

Comparison of feed forward and cascade forward neural networks for human action recognition

Aditi Jahagirdar, Rashmi Phalnikar

School of Computer Engineering and Technology, MIT World Peace University, Pune, India

Article Info

Article history:

Received Jul 6, 2021

Revised Nov 30, 2021

Accepted Dec 9, 2021

Keywords:

CFNN

FFNN

HOF

Human action recognition

HOG

ABSTRACT

Humans can perform an enormous number of actions like running, walking, pushing, and punching, and can perform them in multiple ways. Hence recognizing a human action from a video is a challenging task. In a supervised learning environment, actions are first represented using robust features and then a classifier is trained for classification. The selection of a classifier does affect the performance of human action recognition. This work focuses on the comparison of two structures of the neural network, namely, feed forward neural network and cascade forward neural network, for human action recognition. Histogram of oriented gradients (HOG) and histogram of optical flow (HOF) are used as features for representing the actions. HOG represents the spatial features of the video while HOF gives motion features of the video. The performance of two neural network architectures is compared based on recognition accuracy. Well-known publically available datasets for action and interaction detection are used for testing. It is seen that, for human action recognition applications, feed forward neural network gives better results in terms of higher recognition accuracy than Cascade forward neural network.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Aditi Jahagirdar

School of Computer Engineering and Technology, MIT World Peace University

Pune, India

Email: aditi.jah@gmail.com

1. INTRODUCTION

Data analytics is a current buzzword in the computer industry. With immense development in digital technology, the amount of digital data generated is increasing day by day. Access to easy devices like smartphones and closed-circuit television (CCTV) cameras has contributed to vast increase in the image and video data. Analyzing this data manually has become a tedious and time-consuming task. To tackle this problem various algorithms and methods have been proposed for automatic video and image analysis. This area of research is well known by the name intelligent video analysis and finds applications in intelligent video surveillance, human-computer interaction, robotics, smart health care, smart home [1]. Human action recognition (HAR) is an integral part of intelligent video analytics.

Human action is defined by Herath *et al.* [2] as "Action is the most elementary human-surrounding interaction with a meaning". Human actions are broadly divided into gestures, simple actions, interactions, and group activities. Moving of a palm or nodding of the head is considered as a gesture. A person walking, jumping or bending is considered as a simple action. Handshake by two people or one person pushing other is considered as a human-human interaction. A person walking with a dog or picking a bag is considered a human-object interaction. More than two people talking or dancing is considered a group action [3].

Even after being researched for many years, human action recognition remains a challenging task because of its vast scope. The main challenge in human action recognition is that there is no limit to actions that can be performed by a human being. Actions like jogging, walking, and running can create confusion for an automatic action recognition system. Another obstacle in recognition is that there is a large diversity in the way in which a particular action is performed. This gives rise to high intraclass variation. Other conditions like varied camera angles, camera motion, scale changes, illumination changes, cluttered background, and occlusion add to the challenges faced by the automatic human action recognition system.

Most of the work in this area uses the supervised learning approach. The main steps in the HAR system are feature extraction, feature selection and training a classifier with extracted features [4] for the classification. The choice of features to be extracted for representing the action depends on the type of action. Algorithms proposed for gestures recognition, simple action recognition, and group action recognition differ mainly in the selection of features. Various classifiers like the k-nearest neighbor, support vector machine, and neural network are explored for classification purposes. It is observed that, along with the selection of appropriate feature, selection of appropriate classifier plays an important role in HAR performance. This work focuses on comparison of two neural network architectures, namely, feed forward neural network (FFNN) and cascade forward neural network (CFNN) for human action recognition. Two hand-crafted features, namely, histogram of optical flow (HOF) and histogram of gradients (HOG) are used for representing the action. Experimentation is carried out on well known Weizmann, KTH, UT interaction, and University of Central Florida (UCF) sports action datasets. Recognition accuracy is used as a performance parameter for comparing the architectures. The highest recognition accuracy of 97.59% is achieved for UT 1 interaction dataset with FFNN architecture. The accuracy can be improved further by using different hand-crafted features.

Earlier work on human action recognition shows the use of various hand-crafted features. Hand-crafted features are divided into two categories as local features and global features. Local feature defines the object in parts and then these parts are combined to form a local feature descriptor. Global feature defines an object as a whole. Each type of feature has its advantage and disadvantage. Previous work in this domain has emphasized the need of using multiple features to describe an action. As one type of feature can capture only one of the properties of a video, multiple features always help in describing an action efficiently as proved in [4]. Many researchers have combined local and global features for increasing recognition accuracy. Region of interest is detected before extracting actual features in many approaches [5]-[8].

Bak *et al.* [5] have used a deep learning approach for salient region detection. Various fusion mechanisms are explored for assimilating spatial and temporal features. Abdulmunem *et al.* [6] have proposed the use of an support vector machines (SVM) classifier for classifying objects described by a combination of global and local features. 3D gradient location and orientation histograms (GLOH) vector is proposed by Abdulmunem *et al.* [7]. 3D GLOH combines gradient locations and orientation histogram. In Duta *et al.* [8] new feature encoding methods namely vector of locally aggregated descriptors (VLAD) and spatio-temporal (ST_VLAD) are proposed by Ionut C. and others. The proposed method gives comparable results on datasets used for testing. A detailed study of the bag of visual word model using local features applied to human action recognition is given in [9].

A new bag of visual word framework called hybrid super vector is also proposed in this paper which gives promising results. A new feature descriptor using the fusion of stationary wavelet transform (SWT) and local binary pattern (LBP) is proposed in [10]. A discrete wavelet transform (DWT) based method is proposed in [11]. Four-level DWT is applied to find the features and then the stepwise linear discriminant analysis is applied for finding key features to be used in training. 3D stationary wavelet transform is used to describe the action in [12], [13]. Accumulate motion image (AMI) and motion AMI history image (MHI) are introduced in [14]. DWT features are further extracted from AMI and LBP features are extracted from MHI images to form a feature descriptor. Jyotsna *et al.* [15], HOG along with principal component analysis (PCA) is used to describe the action after applying the segmentation. K nearest neighbor classifier is used as a classifier which gives recognition accuracy of 94% on Weizmann and 91.83% on the KTH dataset. HOG is successfully used for face recognition in intelligent surveillance system in [16]. A combination of HOG and local feature swine confinement worker (SWF) [17] is seen to give high classification accuracy for UT interaction and UCF sports datasets.

Artificial neural networks (ANN) have reformed the domain of machine learning. ANNs are widely used in human action recognition problems because of their capability to map complex inputs to outputs. Several types of ANNs are explored for classifying different types of data. Teixeira and Fernandes [18] have compared performances of feed-forward neural network and cascade forward neural network for time series domain to prove the advantage of cascade forward neural network. Dhanaseely [19] have presented results obtained by feed-forward neural network and cascade forward neural network for face recognition dataset with principal component analysis used as a feature. Badde *et al.* [20] and Goyal [21] use of feed-forward backpropagation networks and cascade forward backpropagation networks are explored in the civil

engineering domain. Cascade forward network is shown to give better accuracy in comparison to feed-forward network.

2. PROPOSED METHOD

The method for human action recognition proposed in this work is given in this section. Block schematic of the proposed method is given in Figure 1. A test video is converted to frames and preprocessed for de-noising. For human action recognition, spatial as well as temporal information is important. A histogram of gradients is used here to represent spatial information of the action. Histogram of optical flow represents the temporal or motion features of the action. Feature selection is done using Principal component analysis. PCA is applied to both the features separately to reduce the dimensionality. HOG and HOF features are then concatenated to form a final feature descriptor. A neural network is then trained with these feature descriptors and used to classify the test video. The following sub-sections describe each step-in detail.

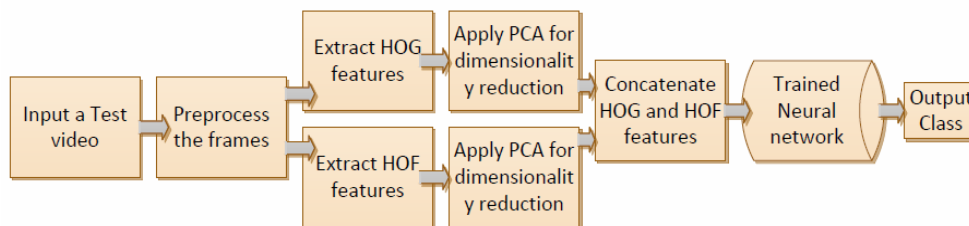


Figure 1. Block schematic for proposed human action recognition system

3.1. Histogram of oriented gradient (HOG)

Histogram of oriented gradients (HOG) describes an object in a frame using its spatial information. HOG was first developed for recognizing human figures from an image. Hence HOG becomes a perfect choice for representing spatial features in the human action recognition task. The original algorithm [22] which was developed for an image is applied to a video here by considering each frame as an image. Each frame is divided into cells which are small spatial regions and for each pixel magnitude and orientation of gradient are calculated. 1D histograms are then used to represent each cell forming the HOG feature.

3.2. Histogram of optical flow (HOF)

Histogram of optical flow is proved to have the capability of representing human body motion [23]. Optical flow is nothing but a pattern of apparent movement in a sequence that represents relative motion between the observer and the sequence. It is obtained by finding changes in the position of the object in two consecutive frames. Here, HOF is represented by optical vectors calculated at each pixel. Using the feature selection method, optical vectors having maximum value are selected to form a feature descriptor.

3.3. Neural network

Two architectures of neural networks are used separately for evaluating the performance of the system. The architectures of feed-forward neural network (FFNN) and cascade forward neural network (CFNN) [24] are shown in the Figures 2 and 3 respectively. FFNN is the most commonly used neural network model in which the input layer, hidden layers, and output layer are used. Figure 2 shows the general architecture of FFNN. All the input nodes are connected to all the nodes in 1st hidden layer and all the hidden nodes of the last hidden layer are connected to the output layer. The direction of data is only in one direction i.e. input to output. The backpropagation algorithm is used to calculate the weights between layers. Multiple layers of neurons and a backpropagation algorithm make it possible for the network to learn linear as well as nonlinear relations between input and output.

Cascade forward neural network is similar to FFNN but is having an extra weighted connection from the input layer to each hidden layer and from each hidden layer to successive layer. This extra connection from input to each layer makes the learning of the network faster. Figure 3 shows the architecture of CFNN.

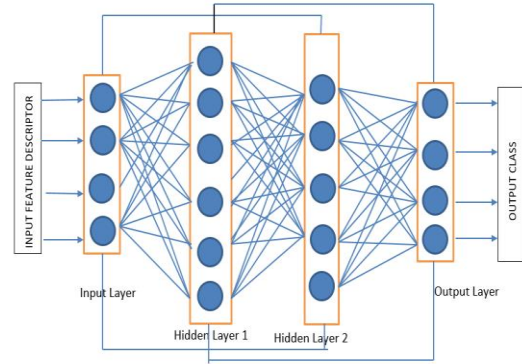
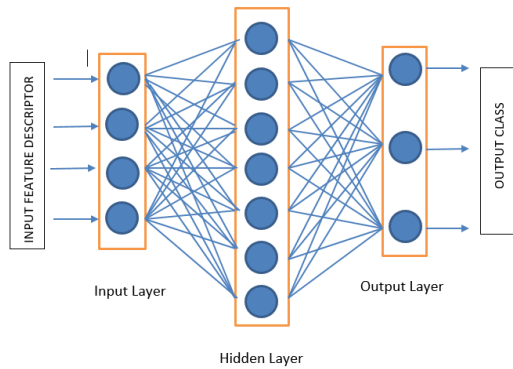


Figure 2. The architecture of neural network: FFNN Figure 3. The architecture of neural network: CFNN

For fair comparison of the performances, same parameters are used for both the neural network architectures. The hyperbolic tangent sigmoid function is used as an activation function in the hidden layer nodes. Linear function is used as the activation function for the output layer. Levenberg-Marquardt backpropagation algorithm, which is preferred for supervised learning and is fastest, is used for learning the weights. Mean square error is used as a loss function.

3.4. Datasets

Publicly available datasets, namely, Weizmann, KTH, UCF sports, and UT interaction action data sets are used for evaluating the performance of FFNN and CFNN architectures. These datasets are selected for evaluation because of their distinct properties. In the Weizmann dataset, videos of ten day-to-day actions like walking, running, and jumping are incorporated. These actions are performed by nine different actors. The recording is done in a controlled environment where only one actor is present in one frame and the background is uncluttered. Total ninety videos are available in this dataset. The complexity of the KTH dataset is more than the Weizmann dataset.

In the KTH dataset, six simple actions like hand clapping, waving, and boxing are performed by 25 different actors. Each action is performed by every actor in four different scenarios. The scenarios used are indoor, outdoor, change in scale, and change in the view angle. This dataset is having 600 videos recorded in a controlled environment. UCF sports dataset is also having one actor in every video but the recording is done at real-time sports events. There are videos of ten different sports like horse riding, golf, and diving. As these videos are recorded in real-time, they have varying backgrounds, varying view angles, illumination changes, and different scales. This increases the complexity of this dataset. UT interaction dataset is different from previously described datasets because it is having two actors in every video. There are six actions like hugging, handshaking, punching, pushing, and kicking performed by 10 different pairs. Only the action of pointing a finger is performed by a single actor. This dataset is divided into two parts as UT interaction 1 (UT 1) and UT interaction 2 (UT 2) dataset. In UT 1 dataset, actions are performed in a controlled environment. In UT 2 dataset same actions are performed with a cluttered background, partial occlusion, illumination changes, and view angle changes. In few videos of the UT 2 dataset, more than two actors are present in a frame, making the recognition task more challenging. Figure 4 shows sample frames from all the datasets used. Figure 4(a) shows a sample frame from Weizmann dataset of action class ‘walk’. Figure 4(b) shows the sample frame of action class ‘walk’ from KTH dataset. Figure 4(c) shows the sample frame from video of action class ‘Swing bench’ from UCF Sports dataset. Figure 4(d) shows the sample frame of the action class ‘shaking hands’ from UT interaction dataset.

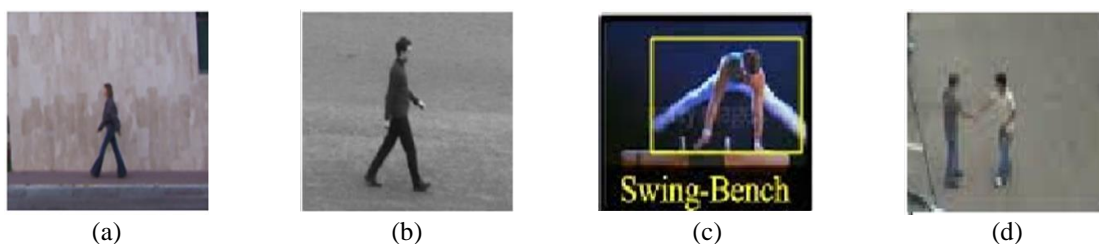


Figure 4. Sample frames from: (a) Weizmann, (b) KTH, (c) UCF Sports, and (d) UT interaction 1 datasets

For evaluating the performance, 80% of samples are used for the training, 10% for validation, and 10% for testing. Stratified sampling is used to keep the number of samples of each class proportional to the number of samples of that class in the main dataset. Each setup is run 6 times considering different samples for training, validation, and testing, and an average of accuracy, precision, and recall are calculated for both neural network models.

3. RESULTS AND DISCUSSION

Extensive testing was done to evaluate the performance of the HOG+HOF descriptor for the human action recognition system. For finding an optimum number of hidden layers to be used, experimentation was performed on all data sets with a different number of hidden layers, and accuracy was calculated. Figure 5 shows a graph of the number of hidden layers plotted against accuracy obtained. The depth of neural networks is increased by increasing the number of hidden layers from 5 to 100.

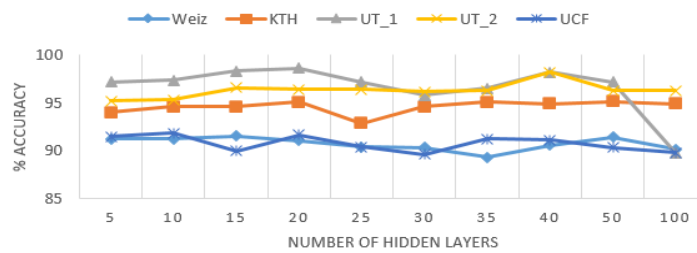


Figure 5. Effect of number of hidden layers on average accuracy

It is observed that recognition accuracy varies from 88% to 97% for different datasets. Also, recognition accuracy changes as the number of layers are changed. The time required for training the network goes on increasing as the number of layers is increased. With 15 number of hidden layers, good accuracy is achieved for all the datasets within optimal time. For all the further evaluations, 15 hidden layers are implemented. Figure 6 shows the performance of the neural network model on the UT_2 interaction data set. Figures 6(a) and (b) show the validation performance of FFNN and CFNN respectively obtained for UT 2 interaction dataset. It is seen that mean square error reduces with the number of epochs and after some epochs, it is almost constant. For FFNN, mean square error (MSE) reduces almost with same rate for training, testing and validation samples. After 10 epochs, MSE converges to the same value and remains constant thereafter. On the other hand, for CFNN, MSE reduces fast for training samples as there is a connection is present from the input layer to intermediate hidden layers. It is seen that a low value of MSE is achieved after only 2 epochs for training samples. For test and validation data samples, MSE does not reduce much and remains constant after only one epoch. This shows that because of connections between the input layer and every hidden layer, overfitting takes place which results in high MSE for validation and test dataset.

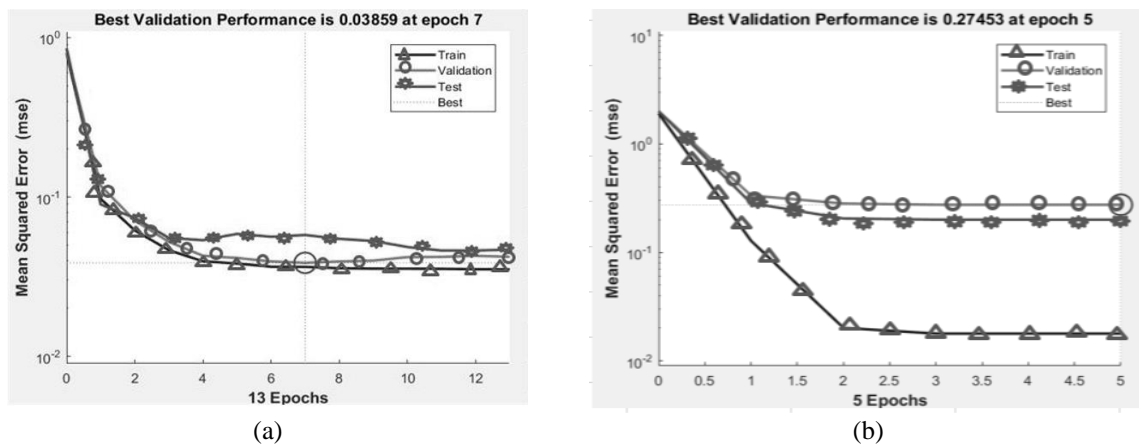


Figure 6. Sample validation performance obtained with: (a) FFNN and (b) CFNN on UT data set

Error histograms is the another measure for evaluating the performance of the classification model. The histogram of classification error plotted against the instances gives the distribution of classification error. It is seen that most of the sample points from training, testing as well as validation data fall in the bin of 0 error for FFNN as well as CFNN. The spread of error histogram is more for FFNN than for CFNN.

The graph in Figure 7 shows recognition accuracy obtained with FFNN and CFNN architectures. The accuracy obtained with both architectures is almost the same for the Weizmann dataset. For the remaining all datasets, the accuracy obtained with FFNN is more than that obtained with CFNN architecture.

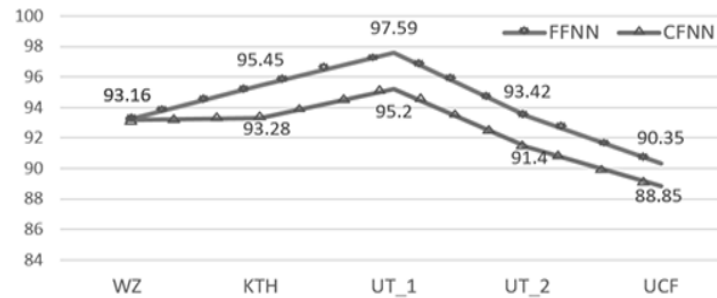


Figure. 7 Comparison of recognition accuracy obtained with FFNN and CFNN architectures

Table 1 shows the comparison of recognition accuracy obtained for UCF sports, UT interaction, Weizmann, and KTH action datasets with state-of-the-art methods presented in the literature. It is seen that for the UCF sports dataset, the proposed method with FFNN gives the highest accuracy. For the UT interaction dataset, the proposed method with FFNN architecture outperforms all other methods based on recognition accuracy. It is observed that for Weizmann and KTH datasets, comparable accuracy is obtained by the method proposed in this work. For all the datasets FFNN architecture gives better accuracy than CFNN. As in CFNN, the input layer is connected to more layers in the network, it tends to overfit reducing the overall recognition accuracy.

Table 1. Comparative results with state of art methods

State of art methods	UCF sports	UT interaction	State of art methods	Weizmann	KTH
Carvajal <i>et al.</i> [25]	88.6	--	Chaarouai <i>et al.</i> [26]	92.28%	96.70
Yi and Lin [27]	90	91.8	Junejo <i>et al.</i> [28]	89%	97.1
Wang and Qi [29]	92	83.3	Chivers [30]	97%	96.9
Weng and Guan [31]	92.8	58.2	Siddiqi <i>et al.</i> [11]	81%	80.33
Cho <i>et al.</i> [32]	89.7	85%	H. Naveed <i>et al.</i> [33]	91.69%	92.28
Nazir <i>et al.</i> [34]	94	--	M. F. Aslan <i>et al.</i> [35]	--	95.33
Ji <i>et al.</i> [36]	--	83.3	S. Zeng <i>et al.</i> [37]	98.7	--
This Work with FFNN	90.35	97.59	This Work with FFNN	93.16	94.08
This work with CFNN	88.85	95.2	This Work with CFNN	93.16	93.28

4. CONCLUSION

Experimental results show that, for human action recognition, FFNN gives higher accuracy than CFNN. It is observed that mean square error reduces fast for training datasets but stabilizes at a higher level for test and validation datasets in the case of CFNN. This shows that, in CFNN, overfitting occurs because of weighted connections present between the input layer and all hidden layers. The recognition accuracy achieved by CFNN reduces as compared to FFNN because of overfitting.

In this paper, a fusion of HOG and HOF features is used to describe human actions. HOG and HOF features are selected for this task as both of these are global features. As HOG and HOF features are extracted from the frame as a whole, the requirement of the crucial task of segmentation and foreground extraction is eliminated. A combination of HOG, which gives spatial information, and HOF which gives motion information, form a strong feature descriptor. The recognition accuracy will vary as per the fetures selected for representing the actions. In this work as the focus is on comparison of neural network architectures, various features are not explored. Comparison of results obtained in this work with the previous state of art methods shows that for Weizmann and KTH datasets, recognition accuracy obtained is comparable with other methods. For UCF sports and UT interaction datasets, which are more complex, recognition accuracy outperforms other methods.





REFERENCES

- [1] V. Ghate, "Hybrid deep learning approaches for smartphone sensor-based human activity recognition," *Multimed. Tools Appl.*, 2021, doi: 10.1007/s11042-020-10478-4.
- [2] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4-21, 2017, doi: 10.1016/j.imavis.2017.01.010.
- [3] K. Anuradha and N. Sairam, "Spatio-temporal based approaches for human action recognition in static and dynamic background: A survey," *Indian J. Sci. Technol.*, vol. 9, no. 5, 2016, doi: 10.17485/ijst/2016/v9i5/72065.
- [4] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image Vis. Comput.*, vol. 55, pp. 42-52, 2016, doi: 10.1016/j.imavis.2016.06.007.
- [5] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatiooral Saliency Networks for Dynamic Saliency Prediction," *IEEE Trans. Multimed.*, vol. 20, no. 7, pp. 1688-1698, 2018, doi: 10.1109/TMM.2017.2777665.
- [6] A. Abdulmunem, Y. K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Comput. Vis. Media*, vol. 2, no. 1, pp. 97-106, 2016, doi: 10.1007/s41095-016-0033-9.
- [7] A. Abdulmunem, Y. K. Lai, and X. Sun, "3D GLOH features for human action recognition," *Proc. - Int. Conf. Pattern Recognit.*, pp. 805-810, 2016, doi: 10.1109/ICPR.2016.7899734.
- [8] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe, "Spatio-temporal VLAD encoding for human action recognition in videos," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10132 LNCS, pp. 365-378, 2017, doi: 10.1007/978-3-319-51811-4_30.
- [9] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Underst.*, vol. 150, pp. 109-125, 2016, doi: 10.1016/j.cviu.2016.03.013.
- [10] H. M. E. Maryam, N. Al-Berry, M. A.-M. Salem, A. S. H. and M. F. Tolba, "Action Classification Using Weighted Directional Wavelet LBP Histograms," in *Advances in Intelligent Systems and Computing*, 2016, vol. 407, pp. 15-24, doi: 10.1007/978-3-319-26690-9_2.
- [11] M. H. Siddiqi, R. Ali, M. S. Rana, E. K. Hong, E. S. Kim, and S. Lee, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," *Sensors (Switzerland)*, vol. 14, no. 4, pp. 6370-6392, 2014, doi: 10.3390/s140406370.
- [12] M. Al-Berry, A.-M. Mohammed, H. Ebied, A. Hussein, and M. Tolba, "Weighted Directional 3D Stationary Wavelet-based Action Classification," *Egypt. Comput. Sci. J.*, vol. 39, no. 2, pp. 83-97, 2015.
- [13] M. N. Al-Berry, H. M. Ebied, A. S. Hussein, and M. F. Tolba, "Human action recognition via multi-scale 3D stationary wavelet analysis," *2014 14th Int. Conf. Hybrid Intell. Syst. HIS 2014*, pp. 254-259, 2014, doi: 10.1109/HIS.2014.7086208.
- [14] L. P. Suresh, S. S. Dash, and B. K. Panigrahi, "Artificial intelligence and evolutionary algorithms in engineering systems: Proceedings of ICAEES 2014, vol. 2," *Adv. Intell. Syst. Comput.*, vol. 325, pp. 309-316, 2015, doi: 10.1007/978-81-322-2135-7.
- [15] E. Jyotsna, P. V Akhil, and A. Kumar, "Silhouette based human action recognition using PCA and ISOMAP," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 11, pp. 4192-4198, 2013.
- [16] R. C. Ng, K. M. Lim, C. P. Lee, and S. F. A. Razak, "Surveillance system with motion and face detection using histograms of oriented gradients," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 14, no. 2, pp. 869-876, 2019, doi: 10.11591/ijeecs.v14.i2.pp869-876.
- [17] A. S. Jahagirdar and M. S. Nagmode, "A Novel Human Action Recognition and Behaviour Analysis Technique using SWFHOG," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 571-580, 2020, doi: 10.14569/IJACSA.2020.0110475.
- [18] J. P. Teixeira and P. O. Fernandes, "Comparison of Artificial Neural Network Architectures in the Task of Tourism Time Series Forecast," *Int. J. Comput. Inf. Eng.*, vol. 6, no. 6, pp. 830-835, 2012, doi: 10.5281/zenodo.1072894.
- [19] A. John Dhanaseely, S. Himavathi, and E. Srinivasan, "Performance comparison of cascade and feed forward neural network for face recognition system," *IET Semin. Dig.*, vol. 2012, no. 4, 2012, doi: 10.1049/ic.2012.0154.
- [20] D. S. Badde, A. Gupta, and V. K. Patki, "Cascade and Feed Forward Back propagation Artificial Neural Network Models for Prediction of Compressive Strength of Ready Mix Concrete," *IOSR J. Mech. Civ. Eng.*, no. 2278-1684, pp. 1-6, 2009.
- [21] S. Goyal and G. K. Goyal, "Cascade and Feedforward Backpropagation Artificial Neural Network Models For Prediction of Sensory Quality of Instant Coffee Flavoured Sterilized Drink," *Can. J. Artif. Intell. Mach. Learn. Pattern Recognit.*, vol. 2, no. 6, pp. 78-82, 2011.
- [22] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886-893, doi: 10.1109/CVPR.2005.177.
- [23] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačić, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1369-1376, 2010, doi: 10.1016/j.patrec.2010.03.024.
- [24] A. S. Jahagirdar and M. S. Nagmode, "Silhouette-based human action recognition by embedding HOG and PCA features," *2nd Springer International Conference on Intelligent Computing and Communication - ICICC 2017 At: Pune, India*, vol. 673, 2018, doi: 10.1007/978-981-10-7245-1.
- [25] J. Carvajal, A. Wiliem, C. McCool, B. Lovell, and C. Sanderson, "Comparative evaluation of action recognition methods via riemannian manifolds, fisher vectors and GMMs: Ideal and challenging conditions," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9794, pp. 88-100, 2016, doi: 10.1007/978-3-319-42996-0_8.
- [26] A. A. Chaaroui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1799-1807, 2013, doi: 10.1016/j.patrec.2013.01.021.
- [27] Y. Yi and Y. Lin, "Human action recognition with salient trajectories," *Signal Processing*, vol. 93, no. 11, pp. 2932-2941, 2013, doi: 10.1016/j.sigpro.2013.05.002.
- [28] I. N. Junejo, K. N. Junejo, and Z. Al Aghbari, "Silhouette-based human action recognition using SAX-Shapes," *Vis. Comput.*, vol. 30, no. 3, pp. 259-269, 2014, doi: 10.1007/s00371-013-0842-0.
- [29] X. Wang and C. Qi, "Action recognition using edge trajectories and motion acceleration descriptor," *Mach. Vis. Appl.*, vol. 27, no. 6, pp. 861-875, 2016, doi: 10.1007/s00138-016-0746-x.
- [30] D. S. Chivers, "Human Action Recognition by Principal Component Analysis of Motion Curves Human Action Recognition by Principal Component Analysis of Motion Curves A dissertation submitted in partial fulfillment by," *Browse all Theses and Dissertations*, p. 642, 2012.
- [31] Z. Weng and Y. Guan, "Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor," *Eurasip J. Image Video Process.*, vol. 2018, no. 1, 2018, doi: 10.1186/s13640-018-0250-5.
- [32] J. Cho, M. Lee, H. J. Chang, and S. Oh, "Robust action recognition using local motion and group sparsity," *Pattern Recognit.*, vol. 47, no. 5, pp. 1813-1825, 2014, doi: 10.1016/j.patcog.2013.12.004.





- [33] H. Naveed, G. Khan, A. U. Khan, A. Siddiqi, and M. U. G. Khan, "Human activity recognition using mixture of heterogeneous features and sequential minimal optimization," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 9, pp. 2329-2340, 2019, doi: 10.1007/s13042-018-0870-1.
- [34] S. Nazir, M. H. Yousaf, and S. A. Velastin, "Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition," *Comput. Electr. Eng.*, pp. 660-669, 2018, doi: 10.1016/j.compeleceng.2018.01.037.
- [35] M. F. Aslan, A. Durdu, and K. Sabanci, "Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization," *Neural Comput. Appl.*, vol. 9, 2019, doi: 10.1007/s00521-019-04365-9.
- [36] X. Ji, C. Wang, and Z. Ju, "A new framework of human interaction recognition based on multiple stage probability fusion," *Appl. Sci.*, vol. 7, no. 6, 2017, doi: 10.3390/app7060567.
- [37] S. Zeng, G. Lu, and P. Yan, "Enhancing human action recognition via structural average curves analysis," *Signal, Image Video Process.*, vol. 12, no. 8, pp. 1551-1558, 2018, doi: 10.1007/s11760-018-1311-z.

BIOGRAPHIES OF AUTHORS



Dr. Aditi Jahagirdar     received her Bachelor's degree in 1993 in Electronics & Telecommunication Engineering and a Master's degree in Electronics Engineering in 1999 from the University of Pune. She completed her Ph.D. in the domain of computer vision in 2021 from Savitribai Phule Pune University, Pune, Maharashtra, India. She is currently working as Assistant Professor in the School of Computer Engineering and Technology, at MIT World Peace University, Pune. She is having a total of 26 years of teaching experience. She is a life member of the Computer Society of India and Institution of Electronics and Telecommunication Engineers. Her research interests include computer vision, Image processing, video processing, and pattern recognition. She can be contacted at email: aditi.jahagirdar@mitwpu.edu.in.



Dr. Rashmi Phalnikar     is an Associate Professor in the School of Computer Engineering and Technology at MIT World Peace University, Pune where she has been a faculty member since 2008. She has completed her Ph.D. (Software Engineering) from Sardar Vallabhai National Institute of Technology, Surat, India. Her research topic was "A Framework for Analysis of Complex User Requirements using Aspect Oriented Use Case Method and Graph Theory". Her published work to date, (more than 50) in reputed Conferences and Journals cover the theoretical and experimental work in the domain related to Software Engineering, Role of Non Functional Requirements, Application Areas of Data Mining and Data Analysis. She is a Life Member of Indian Society for Technical Education (ISTE) and Computer Society of India (CSI). She can be contacted at email: rashmi.phalnikar@mitwpu.edu.in.