Padmaja *et al. J Big Data* (2020) 7:24
https://doi.org/10.1186/s40537-020-00296-8

Journal of Big Data

RESEARCH Open Access

A comparison on visual prediction models for MAMO (multi activity-multi object) recognition using deep learning

Budi Padmaja*, Madhu Bala Myneni and Epili Krishna Rao Patro

*Correspondence:
b.padmaja@gmail.com
Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, Telangana, India

Abstract

Multi activity-multi object recognition (MAMO) is a challenging task in visual systems for monitoring, recognizing and alerting in various public places, such as universities, hospitals and airports. While both academic and commercial researchers are aiming towards automatic tracking of human activities in intelligent video surveillance using deep learning frameworks. This is required for many real time applications to detect unusual/suspicious activities like tracking of suspicious behaviour in crime events etc. The primary purpose of this paper is to render a multi class activity prediction in individuals as well as groups from video sequences by using the state-of-the-art object detector You Look only Once (YOLOv3). By optimum utilization of the geographical information of cameras and YOLO object detection framework, a Deep Landmark model recognize a simple to complex human actions on gray scale to RGB image frames of video sequences. This model is tested and compared with various benchmark datasets and found to be the most precise model for detecting human activities in video streams. Upon analysing the experimental results, it has been observed that the proposed method shows superior performance as well as high accuracy.

Keywords: Multi-activity, Human activity recognition, Computer vision, YOLO, Video sequences

Introduction

There is no denying the fact that a human-like understanding of video surveillance by recognizing object or activity poses a significant challenge for autonomous systems in urban areas. In smart cities or environments, autonomous systems and people exist together in shared public spaces. Technology advancements have enabled the machines/systems to understand or recognize human actions in videos, but accurate and efficient human action recognition is a potential conundrum for the researchers in the field of computer vision. This area is still open for further research to develop systems that can be used productively in a stable and reliable manner. In areas where autonomous systems have to interact with people, it is very important that they have information about what people are exactly doing in their immediate environment. This is especially true if a direct interaction with the human being is to take place. Since human actions are highly

dynamic, it is not only important to predict the actions correctly but also in real-time. Action recognition and prediction are two major tasks in computer vision and action recognition. It is a primary task that recognizes human simple actions based on the complete actions in a video. It plays a key role in many domains and applications including intelligent visual surveillance [1, 2], video retrieval, gaming [3], home behavior analysis, entertainment, autonomous driving vehicle, human–robot interaction, health care and ambient assisted living [4, 5]. Human action recognition in video includes various tasks like human detection, pose estimation, human tracking, and analysis. Basically action recognition can be classified at different levels of abstraction depending on the complexity [6] of visual information. It varies from simple actions such as concept/gesture activity, interaction with object/human to a complex action as a group activity.

Related work

Human action recognition is a crucial and challenging area owing to the accomplishment of the same action in a plethora of ways, even by the same individual. Besides, due to camera view point, occlusions, noise, complex dynamic background, long-distance and low-quality videos, action recognition still remains a challenging problem. A typical action recognition framework consists of two components: action representation and action classification [7]. In action representation, an action video is converted into a series of feature vectors and in action classification; an action label is inferred from the vector [8]. However, in deep networks, the above two steps are merged into a single end-to-end trainable framework by enhancing the classification performance. Action representation is the first and foremost important problem in action recognition, because human actions differ in videos due to motion speed, camera view, pose variation, etc. The major challenges in action recognition arise due to large appearance and pose variations. So, to overcome these challenges, an action video is converted into a feature vector by extracting representative and discriminative information of human actions by minimizing the variations. Action representation approaches are broadly categorized in two ways: holistic features and local features. Holistic representations capture rich and expressive motion information of humans for action recognition, but these methods is sensitive to noise and cluttered background. Bobick et al. [9] presented Motion Energy Image (MEI) and Motion History Image (MHI) framework to encode dynamic human motion into a single image. However, these methods are sensitive to viewpoint changes. Weinland et al. [10] propounded the 3D motion history volume (MHV) to overcome the viewpoint dependency in the final action representation. Local representations overcome the problems in holistic representations by identifying local regions containing salient motion information. Local features depict local motion of a human in space–time regions which are more informative than surrounding areas. Thus, features are extracted from these regions after detection. There are many successful methods such as space–time interest points [11] and motion trajectory [12], which are based on local representations, and these techniques are robust to translation and appearance variation. Bregonzio et al. [13] used Gabor filters to detect spatial–temporal interest points (STIP) and further points was computed using Hessian matrix. Several descriptors were proposed later including 3D SIFT, HOG3D, and local trinary patterns. Laptev et al. [14] worked on local neighborhood to compute optical flow features and aggregated in

histograms, known as histograms of optical flow (HOF). Further, HOF features were combined with histogram of oriented gradients (HOG) features to show complex human activities. The author has identified and used various visual features for automatic sign recognition applications [15, 16].

Action classifiers learn from training samples to determine the accurate class boundaries for various action classes after action representations. There are other classifiers for human interactions and RGB-D videos. Ryoo and Aggarwal [17] used body part tracker to extract human interactions in videos by applying context free grammar to model spatial and temporal relationships between individuals. A human detector was adopted to recognize human interaction by capturing spatio-temporal context of a group of people and spatio-temporal distribution of individuals in videos. This method performed well on collective actions and it was further extended to a hierarchical representation which models the atomic action, interaction, and collective action all together [18]. Due to advancement of Kinect sensor, action recognition from RGB-D videos has received a lot of attention as it provides an additional depth channel compared to conventional RGB videos [19]. Many techniques such as histogram of oriented 4D normals and depth spatio-temporal interest points were proposed using depth data for action recognition task.

In recent years, many deep learning techniques have been popular due to their ability to do powerful feature learning for action recognition from massive labeled datasets [20]. There are two major variables in developing deep networks for action recognition, one is convolution operation and the other is temporal modeling. A 3D CNN is a multi-frame architecture which captures temporal dynamics in very less amount of time and can create hierarchical representations of spatio-temporal data [21]. Multi-stream network architecture contains two-stream network, a spatial ConvNet and temporal ConvNet, where the first stream learns actions from still images and the second one performs recognition based on optical flow field. This network does the fusion of outputs generated from two streams by their respective Softmax function, but it is not appropriate for gathering information over a long period of time [22]. The major drawback in the two-stream approach is that they do not allow interactions between the two streams and this is important for learning spatio-temporal features in videos. Hybrid networks contain a recurrent layer (such as LSTM) on the top of the CNN to aggregate temporal information to get the benefits of both CNNs and LSTMs [23, 24]. It has shown very good performance in capturing spatial motion patterns, temporal orderings, and long-range dependencies. In this paper, we focus on exploring the deep structure You Only Look Once (YOLO) object detection model for action recognition. YOLOv3 is a popular object detection model in real time and used to reduce the pre-training cost, increase the speed without affecting the performance of action recognition. Yan et al. [25] has introduced YOLOv3 framework for human object interaction recognition and results are achieved 93% accuracy on their own multitasking dataset.

Object detection method

YOLOv3

YOLOv3 object detector is became a popular detector due to its outstanding speed (45 frames per second). It is based on Darknet architecture (darknet-53), which has 53 layers stacked on top, giving 106 fully convolution architecture for object detection. YOLO

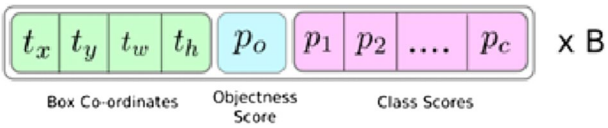

Fig. 1 The attributes of bounding box in YOLOv3

Table 1 Sample bounding box values for different image classes

Class	X	Y	Height	Width
1	0.378446	0.275689	0.526316	0.528822
2	0.428044	0.742382	0.830258	0.457064
3	0.55481	0.438479	0.237136	0.237136
4	0.415512	0.541872	0.67313	0.81626

takes the entire input image (608×608) in a single instance and divides it into an $S \times S$ grid (19×19). Then it predicts center of location (x and y axis), size (width and height) and probability of the object in each grid. For each grid cell, it estimates B bounding boxes along with its confidence score. The confidence score indicates that the probability of the box carries an activity (Threshold: 50%) and the accuracy of the box. There are 5 prediction parameters in each bounding box along with the activity classes such as Pc, x, y, h, w, c1, c1,c3,.....c80. Figure 1 shows the attributes of bounding box, where tx, ty, tw, th are the box co-ordinates, P0 is the objectness score and P1, P2, P3,.....Pc are the class scores, while B is the number of bounding boxes.

Table 1 shows sample bounding box values computed by YOLO for each image per class. The first value indicates the class number followed by values for x, y, h, w. The range of x and y values are always between (0, 0) to (1, 0), but the height and width may be more than 1, if the object fits into more than one grids in the image frame.

The network structure of YOLOv3 for object detection is shown in Fig. 2. This structure has three detection layers to detect different objects such as small, medium and large. It performs prediction in three scales by precisely down-sampling the dimensions of the input image by 32, 16, and 8, respectively. The first detection is made after 82nd layer and after the 81 layers, the image is down sampled by the network with a stride of 32. Then the second detection is made at the 94 layer and the third detection is made at the 106th layer.

CiRA-core

This experiment has implemented using CiRA-core which is based on Robot Operating System (ROS). It is a robot system integration platform developed by Tongloy et al. supported by Thailand Research Fund (TRF) and the National Science and Technology Development Agency (NSTDA). It facilitates the users in manipulating industrial robots by using deep learning. CiRA-core provides a number of modules such as DeepTrain, DeepDetect, DeepCrop, DeepLandmark, etc. In this paper, DeepTrain module is used for feature annotation (labeling the actions in the image) and training deep neural networks. Then DeepDetect module is used to recognize the human actions in the image

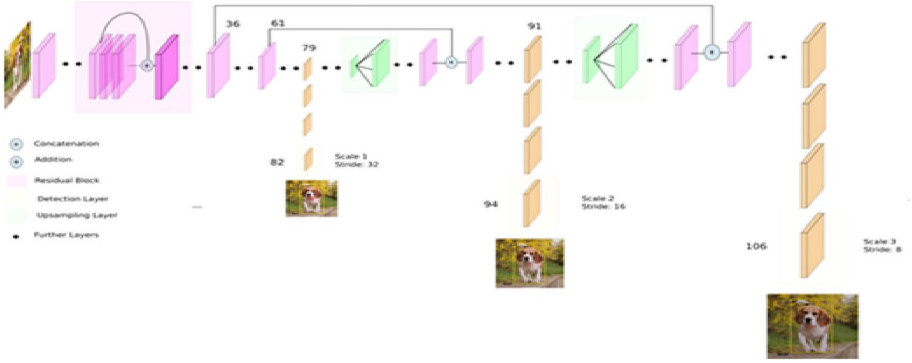

Fig. 2 Network architecture of YOLOv3

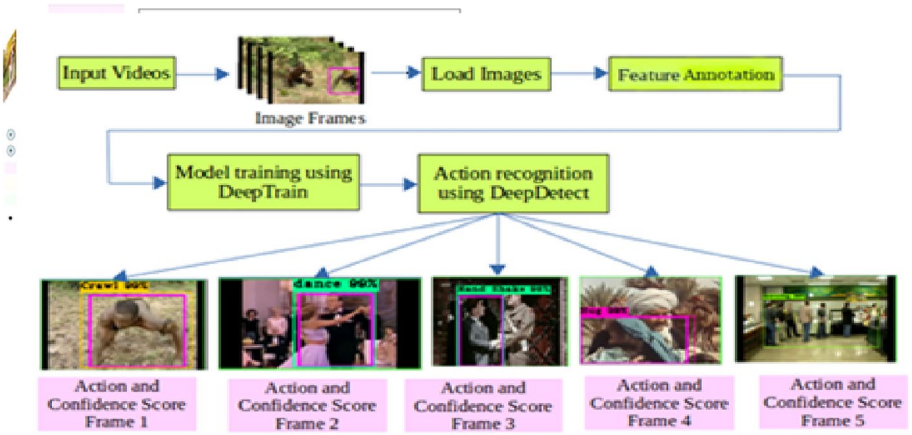

Fig. 3 Schematic diagram of proposed architecture

using deep learning weight file. For identifying human object interaction, DeepCrop module is used to crop the objects from the images, and then DeepLandmark module to detect the human-interactions with objects more accurately.

Proposed MAMO recognition visual system

Multi activity multi object recognition system (MAMO) is proposed by using YOLOv3 standard framework. Figure 3 portrays the block diagram of our proposed architecture. It is implemented in two modules such as DeepTrain and DeepDetect modules.

DeepTrain

The input video sequences are converted into image frames, and then it is loaded into the DeepTrain module. In the feature annotation step, these images are manually labeled with the action classes for various activities, and then the ground truth file (e.g. activity.gt) is prepared. Then using auto gen feature, all the images are rotated with an angle of 45 degree variation from -180 to $+180$ degree and the bounding box values for each image are computed. The image frames are trained using batch size $=64$ and sub

division = 16 to generate the object files (obj.data, obj.names), configuration file for train (train.cfg) and test (test.cfg) and weight file (train.weights). Figure 4 shows DeepTrain module on image frames of AVA dataset.

DeepDetect

The weight and configuration files are loaded to DeepDetect to check the action label and confidence score for each action in that image. Assumption of the average loss value to be 0.05 for preparing the weight file and 50% confidence score as threshold for action detection are considered.

The accuracy of human–object interaction has improved by using DeepCrop module to crop the labeled action from the image frame and then trained using DeepTrain module. Further, DeepDetect module is used to detect the interactions from the cropped images. It is observed that, there is a drastic improvement in confidence score for the human–object interactions from the cropped images. Cropped images enhance action detection more accurately than the entire image frame due to variation in background, brightness, clutter, and noise present in the image. Sometimes, certain human interactions with small objects (e.g. "cutting with a knife") can't be detected more accurately from an entire image, so, DeepCrop and DeepLandmark modules are required to improve the action recognition. To detect the small objects, DeepLandmark module has used. It is a 2-stage YOLO working on two different weight files, one on uncropped image weight file and cropped image weight file, to improve the detection of the human interactions more accurately. The Fig. 5 shows the block diagram of our work on DeepCrop and DeepLandmark.

Data description

The video data used for this experiment includes various levels of objects and interactions by considering standard datasets like KTH-6 activities, UCF-11 or YouTube Action Dataset-11 activities, AVA Dataset-80 classes in 3 categories and Collective Action Dataset-6 activities.

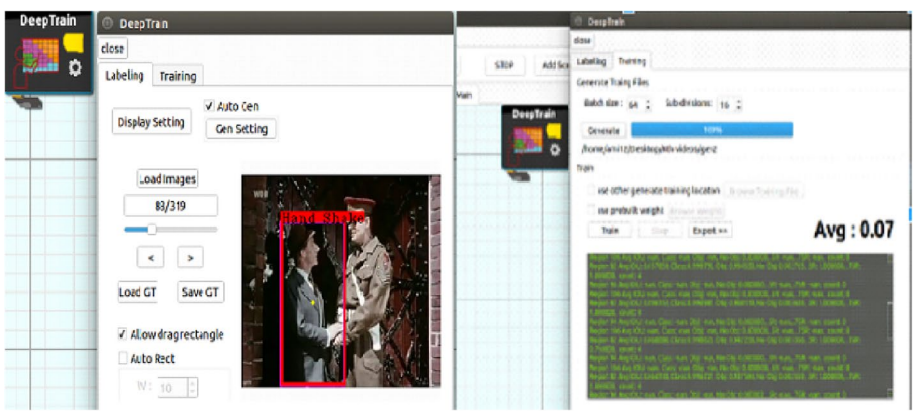

Fig. 4 Demonstration of DeepTrain module on image frames on AVA dataset

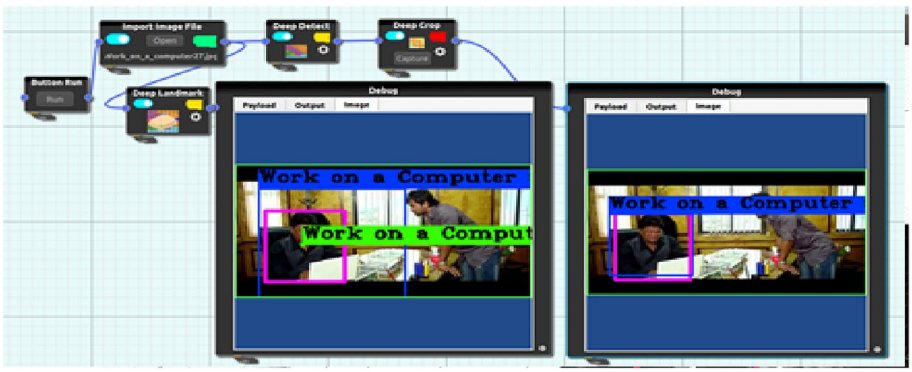

Fig. 5 The DeepCrop and DeepLandmark model for human-object interaction

KTH dataset

This dataset comprises 6 types of human actions such as walking, running, boxing, jogging, waving and clapping given in Table 2.

These are performed multiple times in 4 distinct scenarios. All the video sequences were captured with a still camera (25 fps frame rate) and over homogeneous background. Figure 6 shows human action images from KTH dataset.

UCF dataset

This is an unconstrained dataset which contains 11 types of action categories such as basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog are given in Table 3.

There are numerous issues including variations in camera motion, object appearance as well as pose, illumination, and cluttered background. All videos are categorized into 25 groups with more than 4 action clips in it and for our experiment; we have taken samples from all action clips. Figure 7 shows human action images from YouTube action dataset [26].

Table 2 Action class categories in KTH dataset with abbreviation

Action class	Abbreviation	Type of interaction
Walking	WA	Human atomic actions
Running	RU	
Boxing	BO	
Jogging	JO	
Hand waving	HW	
Hand clapping	HC	

Fig. 6 Human actions in KTH dataset

Table 3 Action class categories in YouTube action (UCF-11) dataset with abbreviation

Action class	Abbreviation	Type of interaction
Basketball shooting	BS	Human–object interaction
Biking/cycling	BC	
Diving	DV	
Golf swinging	GS	
Horse riding	HR	
Soccer juggling	SJ	
Swinging	SW	
Tennis swinging	TS	
Trampoline jumping	TJ	
Volleyball spiking	VS	

Fig. 7 Human actions in YouTube action dataset

AVA dataset

This dataset is taken from 15 to 30 min videos of 430 different popular movies, with a sampling frequency of 1 Hz and a total of 900 key frames for each movie. It contains 80 annotated atomic visual actions, where each action is localized in space and time. All the 80 classes of this dataset are given in Table 4.

It is divided into 3 categories: simple actions (14 classes), human–object interaction (49 classes), human-to-human interaction (17 classes). In each frame, a person is localized using a bounding box and the corresponding label is attached. This dataset contains classes related to atomic actions such as bending, crawling, sitting, jumping, etc. as well as interactions classes with objects and humans such as climbing, cooking, cutting, kissing, hugging, fighting, etc. Figure 8 shows action images from different categories.

Collective action dataset

This is a group action dataset which is given by the University of Michigan and contains 6 distinct collective activities: crossing, waiting, talking, queuing, dancing, and jogging given in Table 5.

Table 4 Action class categories in AVA action dataset with abbreviation

Action class	Abbreviation	Type	Action class	Abbreviation	Type
Bend	BE	Human atomic actions	Play musical instrument	PM	
Crawl	CR		Play with pets	PP	
Dance	DN		Point to	PT	
Fall down	FD		Press	PR	
Get up	GU		Pull	PU	
Jump	JU		Push	PS	
Kneel	KN		Put down	PD	
Martial Art	MA		Read	RD	
Run	RU		Ride	RI	
Sit	SI		Row boat	RB	
Sleep	SL		Sail boat	SB	
Stand	ST		Shoot	ST	
Swim	SW		Shovel	SH	
Walk	WA		Smoke	SM	
Answer phone	AP	Human–object interaction	Stir	ST	
Brush teeth	BT		Take a photo	TP	
Carry or hold	CH		Text on/look at a cellphone	TC	
Catch	CA		Throw	TH	
Chop	CH		Touch	TO	
Climb	CL		Turn	TU	
Clink glass	CG		Watch	WA	
Close	CS		Work on a Computer	WC	
Cook	CK		Write	WR	
Cut	CU		Fight or hit	FH	Human–human interaction
Dig	DG		Give or serve	GS	
Dress	DS		Grab	GR	
Drink	DN		Hand clap	HC	
Drive	DR		Handshake	HS	
Eat	EA		Hand wave	HW	
Enter	EN		Hug	HU	
Exit	EX		Kick	KI	
Extract	ET		Kiss	KS	
Fishing	FI		Lift a person	LF	
Hit	HI		Listen to a person	LP	
Kick	KI		Play with kids	PK	
Lift	LI		Push another person	PP	
Listen	LT		Sing	SI	
Open	OP		Take an object from a person	TO	
Paint	PA		Talk to a person	TP	
Play board game	PG		Watch a person	WP	

These are performed by people from 44 short video sequences. These videos are recorded by consumer hand-held camera from different view-points. In all video sequences, every 10th frame is annotated with image location of the person, activity id and pose direction. Figure 9 shows group activity images from collective action dataset.

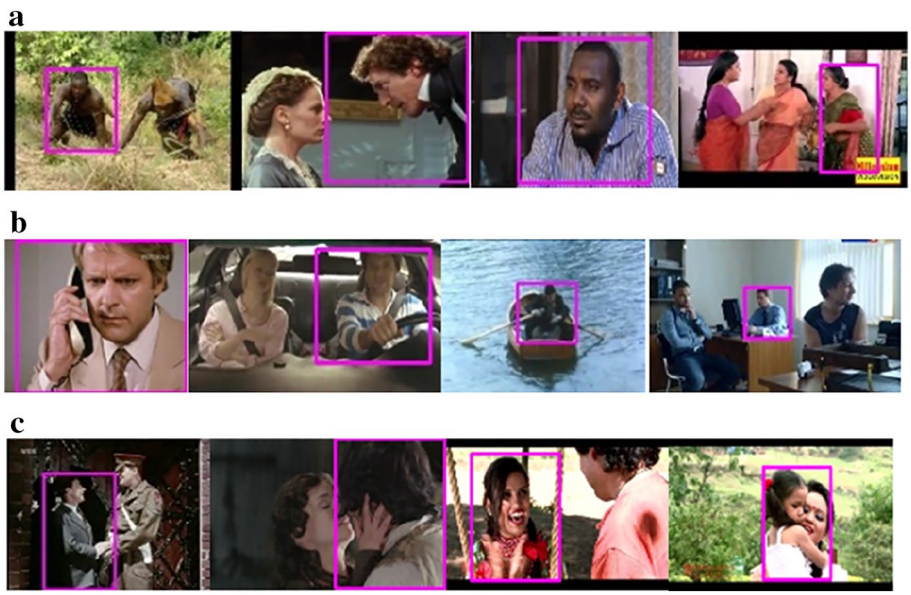

Fig. 8 Human activities from AVA datasets: **a** human atomic movements (crawl, bend, sit, stand); **b** human–object interactions (answer phone, driving, row boat, work on a computer) and **c** human–human interactions (hand shaking, kiss, hand clap, lift a person)

Table 5 Action class categories in collective action dataset with abbreviation

Action class	Abbreviation	Type of interaction
Crossing	CR	Group-actions
Waiting	WA	
Talking	TA	
Queuing	QU	
Dancing	DA	
Jogging	JO	

Results and discussions

In this experiment, the model has trained to 10,000 iterations with an average loss of 0.05 with a batch size of 64 and 16 subdivisions. We have taken different challenging datasets (KTH, UCF-11, AVA, collective action) to train our model and predict the accuracy of human activities. The model performance has assessed through Intersect over union (IoU) measure. We have chosen the weights file with the highest IoU and then used this weight file for detection (train_10000.weights). The experiments were performed on Intel(R) Core(TM) i5-8600 CPU@ 3.10 GHz with GPU(GeForce GTX 1070 Ti), Graphics card RAM size of 8 GB, and on Ubuntu 16.04LTS (64 bit) operating system. Table 6 shows some of the training parameters used by DeepTrain model on sample datasets.

Our primary evaluation metric is prediction accuracy on different interactions (actions) taken from AVA image datasets. Figure 10a–c shows the accuracy at each

Fig. 9 Collective human actions (crossing, queuing, talking, waiting)

Table 6 Training parameters generated by DeepTrain model

Sample training parameters (train.cgf) generated by DeepTrain	
Batch = 64	Exposure = 1.5
Subdivisions = 16	Hue = 0.1
Width = 608	Learning_rate = 0.001
Height = 608	Burn_in = 1000
Channels = 3	Max_batches = 500200
Momentum = 0.9	Policy = steps
Decay = 0.0005	Steps = 400,000, 450,000
Angle = 0	Scales = 0.1, 0.1
Saturation = 1.5	

class along with linearity among classes of 3 categories of actions such as human atomic actions, human–object interactions and human-to-human interaction.

In AVA dataset, it is observed that various classes like sit (SI), sleep (SL), bend (BE) from human atomic actions, answer phone (AP), hit (HI) from human–object interactions and give or serve (GS), grab (GR) from human-to-human interactions have more mismatched classification done by model. Figure 11 shows the confusion matrix for 14 category subset by using deepTrain model of action based classification. This is used to calculate precision, recall, specificity, F1-score, and overall accuracy measures of the model. Figure 12 shows automatic visual predictions of human actions with confidence score fromdifferent datasets.

The results obtained are compared quantitatively with the state-of-the-art techniques proposed so far for the datasets used in this paper and it is presented in Table 7. Based on the comparison, it is observed that YOLOv3 shows the best results for action recognition on various challenging datasets.

Conclusion

In this paper, we proffer a real-time model for human action recognition in videos by employing the avant-garde object detector YOLOv3. It is observed that YOLOv3 detects the activities more accurately by taking even small number of frames with high confidence score. And this model performs action recognition very well irrespective of occlusion, cluttered background, variation in viewpoint, inter and intra-class similarities present the image frames. We focused on detecting simple to complex human actions using gray scale to RGB image frames taken from video sequences.

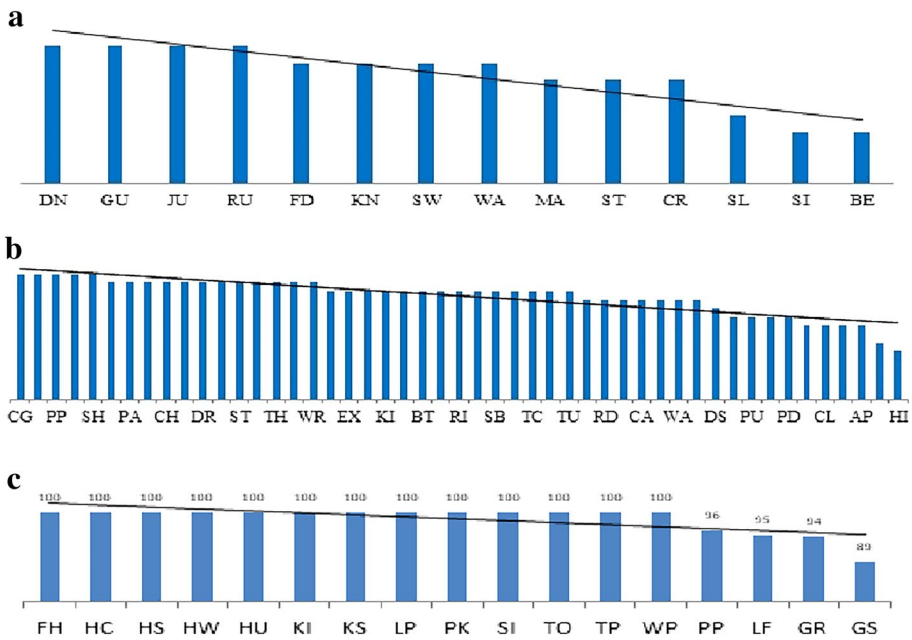

Fig. 10 Prediction accuracy of 3 categories of interactions using YOLO v3 on various classes in AVA dataset; **a** human atomic actions; **b** human–object interactions and **c** human-to-human interaction

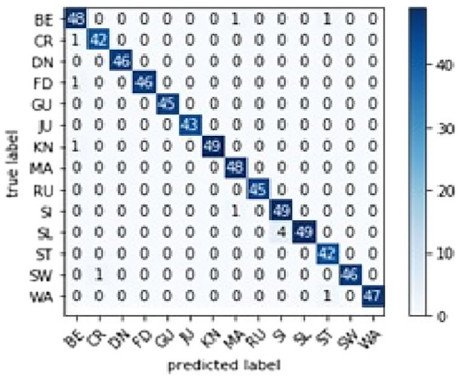

Fig. 11 Confusion matrix for 14 category subset by using deepTrain model of action based classification

DeepLandmark model more accurately detects human activities performed with small objects such as "smoking a cigar", "cutting with a knife", "fishing" etc. It is also observed that YOLOv3 take more time for training on large datasets. The aim of this paper is to focus on complete human behavior analysis through human actions and we found YOLOv3 to be the more accurate human action detector so far.

Fig. 12 Automatic visual predictions of human actions with confidence score from different datasets, **a** KTH; **b** UCF-11; **c** AVA

Table 7 Comparison of accuracy on different datasets (KTH, UCF-11, AVA, collective action)

Dataset	References	Technique used	Earlier result (%)	Proposed result (%)
KTH	Jhuang et al. [27]	SVM	91.70	98.47
	Lin et al. [28]	k-NN	93.43	
	Liu et al. [29]	Adaboost with C.45	93.80	
	Kim et al. [30]	NN	95.33	
UCF-11	Liu et al. [26]	Adaboost with C.45	71.2	99.52
	Cho et al. [15]	SVM and Kernel group sparsity	84.2	
	Ravanbakhsh et al. [31]	CNN	88.0	
			89.5	
AVA	Ulutan et al. [32]	Actor conditioned attention maps	97.10	99.74
			99.74	
Collective action	Choi and Savarese [33]	Multiclass SVM	79.2	99.97
	Choi et al. [34]	Randomized spatio-temporal volume (RSTV)	82.0	

Acknowledgements
The research work presented in this paper is carried out at CiRA (Center of Industrial Robots and Automation) lab, College of Advanced Manufacturing Innovation, KMITL, Bangkok. We gratefully acknowledge Dr. Siridech Boonsang (Dean) and Dr. Santhad Chuwongin (Head of CiRA) for their guidance and encouragement. We would like to thank Dr. L. V. Narasimha Prasad, Principal and Dr. D. Shobha Rani, Dean International Affairs, IARE for their support and encouragement. We would like to thank Neha for her assistance in editing and proofreading the paper.

Authors' contributions
BP made substantial contributions in formulating the concept, design, and carried out experimentation, analysis and interpretation of data; MMB participated in statistical analysis of result and investigation process. EKRP supported in coding, testing and drafting the manuscript. All authors read and approved the final manuscript.

Authors' information
Ms. B. Padmaja is a faculty member in the Department of Computer Science, Institute of Aeronautical Engineering, Hyderabad, Telangana, India. She has received her B.Tech from North Eastern Regional Institute of Science and Technology (NERIST), Arunachal Pradesh, India in 2001. She completed her M.Tech from School of IT, JNTUH, Hyderabad, India. Currently she is pursuing her research in "Reality Mining: Smart Phone Based Human Behavior Analysis" from JNTUH, Hyderabad. She is a member of ISTE and CSI. She has more than 18 years of teaching experience and published 20 researched papers in various International Journals and Conferences.

Dr. Myneni Madhu Bala is a professor in Computer Science and Engineering, Institute of Aeronautical Engineering. She holds her Ph.D. in Image Mining from Jawaharlal Nehru Technological University, Hyderabad in 2015. Her major research areas are Data Analytics, Image Processing, Natural Language Processing and Data Mining. She has more than 20 years of teaching experience and published more than 40 research papers in various International Journals and Conferences. She has published 2 patents, 2 book chapters and 1 research project to her credit. She was felicitated as "IWN Unsung Hero" for outstanding contribution and achievements in Research and Innovations in the field of Engineering in Telangana Leadership Conclave, Consortium of Indian Industries (CII) and Indian Women Network (IWN) held at Hyderabad, 2019.

Mr. E. Krishna Rao Patro is a faculty member in the Department of Computer Science, Institute of Aeronautical Engineering, Hyderabad, Telangana, India. He has received his MCA from Anna University, Chennai and M.Tech in Computer Science and Engineering from JNTUH, Hyderabad, India. Currently he is pursuing his Ph.D. in Intrusion Detection System using Machine Learning from VelTech University, Chennai. He has more than 22 years of teaching experience and published 10 researched papers in various International Journals and Conferences.

Funding
This research work is not funded by any organization.

Availability of data and materials
All the datasets are publicly available. KTH dataset: http://www.nada.kth.se/cvap/actions/. AVA dataset: https://research.google.com/ava/. UCF-11 dataset: https://www.crcv.ucf.edu/data/UCF101.php. Collective Activity dataset: http://vhosts.eecs.umich.edu/vision//activity-dataset.html

Competing interests
The authors declare that they have no competing interests.

Received: 3 September 2019 Accepted: 21 February 2020
Published online: 18 March 2020

References

1. Singh S, Velastin SA, Ragheb H. Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: 2010 seventh IEEE international conference on advanced video and signal based surveillance (AVSS); 2010. p. 48–55.
2. Xiang T, Gong S. Video behavior profiling for anomaly detection. IEEE Trans Pattern Anal Mach Intell. 2008;30(5):893–908.
3. Shirai A, Geslin E, Richir S. Wiimedia: motion analysis methods and applications using a consumer video game controller. In: Proceedings of the 2007 ACM SIGGRAPH symposium on video games, 2007. New York: ACM; 2007. p. 133–40.
4. Pantic M, Pentland A, Nijholt A, Huang T. Human computing and machine understanding of human behavior: a survey. Artifical intelligence for human computing. Berlin: Springer; 2007. p. 47–71.
5. Kidd C, Orr R, Abowd G, Atkeson C, Essa I, MacIntyre B, Mynatt E, Starner T, Newstetter W. The aware home: a living laboratory for ubiquitous computing research. Cooperative buildings: integrating information, organizations, and architecture. Berlin: Springer; 1999. p. 191–8.
6. Poppe R. A survey on vision-based human action recognition. Image Vis Comput. 2010;28:976–90.
7. Shi Q, Cheng L, Wang L, Smola A. Human action segmentation and recognition using discriminative semi-markov models. IJCV. 2011;93:22–32.
8. Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal multiplier networks for video action recognition. In: IEEE conference on computer vision and pattern recognition (CVPR); 2017. p. 7445–54.
9. Bobick A, Davis J. The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell. 2001;23(3):257–67.
10. Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst. 2006;104(2–3):249–57.
11. Klaser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. In: BMVC, 2008; 2008.
12. Wang H, Kläser A, Schmid C, Liu CL. Action recognition by dense trajectories. In: IEEE conference on computer vision & pattern recognition, Colorado Springs, United States; 2011. p. 3169–76.
13. Bregonzio M, Gong S, Xiang T. Recognizing action as clouds of space-time interest points. In: CVPR, 2009; 2009.
14. Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: CVPR, 2008; 2008.
15. Cho J, Lee M, Chang HJ, Oh S. Robust action recognition using local motion and group sparsity. Pattern Recognit. 2014;47(5):1813–25.
16. Padmaja B, Rao PN, Bala MM, Patro EKR. A novel design of autonomous cars using IoT and visual features. In: 2018 2nd international conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018. New York: IEEE; 2018. p. 18–21.
17. Ryoo M, Aggarwal J. Recognition of composite human activities through context-free grammar based representation. In: CVPR, vol 2; 2006. p. 1709–18.

18. Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition. In: ECCV. Berlin: Springer; 2012. p. 215–30.
19. Hadfield S, Bowden R. Hollywood 3D: recognizing actions in 3D natural scenes. In: CVPR, Portland, Oregon; 2013.
20. Donahue J, Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: CVPR, 2015; 2015.
21. Taylor GW, Fergus R, LeCun Y, Bregler C. Convolutional learning of spatio-temporal features. In: ECCV, 2010; 2010.
22. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: NIPS, 2014; 2014.
23. Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: deep networks for video classification. In: CVPR, 2015; 2015.
24. Padmaja B, Rama Prasad VV, Sunitha KVN, Vineeth Reddy G. Deep RNN based human activity recognition using LSTM architecture on smart phone sensor data. J Fundam Appl Sci. 2018;10(5S):1102–15.
25. Yan W, Gao Y, Liu Q. Human-object interaction recognition using multitask neural network. In: 2019 3rd international symposium on autonomous systems (ISAS), Shanghai, China; 2019. p. 323–8.
26. Liu J, Luo J, Shah M. Recognizing realistic actions from videos in the wild. In: CVPR 2009, Miami, FL; 2009.
27. Jhuang H, Serre T, Wolf L, Poggio T. A biologically inspired system for action recognition. In: IEEE 11th international conference on computer vision; 2007. p. 1–8.
28. Lin Z, Jiang Z, Davis LS. Recognizing actions by shape-motion prototype trees. In: IEEE 12th international conference on computer vision; 2009. p. 444–51.
29. Liu J, Luo J, Shah M. Recognizing realistic actions from videos in the wild. In: IEEE conference on computer vision and pattern recognition; 2009. p. 1996–2003.
30. Kim TK, Wong SF, Cipolla R. Tensor canonical correlation analysis for action classification. In: IEEE conference on computer vision and pattern recognition; 2007. p. 1–8.
31. Ravanbakhsh M, Mousavi H, Mohammad R, Murino V, Davis LS. Action recognition with image based CNN features. In: IEEE conference on computer vision and pattern recognition (CVPR), December 2015; 2015.
32. Ulutan O, Swati R, Srivatsa M, Torres C, Manjunath BS. Actor conditioned attention maps for video action detection. Computer Vision and Pattern Recognition; 2019.
33. Choi W, Savarese S. Understanding collective activities of people from videos. IEEE Trans Pattern Anal Mach Intell. 2014;36:1242–57.
34. Choi W, Shahid K, Savarese S. Learning context for collective activity recognition. In: IEEE conference on computer vision and pattern recognition (CVPR); 2011.

Publisher's Note