.1%, and 71.2% by other models. Future iterations of this research will harness larger batch sizes and epochs, alerting users upon detecting anomalies in human behavior.

Butt *et al.* [23] illustrated the creation of novel codebooks, focusing on video motion representation by melding hybrid encoding with spatio-temporal CNN architectures. Their method, favoring agglomerative clustering, utilizes global and class-specific codebooks, with performance metrics on HMDB51 and UCF101 standing at 72.6% and 96.2%, respectively. Their efforts seek to refine the bag-of-words codebook and feature encoding, leveraging two-stream 2D and 3D ResNets to extract deep ConvNet features. To augment computational performance, they embed the encoding within a holistic ConvNet, determining fusion weights for each stream.

Kahlouche *et al.* [24] turned to deep learning for classifying seven activities. Their approach harnesses the efficiency of the human activity recognition algorithm, with training on Microsoft Kinect 3D skeletal data focusing on both spatial and temporal attributes. The combined strength of CNN-long short term memory (LSTM) aids in mastering these features [25], [26]. LSTM and CNN architectures enhance outcomes, and data preprocessing ensures view invariance. Upcoming work will enhance human action recognition, factoring in skeletal sequencing and the visual appearance of objects using both RGB and depth modalities.

3. METHODOLOGY

Figure 1 illustrates the four core stages of the proposed method: pre-processing, feature extraction, feature fusion, and training. The pre-processing step encompasses frame extraction, resizing, and contrast enhancement through histogram equalization, as detailed in section 3.1. From these enhanced frames, features are gleaned using the spatio-temporal domain and then channeled through the pre-established ResNet50 and VGG16 networks, as delineated in section 3.2. Subsequently, features derived from the FC1000 layer of these networks are merged (or concatenated) employing an averaging strategy, a process further elaborated upon in section 3.3. These amalgamated features then undergo training via a multiclass nonlinear support vector machine (SVM) classifier, detailed in section 3.4.

3.1. Pre-processing

Preprocessing plays a pivotal role in enhancing the quality of video frames, paving the way for more effective analysis. It mitigates unwanted distortions and amplifies specific features vital for the given application. These essential features can differ based on the application in focus. The proposed methodology incorporates frame extraction, resizing, and contrast augmentation. These stages are further elaborated and their procedural flow is depicted in Figure 2.

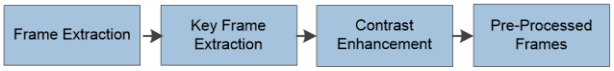

Figure 2. Pre-processing flowchart

3.1.1. Frame extraction

The initial step is “frame extraction” from the video. The primary objective of this extraction process is to isolate essential frames from the video, eliminate redundancy, and streamline processing. Frames are then resized to a resolution of 128×128. Keyframes from these videos have the potential to forecast human behavior. The keyframe process prioritizes the selection of summary frames that best represent the video’s content.

3.1.2. Contrast enhancement using histogram equalization

Contrast enhancement boosts the visibility of objects by amplifying brightness disparities. This enhancement can be applied in either a single step or multiple stages. One popular method in image processing for improving contrast is histogram equalization, prized for it is straightforwardness and efficiency. This method adjusts an image’s dynamic range and contrast to align with it is intensity histogram by expanding prevalent intensity levels and redistributing the values, resulting in a uniform image histogram. The fundamental formula for histogram equalization is depicted in (1), with the protocol detailing each subsequent step.

$$h(v) = round\left(\frac{cdf(v) - cdf_{min}}{(M \times N) - cdf_{min}} \times (L-1)\right) \quad (1)$$

where v is value, cdf_{min} is minimum non-zero value of the cumulative distribution function, $(M \times N)$ is image's number of pixels, and L is number of gray levels used.

3.2. Feature extraction

Figure 3 illustrates the feature extraction process, divided into three phases. The superfluous temporal data from interframes are deducted from the intraframe and then processed through 3D ResNet50 and 3D VGG16 pre-trained networks. Spatio-temporal characteristics are drawn from the fully connected layer of these pre-trained networks. Subsequently, these features are merged via the averaging technique and trained with a multiclass nonlinear SVM.

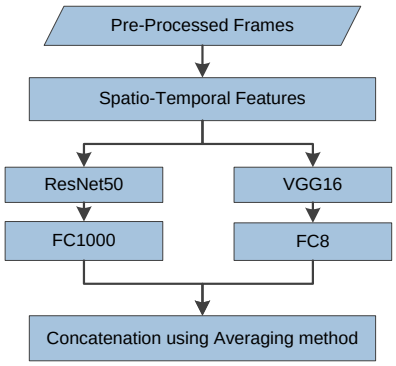

Figure 3. Feature extraction and fusion flowchart

3.2.1. Spatio-temporal features

A set of 15 frames, termed group of pictures (GOP), is considered. Repeated data is eliminated by subtracting the temporal information in interframes from the intraframe. The spatio-temporal characteristics of this refined data are then extracted using the pre-trained 3D ResNet50 and 3D VGG16.

a. ResNet50

The ResNet [27] architecture serves as the basis for feature extraction. Specifically, the pre-trained ResNet50 network, which comprises 49 convolutional layers arranged in three layers across five sections, and one fully connected layer, is utilized for initial feature extraction. The spatio-temporal features derived from the GOP are input into this network, and the features from the FC1000 layer are subsequently harvested. A comprehensive breakdown of the ResNet50 structure can be found in Table 1. The input dimension for the pre-trained network has been adjusted to 112×112.

Table 1. ResNet50 layers detail

Layer name	Output size	Layers	
Conv1	112×112	7×7, 64, stride 2	
Conv2	56×56	3×3 Max pool, stride 2	
		1×1, 64	×3
		3×3, 64	
		1×1, 256	
Conv3	28×28	1×1, 128	×4
		3×3, 128	
		1×1, 512	
Conv4	14×14	1×1, 256	×6
		3×3, 256	
		1×1, 1024	
Conv5	7×7	1×1, 512	×3
		3×3, 512	
		1×1, 2048	
	1×1	Average Pool	
		1,000-d Fully Connected	
		SoftMax	

The mathematical expression for the convolution layer in a pre-trained network is described as in (2):

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau \quad (2)$$

The mathematical expression for the rectified linear unit (ReLu) layer in a pre-trained network is described as in (3):

$$f(x) = max\ (0, x) \tag{3}$$

where x is input to a neuron. The mathematical expression for the SoftMax layer in a pre-trained network is described as in (4):

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} = \ \ for\ i = 1 \dots, K\ and\ z = (z_1, \dots \dots, z_K)\ \epsilon\ R^K \tag{4}$$

where z is input vector and K is real number.

b. VGG16

The VGG16 network, already pre-trained, encompasses 13 convolutional layers divided into five sections, along with three fully connected layers. This network is employed in the secondary phase of feature extraction. spatio-temporal attributes from the GOP are funneled into this network, with the FC8 layer being the source of extracted features. An in-depth depiction of the VGG16 layout is provided in Table 2. The dimensions of the input layer in this pre-trained network are adjusted to 224×224.

Table 2. VGG16 layers detail

Layer name	Output size	Layer size
Conv_1	112×112	224×224×64
		224×224×64
MaxPool		112×112×128
Conv_2	56×56	112×112×128
		112×112×128
MaxPool		56×56×256
Conv_3	28×28	56×56×256
		56×56×256
		56×56×256
MaxPool		28×28×512
Conv_4	14×14	28×28×512
		28×28×512
		28×28×512
MaxPool		14×14×512
Conv_5	7×7	14×14×512
		14×14×512
		14×14×512
MaxPool		7×7×512
Fully_Connected_1 (FC6) 1×1×4096		
Fully_Connected_2 (FC7) 1×1×4096		
Fully_Connected_3 (FC8) 1×1×1000		
SoftMax layer		

3.3. Feature fusion

Features derived from the FC1000 layer of ResNet50 and the FC8 layer of VGG16 are integrated using the averaging technique. The 2048 attributes obtained from the FC1000 layer of ResNet50 undergo downsampling to align with the 1,000 attributes from VGG16's FC8 layer. A feature vector is formed via the averaging process utilizing these attributes. This vector becomes instrumental in the training process, further elucidated in the subsequent section.

3.4. Training-multiclass nonlinear support vector machine

The feature vector resulting from feature fusion is trained using a multiclass nonlinear SVM classifier equipped with radial basis function (RBF) kernel. SVMs, commonly employed for classification and regression [28], utilize hyperplanes in multi-dimensional spaces to separate target categories. The goal is establishing the optimal decision boundary with the maximum margin for classifying new data points. However, when decision boundaries in multi-dimensional space are inadequate for data classification, the RBF kernel comes into play. This kernel functions similarly to the KNN approach. Importantly, during the training phase, the kernel can reduce space complexity by retaining only the support vectors instead of the entire dataset. The RBF kernel is defined by (5):

$$k(\vec{x}_i, \vec{x}_j) = exp(-\gamma \|\vec{x}_i - \vec{x}_j\|^2)\ for\ \gamma > 0 \tag{5}$$

where $\vec{x}_i, \vec{x}_j$ is feature vectors, γ is $\frac{1}{2\sigma^2}$, and σ is free parameter.

3.5. Database

The UCF101 dataset features upcoming YouTube videos spanning 101 action categories, totaling 13,320 individual clips. These action categories can be grouped into five types: human-object interaction, body-motion only, human-human interaction, playing musical instruments, and sports. The UCF101 extends the UCF50 dataset, encompassing 50 categories [29]. For our experiments, we selected the third category. This subset consists of five specific classes: "BandMarching," "Haircut," "HeadMassage," "MilitaryParade," and "SalsaSpin." These classes contain 155, 130, 147, 125, and 133 videos. All videos are presented in a 320×240 resolution at 25 fps. We've divided the database into three segments: training, validation, and testing, with a ratio of 70:10:20.

4. EXPERIMENTAL RESULTS

This section outlines the outcomes post-experimentation, aiming to validate or refute the hypothesis. It includes data representations like tables and graphs, along with an analysis comparing the results to previous studies. While this section presents data-derived insights, a more in-depth analysis.

4.1. Experimental setup

For this model, we utilized a system with a 10th-generation i-7 processor running on a Windows 10 64-bit OS. The system boasts 16 GB RAM, an Nvidia RTX graphics card, and an additional 8 GB dedicated RAM. All experiments were conducted using MATLAB 2019b.

4.2. Performance evaluation parameters

We assessed the model's efficacy using several metrics: specificity, accuracy, precision, recall, F1 score, and the receiver operating characteristic (ROC) curve. In (6) through 10 provide the mathematical formulations for specificity, accuracy, precision, recall, and F1 score, respectively [30]. Figure 4 illustrates the confusion matrix, displaying the outcomes of machine learning classification based on various combinations of predicted and actual values. The results of these performance measurements are summarized in Table 3.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (6)$$

$$Accuracy = \frac{Total\ number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions} \quad (7)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (8)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (9)$$

$$F1\ score = 2 * \left(\frac{precision * recall}{precision + recall}\right) \quad (10)$$

The ROC curve is depicted by plotting the true positive rate (TPR) against the false positive rate (FPR). This curve can be observed in Figure 5. Table 4 contrasts these findings with other methodologies. As the loss function diminishes, there's a noticeable uptick in accuracy during the training phase. A consistent trend of increasing accuracy is evident with each passing epoch. The peak accuracy recorded during training was 95%. Table 5 highlights the competitive advantage of the proposed approach, which surpasses the accuracy of prior state-of-the-art methods. This underscores the potential significance of the method for various applications, including personal security and senior care, where accurate human activity recognition is essential.

Table 3. Performance parameter measures

Parameters	Percentage
Specificity	97.57
Accuracy	89.71
Precision	96.84
Recall	89.03
F1-score	92.77

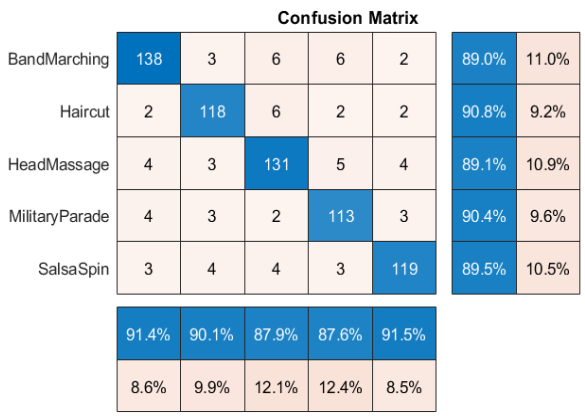

Figure 4. Confusion matrix

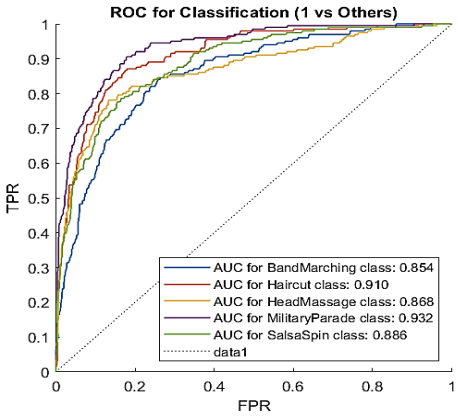

Figure 5. ROC curve

Table 4. Comparison with different methods for the UCF-101 dataset

Methods	Accuracy (%)
SVM	64.73
VGG16	72.98
ResNet50	77.52
Proposed	89.71

Table 5. Comparison with the state-of-the-art

Papers	Accuracy (%)
Carreira *et al.* [31]	75.7
Simonyan and Zisserman [32]	87.0
Proposed	89.71

5. CONCLUSION

In this research, pre-trained ResNet50 and VGG16 models were employed to recognize human activities from videos. Using the UCF YouTube action dataset for training and testing, these networks effectively discerned various human actions such as basketball playing, cycling, diving, golf swinging, horseback riding, soccer juggling, tennis strokes, trampoline jumps, volleyball spikes, and dog walking. The dimensions for the frames in ResNet50 were 112×112×15, and for VGG16, they were 224×224×15. This distinction could influence the accuracy of the results. Nevertheless, the 3DCNN models exhibited remarkable proficiency in video classification with minimal errors.

A myriad of human activities in videos were accurately classified using pre-trained networks. According to the gathered data, the training phase reduced the loss function to 0.08 and elevated the accuracy rate to 95.6%. The 3DCNN model's inherent strength lies in its ability to seamlessly process spatial and temporal information, ensuring consistent analysis across video frames. The tests yielded an accuracy of 89.71%, a recall rate of 89.03%, and F1 score of 92.77%. This suggests that the pre-trained network effectively discerns and classifies human activities. Such advancements in using pre-trained neural networks hold promise for enhancing human activity recognition, which is crucial for areas like personal security, senior care, and child safety.

REFERENCES

[1] S. R. Ke, H. L. U. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo, and K. H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013, doi: 10.3390/computers2020088.

[2] F. Rustam *et al.*, "Sensor-based human activity recognition using deep stacked multilayered perceptron model," *IEEE Access*, vol. 8, pp. 218898–218910, 2020, doi: 10.1109/ACCESS.2020.3041822.

[3] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, no. 2, pp. 690–712, 2021, doi: 10.1007/s10489-020-01823-z.

[4] K. Shreedarshan and S. S. Selvi, "Crowd recognition system based on optical flow along with SVM classifier," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2451–2459, 2019, doi: 10.11591/ijece.v9i4.pp2451-2459.

[5] A. C. S. E. Santos and H. Pedrini, "Human action recognition based on a spatio-temporal video autoencoder," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 11, pp. 1-29, 2020, doi: 10.1142/S0218001420400017.

[6] H. Zhang and L. E. Parker, "4-Dimensional local spatio-temporal features for human activity recognition," *IEEE International Conference on Intelligent Robots and Systems*, pp. 2044–2049, 2011, doi: 10.1109/IROS.2011.6048130.

[7] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[8] G. Varol, I. Laptev, C. Schmid, and A. Zisserman, "Synthetic humans for action recognition from unseen viewpoints," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2264–2287, 2021, doi: 10.1007/s11263-021-01467-7.

[9] H. Kim, S. Lee, and H. Jung, "Human activity recognition by using convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5270–5276, 2019, doi: 10.11591/ijece.v9i6.pp5270-5276.
[10] W. Liu, S. Fu, Y. Zhou, Z. J. Zha, and L. Nie, "Human activity recognition by manifold regularization based dynamic graph convolutional networks," *Neurocomputing*, vol. 444, pp. 217–225, 2021, doi: 10.1016/j.neucom.2019.12.150.
[11] Y. F. Tan, X. Guo, and S. C. Poh, "Time series activity classification using gated recurrent units," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3551–3558, 2021, doi: 10.11591/ijece.v11i4.pp3551-3558.
[12] K. K. Reddy, N. Cuntoor, A. Perera, and A. Hoogs, "Human action recognition in large-scale datasets using histogram of spatiotemporal gradients," in *2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 106–111, doi: 10.1109/AVSS.2012.40.
[13] S. Dubey, A. Boragule, and M. Jeon, "3D ResNet with ranking loss function for abnormal activity detection in videos," *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 2019, doi: 10.1109/ICCAIS46528.2019.9074586.
[14] Z. Yu and W. Q. Yan, "Human action recognition using deep learning methods," *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1-6, doi: 10.1109/IVCNZ51579.2020.9290594.
[15] P. S. A. L. Haroon and U. Eranna, "An efficient activity detection system based on skeleton joints identification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 6, pp. 4995–5003, 2018, doi: 10.11591/ijece.v8i6.pp4995-5003.
[16] N. A. Sakr, M. A. -ElKheir, A. Atwan, and H. H. Soliman, "A multilabel classification approach for complex human activities using a combination of emerging patterns and fuzzy sets," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2993–3001, 2019, doi: 10.11591/ijece.v9i4.pp2993-3001.
[17] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Generation Computer Systems*, vol. 81, pp. 307–313, 2018, doi: 10.1016/j.future.2017.11.029.
[18] G. A. Oguntala *et al.*, "SmartWall: novel RFID-enabled ambient human activity recognition using machine learning for unobtrusive health monitoring," *IEEE Access*, vol. 7, pp. 68022–68033, 2019, doi: 10.1109/ACCESS.2019.2917125.
[19] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: a review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 1309–13040, 2021, doi: 10.1109/JSEN.2021.3069927.
[20] P. S. A. L. Haroon and U. Eranna, "A simplified machine learning approach for recognizing human activity," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 3465–3473, 2019, doi: 10.11591/ijece.v9i5.pp3465-3473.
[21] P. Y. Han, S. A. P. R. Sekaran, O. S. Yin, and T. T. Guang, "Principal coefficient encoding for subject-independent human activity analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 4391–4399, 2022, doi: 10.11591/ijece.v12i4.pp4391-4399.
[22] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "Human activity classification using the 3DCNN architecture," *Applied Sciences*, vol. 12, no. 2, 2022, pp. 1-17, doi: 10.3390/app12020931.
[23] A. M. Butt, M. H. Yousaf, F. Murtaza, S. Nazir, S. Viriri, and S. A. Velastin, "Agglomerative clustering and residual-VLAD encoding for human action recognition," *Applied Sciences*, vol. 10, no. 12, pp. 1-14, 2020, doi: 10.3390/app10124412.
[24] S. Kahlouche, M. Belhocine, and A. Menouar, "Real-time human action recognition using deep learning architecture," *International Journal of Computational Intelligence and Applications*, vol. 20, no. 4, pp. 1-25, 2021, doi: 10.1142/S1469026821500267.
[25] L. Alawneh, T. Alsarhan, M. Al-Zinati, M. Al-Ayyoub, Y. Jararweh, and H. Lu, "Enhancing human activity recognition using deep learning and time series augmented data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 12, pp. 10565–10580, 2021, doi: 10.1007/s12652-020-02865-4.
[26] K. Muhammad *et al.*, "Human action recognition using attention-based LSTM network with dilated CNN features," *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021, doi: 10.1016/j.future.2021.06.045.
[27] M. Ronald, A. Poulose, and D. S. Han, "ISPLInception: an inception-ResNet deep learning architecture for human activity recognition," *IEEE Access*, vol. 9, pp. 68985–69001, 2021, doi: 10.1109/ACCESS.2021.3078184.
[28] M. A. Khan *et al.*, "A fused heterogeneous deep neural network and robust feature selection framework for human actions recognition," *Arabian Journal for Science and Engineering*, vol. 48, no. 2, pp. 1-16, 2023, doi: 10.1007/s13369-021-05881-4.
[29] T. H. Thi, J. Zhang, L. Cheng, L. Wang, and S. Satoh, "Human action recognition and localization in video using structured learning of local space-time features," in *IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2010*, 2010, pp. 204–211, doi: 10.1109/AVSS.2010.76.
[30] S. S. Hiremath, J. Hiremath, V. V. Kulkarni, B. C. Harshit, S. Kumar, and M. S. Hiremath, "Facial expression recognition using transfer learning with ResNet50," in *Inventive Systems and Control*, Singapore: Springer, vol. 672, pp. 281–300, 2023, doi: 10.1007/978-981-99-1624-5_21.
[31] J. Carreira, and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724-4733, doi: 10.1109/CVPR.2017.502.
[32] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.

BIOGRAPHIES OF AUTHORS

Seemanthini Krishne Gowda serves as an assistant professor of machine learning, having earned her Ph.D. in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum, in 2020. With a solid teaching background spanning over 14 years and more than 7 years in research, her expertise encompasses AI, machine learning, computer vision, deep learning, and neural networks. She has made significant contributions to the field, authoring over 25 papers in international conferences and journals. She plays a vital role in the Convener Institution Innovation Council, MoE Cell, and AICTE, and has successfully completed 3 projects funded by DST. Additionally, as an AICTE-certified innovation ambassador trainer, she is also a member of the technical committee for a national conference. She can be contacted at email: seemanthinik.mel@bmsce.ac.in.

Shobha Narasimha Murthy holds the position of associate professor in the Department of Computer Science and Design at Dayananda Sagar College of Engineering, Bengaluru. She completed her Ph.D. in Computer and Information Science from Visvesvaraya Technological University, Karnataka. An accomplished academic, she has contributed around 11 publications to international and national journals and conferences. Her research primarily focuses on data mining, machine learning, and data science. She can be contacted at email: shobha-csd@dayanandasagar.edu.

Jayaprada S. Hiremath who is working towards her Ph.D. in Computer Science and Engineering at Visvesvaraya Technological University, also serves as an assistant professor at Sai Vidya Institute of Technology. She brings with her a decade of professional industry experience, primarily in the fields of multimedia and networking applications. Her research interests are focused on image processing, machine learning, and artificial intelligence. She has presented her findings at various IEEE conferences and journals. Additionally, she has continually expanded her knowledge through active participation in numerous seminars and workshops. She can be contacted at email: jayaprada.researcher@gmail.com.

Sowmya Lakshmi Belur Subramanya currently serving as an assistant professor in the Department of Machine Learning at BMS College of Engineering, obtained her Ph.D. from Visvesvaraya Technological University in 2021. Her doctoral thesis, titled 'cross-language information retrieval for code-mixed kannada-english queries,' paved the way for her academic career. She commenced her role at BMS College of Engineering in October 2021, shortly after completing her doctorate. She has made significant contributions to the field of natural language processing and information retrieval, authoring 13 papers in various international journals and conferences. She can be contacted at email: sowmyalakshmibs.ise@bmsce.ac.in.

Shantala S. Hiremath is M.Tech. graduate in electronics, brings over 14 years of rich industry experience to the table. She spent 9 years at Sony Software Ltd in Bangalore, where she excelled as a domain specialist in image processing and DIPs algorithm optimization. While at Sony, she was an active participant in numerous Open Houses and idea submission challenges, showcasing her expertise and innovative thinking. Currently, she is a part of Tata Elxsi in Bangalore, serving as a subject matter expert in digital image processing. She can be contacted at email: shantala.hiremath@gmail.com.

Mrutyunjaya S. Hiremath an M.Tech. graduate in digital communication, is currently pursuing his Ph.D. in Electronics and Communication. With a professional background encompassing three years in teaching, eight years in the industry, and over ten years dedicated to research, he possesses a profound expertise in multimedia, and image and video processing applications. His academic pursuits are primarily focused on image processing, machine learning, deep learning, and artificial intelligence. His significant contributions to field are demonstrated through the four scholarly papers he has published in a range of conferences and journals. She can be contacted at email: mrutyunjaya.publications@gmail.com.