Survey On Past and Current Trends in Applying Deep Learning Models in Estimating Human Behaviour

Sajeev Ram[1] . C S Shylaja[2] . K.Kalaivani[2] . Ulagapriya[2]

[1]Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore.
[2]Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, Chennai.

Received: 10 July 2022 / Revised: 23 July 2022 / Accepted: 06 Aug 2022
©Milestone Research Publications, Part of CLOCKSS archiving

Abstract –In recent times application based on computer vision are widely used in many fields starting from lifesaving medical devices to home entertainment systems. One of the challenging problems in it is human behaviour estimation. This process of human behaviour prediction systems requires multiple technology integration. Hence it become more important to narrate out the past and the current trends used in human behaviour predictions. This article briefs out the major steps used in the detection process, a literature survey on various contributions and techniques used by different researchers in different period of time and the dataset used by them in training and testing the system. All the systems are compared and the pros and cons were analysed in detail and the research gaps were also discussed. The application of human behaviour estimation extends in various platforms and one of it being the monitoring the use of mobile phones, which are becoming a serious issue in recent times.

Index Terms – Computer Vision, Deep Learning Methods, Image Processing, Human Detection, Action Detection, Interaction Recognition

I.INTRODUCTION

The phenomenal growth in computer vision detection of objects, humans, their activities were made possible. Hence detecting the behaviour is not anymore a faraway thing for us. The detection of human behaviour was happening for the past few years. It was useful in many applications including Gaming, Disaster management, human assistance systems, health devices, communication devices, visual recording devices, security devices and lot more. They have been the base for many software and industrial products nowadays. Although it sounds to be a widely used product, it is not that much easy to proceed with these detection algorithms. The object and human identification process require many processing steps to take place, and the significant steps include human identification, interaction analysis and activity analysis. This detection process includes different factors to be considered, camera position, gesture movements and usual activities like head position changes and joint movement.

Recognising the specific behaviour requires many templates and classes to which they are associated have to be defined. Not all the behaviours of humans are possibly detected; most of them could be detected with the help of advanced detection algorithms, those behaviours which are not generally detected are termed as anomaly behaviour. These anomalies could also

be helpful since these behaviours could be neglected during the process of detecting a particular behaviour.

The article is organized as follows; Definitions on human action behaviour estimation and its stage are explained in chapter 2. Chapter 3 explains about the different methodologies used in human detection process and the dataset available for human detection are also presented. Action recognition techniques and the related datasets are mentioned in chapter 4. Chapter 5 deals with the interaction recognition systems and its datasets finally chapter 6 Concludes the paper with the future perspectives in this research area.

II.DEFINITIONS

Human activities can be broadly classified into four main categories: 1. Gestures, 2. Actions, 3. Behaviours, and 4. Interaction visually represented inFig. *1*. Movements by a human which is a general action or used to comminute in sign language are called *gestures*. Some of the examples of gestures include rising the arm/leg showing thumb for like. Gestures are usually considered when there are in a position and not in motion.*Actions* are a simple movement which could be found in humans as simple movement patterns. The observer could quickly identify these by others doing it. Some of the examples of activities include walking, jumping, turning around, climbing, sitting.

Behaviours are things which they do externally as work, and these can be identified by the surrounding and the objects along with them. Some of the examples of behaviour include hugging, sweeping, taking class, reading news articles. *Interactions* are the way of communicating with one another; it may be a one to one communication, or many people communicate with each other. Example of interaction includes talking with one another or as a group. Interactions with objects are quite different from social interaction, and these include ATM transaction, cooking. Those interactions which are made with transactions are mostly included in actions or behaviour.

The toughest thing is when the number of persons on screen increases, it is hard to identify the behaviour, as they might be talking as a group or each member will be doing different activities. In this scenario, it is tough for the system to detect the behaviour of humans.

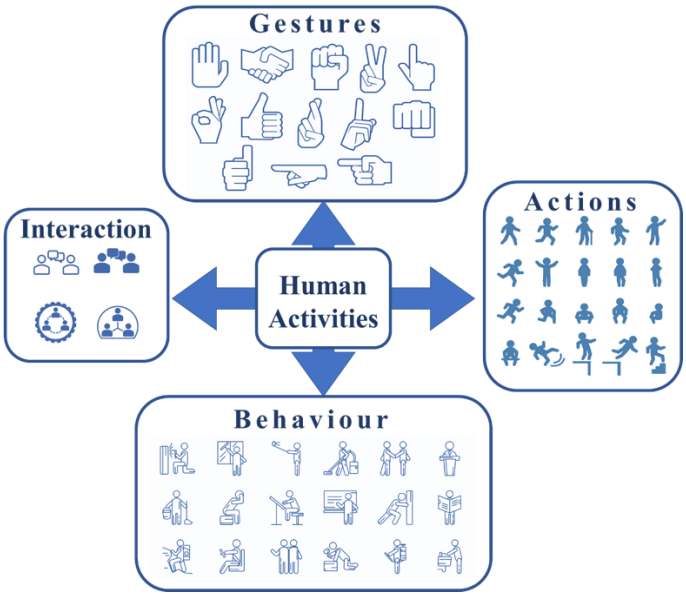

Fig. 1 Classification of Human Activities

III.Human Detection

Video-based Human detection is classified into three main classes, namely,1.Appearance Based Methods, 2. Motion-based Methods and 3. Hybrid methods.All three methods are different from each other by their methods. Appearance-based methods are performed in photographs and use edge detection-based techniques to do it, mostly used in attendance entry devices and similar areas. Motion-based human identification is based on the movement of the human by his movement in legs or hands or both together; these methods are used in many applications such as automatic recorders, human helping aids. Hybrid methods are a combination of the other two methods, the videos are converted into multiple time frame images, and the detection is performed. Hybrid detection is used in the application of surveillance cameras, obtaining responses from a group of audience.

MILESTONE RESEARCH PUBLICATIONS

3.1 Appearance Based Methods

Appearance-based detection methods can be used in non-static cameras; these methods generally include a histogram of oriented gradients (HOG) combined with classification algorithms like SVM. The HOG can be used for detecting the human by identifying the edges of the image captured in the screen. Apart to this method of extracting features from the segmented image can be passed on to the classifier and could be used to detect a human in an image. All these detection algorithms are initially trained with the available public databases then used for detection purposes.

3.2 Motion-based methods

Detection of humans based on the movement by the movement in the humans from the foreground objects is termed as motion-based detection methods. It is also done on the basis of change in the postions, objects and the background. The pixels of the image are closely monitored and the oscillation in pixel. Many works have adapted Discrete Wavelet Transforms for the detection process at the earlier days, which monitors the oscillation of the pixels. Few methods involving calculating the similarity index for every frame and the variance is calculated by which the motion is being predicted. Histogram based methods could also be used in this type of detections as there might be a change in the histogram of every pattern in the image.

People detection could also be performed by analysing the infrared domain in the image spectrum. Also, it should be noted that analysing the brightness levels in the image is inefficient in the process of human detection.

3.3 Hybrid Methods

Developing two different approaches which identify the appearance and motion of the human separately and later their results are merged accordingly with the decision function.In this method, the human in motion is not only detected perhaps all the humans in the frame will be deliberately identified. The significant disadvantages also relay on the same factor as the process is taking to analyse the entire frame the computational complexity and time also increases.

As the technology grows up the computational complexities are getting reduced using the deep processing algorithms, and significantly reduce in the False Positive rate could be noticed. The comparison of various methods and the detection rates are listed in Table ***1***.

Table 1 Performance of Human Detection Approaches

Ref	Features	Classifiers used	Detection Rate
[1]	HOG and Appearance	SVM	89%
[2]	Appearance	SVM	82%

[3]	Histogram based edge detection	SVM	80%
[4]	Shape-based features	Top-down segmentatior	71%
[5]	Motion history image	Thresholding based system	45%
[6]	Spectral analysis	Similarity comparison	91%
[7]	Contour motion	Boosting	90%
[8]	HOG	SVM	95%
[9]	3D surface model	Model fitting	98.4%
[10]	HOG	Multi-level classifier	96%

3.4 Datasets

For working with deep learning algorithms, it is necessary that we must have a data set, as we have to train the machine, and test the machine. There are many public dataset available for pedestrian detection and the names and the Avilabile information on the data set are mentioned in Table 2

Table 2 Dataset for pedestrain

Dataset	Information About the Dataset
MIT Pedestrian Data [11]	• 64 x128 PPM form images • 924 files
INRIA Person Dataset [12]	• 2,478 positive samples and 1,218 negative images for training, and 1,128 positive samples and 453 negative images for testing
CALTECH Databas [13]	• Pictures of objects belonging to 101 categories • 40 to 800 images pe category. • The size of each image is roughly 300 x 200 pixels
CAVIAR Test Case Scenarios [14]	• Several sequences o videos in mpeg2 format
ETH Dataset [15]	• Pictures of objects belonging to 8 categories. • The size of each image is roughly 640 x 480 pixels

MIT pedestrian Dataset is one of the earliest datasets which is available in public domain, the images in the dataset is having pedestrians in front and rear views. The images are resized to 64X128 pixels and also adjusted to keep the pedestrian at middle of the image.Similarly like MIT, INRIA also comes with large possible variety, and also in different posses when compared with MIT. Some advantage of using INRIA is it available along with annotation.

PETS Data sets contains the pictures that are taken in indoor as well as in out door too. ETH dataset has the annotation as well as the

MILESTONE RESEARCH PUBLICATIONS

calibration information along with the image. Samples from MIT and INRIA are shown inFig.2.

(a)

(b)

Fig.2 Images from (a) MIT Dataset (b) INRIA Dataset

IV.ACTION RECOGNITION

Action recognition are generally proceeded in three perspectives. 1. Low-level features, 2.Mid and high-level representation, 3. Silhouettes. Each method performs the process in different manner, Low level featuring methods uses dense optical flow features and Spatio interest points. Mid and high level representation uses semantic features for detection and Silhouettes are the solid dark shade of the human and with this the action recognition is performed. The three methods are narrated briefly and different approaches made in action recognition is compared inTable 3.

4.1 Low-level features

Dense optical flow, and spatial interest points (sip) are the commonly used approaches. As technology started to drive fast few researchers tried to implement sip in the edge detection methods which later strived the applications to come up with action recognition. Pre filtration process such as Gabor filer was also implemented to differentiate the intensity levels between layers.

Dense optic flow being successful for action recognition, it has been implemented in moving cameras. These cameras are widely used in capturing the complex motion pictures. These methods are trained and tested using sport and youtube data sets. The details of the various datasets are briefed out in section 4.4.

4.2 Mid and high-level representation

Long-term tracked trajectories and semantics are the higher-level features that are exploiting during action detection process. Activities including walking, talking and queuing are analysed well by these systems and have provided higher accuracy rate in detection. Early-stolcke-algorithms can be applied inorder to analyse the character behaviour of the specific actions and could be marked as context free grammar.

Another interesting methodology which are used in recent times for the detection of actions is poselets. Poselets are the detection method of actions of humans using 3D human poses along with annotations. The models developed using these type of algorithms summerize a storyline compromising the pose and the annotation making the action detection process easier.

4.3 Silhouettes

Silhouettes are the dark solid region of a human in an image. The main use of using it is, the human movements are being represented as continuous progression. These approaches mainly relay on the traditional segmentation methods. Sequences of silhouettes form the action descriptors as continuous frames and they are further analysed and recognised by the traditional classifiers. From the characteristics of each silhouette a dynamic model is alternatively build to help us in some tough situations. These tough situations might occur in cases, when the human is very close to the capturing device or in the extreme borders of the image.

The advanced deep learning methods also predicts the human actions by analysing the skeleton body structure is developed based on the image and the actions are monitored and with the help of the dataset being trained the actions of the pedestrian are also identified in few systems.

Table 3 : Comparative analysis of action recognition algorithms

Ref	Features	Classifier used	Findings
[16]	MHI and MEI	Monoboli distance based	Findings in mul view and real time
[17]	Spatio points inde	GMM	Using GMM features transformed into grey level image
[18]	Action graphs and 3D points	Max. likelihood decoding	3D points are identified for formation of 3D shape
[19]	Self similarity	KNN	Able to give 100% accuracy
[20]	Motion trajectory	Multi Thread Parsing	automatic rules induction
[21]	Motion trajectory	SVM	Classification or trajectory grouping
[22]	Low or mid-level features	SVM	Classification based on pedestrian actions
[23]	3D image patches	KNN	extract action primitives
[24]	Action detection-windows	3D space-time window	OJLA with multiple labels for comparison

4.4 Dataset

The dataset used for detection of human actions may be natural are target specific data. Instances such as walking, running, jogging, hand waving, boxing, and clapping the hands which are performed by different subjects in different environment are available in KTH dataset[25] and Weizmann dataset[26]. Some of the realistic data could be available with UCF50[27] and Hollywood[28]. Multi camera view data are available in few datasets including UCF50 and Hollywood. Camera motion data are available with YouTube dataset[29].

Hollywood Human Action datasets have huge collection with annotations. 3D surveillance dataset are available in 3DPES[30], MIT trajectory dataset[31]and the Edinburgh Informatics Forum Pedestrian Database (EIFPD)[32]. Gesture recognition process which is also comes under action detection also requires dataset for training and testing and those datasets can also be found in public and includes ASL[33], FGnet[34], Pointing'04 [7], Cambridge Gesture Database [35], and the ChaLearn Gesture Challenge dataset[36]. The RGBD-HuDaAct [37]and the ChaLearn Gesture Challenge [36]datasets include data about RGB-D datasets.

V.INTERACTIONS RECOGNITION

Interactions are generally classified as one to one interaction and one to many interaction or group interactions. Both the interactions are to be considered in different categories and hence the detection process also varies accordingly. In the process of interaction detection the sociological and psychological features are to be collected and processed as per the requirements. The performance of various interaction detection systems is compare in Table ***4.***

5.1 One-to-one interactions

The detection process is carried over by calculating the energy function among the axis which connects both of them. Single moving agents entities are called as agents. Each agent is corelated with features like position, speed and direction. a Linear Trajectory Avoidance (LTA) model can be derived from the energy levels based on the predictions. The one to one predictions could not be achieved in longer levels.

5.2 Group Interactions

Extending the one to one interaction methods to several subjects makes the prediction of group interaction. The dynamics rule for group interaction will be different from the other interaction different methods.In further the group interaction can be further classified in to small group and large group. Small group has three categories namely, self, pair and group causalities, all these categories are identified based on the features of the subjects.

For every person in the frame a unique hypotheses is generated, and using continuous random fields the features of each hypotheses are collected and mapped with each other. An tracker is also assigned to them individually and the performance and energy of each trackers are extracted. Finally, with the energy level, the relationship among people is identified.

Table 4Comparative analysis of interaction recognition algorithms

Ref	Feature	Classifier used	Findings
[38]	Optical flow (OF	SVM	Able to find anomaly detections
[39]	Energy Potentials	Energy minimizatic algorithm	Able to identif pedestrian dynamics
[40]	Distance and velocity	SVM	Interaction between two and three persons identified
[41]	Distance	EM	Interpersonal social distance
[42]	Energy	CRF	Tracking amor groups
[43]	Energy	SVM	Detection of events
[44]	long-tern motion patterns	Graph convolutior network	I3D network with a tracking module

5.3 Dataset

Interaction detection datasets includes BEHAVE dataset [45]have data of interactions between multiple pedestrians, CMU MOCAP [46], Interaction between humans and television are available with TV Human Interactions Dataset [47] and UT Interaction Dataset [48]. Videos with focus on interactions like handshaking and discussing is available in LIRIS [49]. Videos taken in day to day activities are available in CAD60 [50]and CAD-120 [51]. All the stated datasets are annotated and contains videos which helps in detecting the interaction between humans.

VI.CONCLUSIONS

The intention of writing this extensive survey chapter is to present a understanding of how human behaviour is identified using deep learning algorithms. The major goal of the chapter is to make sure that the reader understands the steps involved in detection process, hence represented each step as separate chapter with the necessary algorithms applied in the process. This survey also gives the idea of different approaches and various features that are extracted from the video frames in each and every step and the possible research gaps are also mentioned, which helps the researchers to improve the human behaviour systems.

Dataset which plays a major role in every classification system and hence identifying the right dataset with annotation becomes more important to make accurate predictions. Hence we were to conscious in providing with the right databases which are applied in previous systems and the databases that are publicly available with annotation have been listed out in every stages of the detection process. The major drawback which was identified while doing the analysis process is, we could not identify a combined dataset which could be used for detection, analysis and interaction process and hence a public available data set having data of all the three will also be helpful for future researchers.We strongly

MILESTONE RESEARCH PUBLICATIONS

believe that we have provided the enough information that are required for future researchers to research and develop a more accurate and advance human behaviour prediction system.

REFERENCES

1. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.
2. Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., & Schiele, B. (2011, June). Learning people detection models from few training samples. In *CVPR 2011* (pp. 1473-1480). IEEE.
3. Lin, Z., & Davis, L. S. (2008, October). A pose-invariant descriptor for human detection and segmentation. In *European conference on computer vision* (pp. 423-436). Springer, Berlin, Heidelberg.
4. Leibe, B., Seemann, E., & Schiele, B. (2005, June). Pedestrian detection in crowded scenes. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 878-885). IEEE.
5. Johnsen, S., & Tews, A. (2009, May). Real-time object tracking and classification using a static camera. In *Proceedings of IEEE International Conference on Robotics and Automation, workshop on People Detection and Tracking* (p. 25).
6. Meng, Q., Li, B., & Holstein, H. (2006). Recognition of human periodic movements from unstructured information using a motion-based frequency domain approach. *Image and Vision Computing*, *24*(8), 795-809.
7. Liu, Y., Chen, X., Yao, H., Cui, X., Liu, C., & Gao, W. (2009). Contour-motion feature (CMF): A space–time approach for robust pedestrian detection. *Pattern Recognition Letters*, *30*(2), 148-156.
8. Walk, S., Majer, N., Schindler, K., & Schiele, B. (2010, June). New features and insights for pedestrian detection. In *2010 IEEE Computer society conference on computer vision and pattern recognition* (pp. 1030-1037). IEEE.
9. Xia, L., Chen, C. C., & Aggarwal, J. K. (2011, June). Human detection using depth information by kinect. In *CVPR 2011 workshops* (pp. 15-22). IEEE.
10. Davis, M., & Sahin, F. (2016, October). HOG feature human detection system. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 002878-002883). IEEE.
11. Sreedhar Kumar, S., Ahmed, S. T., & NishaBhai, V. B. Type of Supervised Text Classification System for Unstructured Text Comments using Probability Theory Technique. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(10).
12. Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017, July). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 3226-3229). IEEE.
13. Fei-Fei, L., Fergus, R., & Perona, P. (2004, June). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop* (pp. 178-178). IEEE.
14. Ess, A., Leibe, B., Schindler, K., & Van Gool, L. (2008, June). A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.
15. Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, *23*(3), 257-267.
16. Chen, D. Y., Shih, S. W., & Liao, H. Y. M. (2007, July). Human action recognition using 2-D spatio-temporal templates. In *2007 IEEE International Conference on Multimedia and Expo* (pp. 667-670). IEEE.
17. Ahmed, S., Guptha, N., Fathima, A., & Ashwini, S. (2021). Multi-View Feature Clustering Technique for Detection and Classification of Human Actions.
18. Li, W., Zhang, Z., & Liu, Z. (2010, June). Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 9-14). IEEE.
19. Sun, C., Junejo, I., & Foroosh, H. (2011, November). Action recognition using rank-1 approximation of joint self-similarity volume. In *2011 International Conference on Computer Vision* (pp. 1007-1012). IEEE.
20. Zhang, Z., Tan, T., & Huang, K. (2010). An extended grammar system for learning and recognizing complex visual events. *IEEE transactions on pattern analysis and machine intelligence*, *33*(2), 240-255.
21. Raptis, M., Kokkinos, I., & Soatto, S. (2012, June). Discovering discriminative action parts from mid-level video representations. In 2012 IEEE conference on computer vision and pattern recognition (pp. 1242-1249). IEEE.
22. Calderara, S., Heinemann, U., Prati, A., Cucchiara, R., & Tishby, N. (2011). Detecting

MILESTONE RESEARCH PUBLICATIONS