iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38571

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

A Literature Review on Pedestrian's Intent Detection

Kishanprasad Gunale[1], Advait Samant[2], Mansi Joshi[3], Sakshi Belure[4], Aniket Kulkarni[5]
[1, 2, 3, 4, 5]*Electronics and Communication Engineering, MIT World Peace University, Pune*

Abstract: ***The ability to predict the intent of the human beings walking on the street or crossing the road is one of the important safety features for intelligent cars. It has drawn the attention of automotive industry in the past few years. Estimating whether the pedestrian is going to cross the road or not is a very challenging task because there are multiple factors involved such as the speed at which the pedestrian is walking, whether he/she is walking alone or walking with someone, are they aware of the vehicle or not, etc. In this paper, we explore the different factors that affects the pedestrian's intent while walking or crossing the road. We also explore how different techniques are used to predict the intent of the pedestrian considering those factors. This study will be helpful for designing the system which predicts pedestrian's intent as good as the human ability to interpret the behaviour of the pedestrian.***

Keywords: Predict, Intent, Pedestrian, Behaviour, Interpret

I. INTRODUCTION

Pedestrians are involved in 22% of the worldwide 1.24 million deaths caused by traffic accidents every year. In Most of the cases, pedestrians are crossing a street at sunset and may be caused by poor visibility and driver's fatigue.[3] Many researchers have focused on the development of algorithms that estimate pedestrian's intent while walking on the streets or crossing the roads. However, the problem is still challenging, as the pedestrians may change their direction suddenly. Also, the fast speed of the car means that the decision making has a smaller time window available. [14][15]

The algorithms that are used for predicting the intent of the pedestrian mainly focuses on the parameters like pose estimation, trajectory estimation, head orientation, skeleton detection, speed estimation etc. [12] By using pose estimation and skeleton detection, the difference between the moving and standing person can be identified. These two methods are also useful for distinguishing the person sitting or riding motorcycle and the person who is walking on the street. Trajectory estimation and head orientation along with skeleton detection are generally used for identifying the direction of motion of the pedestrian. The fig 1 shows the flow of the intent detection algorithm

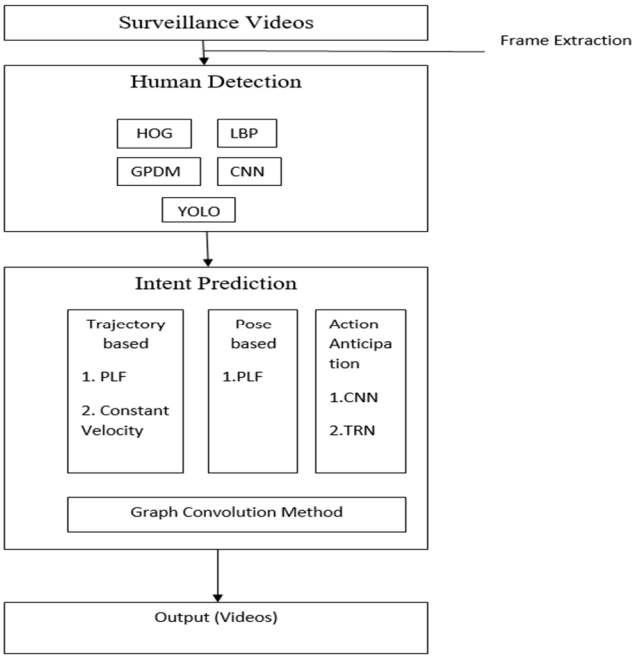


Fig. 1: Flow of the Intent Detection Algorithm

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

II. HUMAN DETECTION

Ideally, the system should not miss a single person present in the frame at that instant as it directly affects the accuracy of the whole algorithm. Also, the algorithm has one of the applications in the safety feature of the cars, it is not at all acceptable to miss even a single person present in the frame at any instance. Hence the human detection algorithm is very important for designing pedestrian's intent detection.[2]

A. HOG

The main application of the Histogram of oriented gradients (HOG) is the human detection process for feature extraction, while linear support vector machines (SVM) are used for human classification. A set of tests are conducted to find the classifiers which optimize recall in the detection of persons in visible video sequences. After this, the same classifiers are used to detect people in infrared video sequences and for obtaining results. Histogram of oriented gradients (HOG) consists of several steps that provide an array of image features representing the objects contained in an image in a schematic manner. The image features can then be used to detect the same objects in other images. The focus of the HOG descriptor is on the structure or the shape of an object. Another added advantage of HOG is that it can provide the edge direction as well. This is done by extracting the gradient and orientation (magnitude and direction) of the edges. Additionally, these orientations are calculated in 'localized' portions. The complete image is then broken down into smaller regions and for each region, the gradients and orientation are calculated. Histograms for these regions would be generated by HOG separately. The gradients and orientations of the pixel values are used for creating histograms.[16]

Fig. 2: HOG features superimposed on the original image.

B. LBP

Face detection, face recognition, facial expression recognition, pedestrian detection, to remote sensing and texture classification with the aim to build powerful visual object detection systems serve as various applications of LBP.

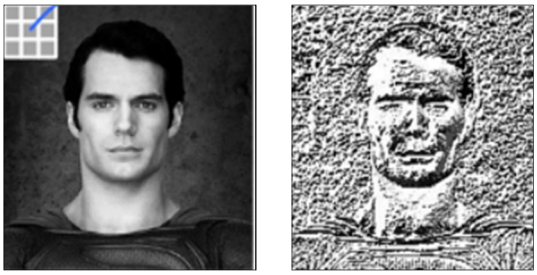


Fig. 3: Image assembled with Local Binary Patterns

LBP (local binary pattern) was originally proposed by Ojala et al. aiming at texture classification, and then extended for various fields, including face recognition, face detection, facial expression recognition etc. LBP is used for classification in computer vision. LBP when combined with the Histogram of oriented gradients (HOG) descriptor, the detection performance considerably on some datasets is improved and it also acts as one of the powerful features for text classification.[17]

©IJRASET: All Rights are Reserved

C. CNN

Fig. 4(a) & (b): Detection of Human Beings from the image using CNN.

As suggested in [5], a model can be trained for intelligent object detection using Convolutional Neural Network (CNN). They named it as ID-CNN i.e., Intelligent Detection using CNN. Using ID-CNN many different objects can be detected, detection of human beings from the frames of the video is one of them, as designed in [32].

ConvNet is a Deep Learning algorithm. It takes in an input image and has a role of assigning importance to various aspects/objects present in the image and hence can differentiate one from the other. In comparison to other classification Algorithms, the pre-processing required in a ConvNet is much lower.[18] The building blocks of CNNs are filters known as kernels. The relevant features from the input using the convolution operation are extracted using Kernels. The filters automatically learned by Convolution Neural Network without mentioning them explicitly. The use of these filters is to help in extracting the right and relevant features from the input data and to capture the spatial features from an image. Spatial features are defined as the arrangement of pixels and the relationship between them in an image. Spatial Features are useful in identifying the object accurately, the location of an object, and its relationship with other objects in an image.[5]

D. GPDM [4]

GPDM stands for Gaussian Process Dynamical Model. GPDM is the Machine Learning algorithm trained for detecting the Human being's path & pose. GPDM considers each body part while making the decision. In the 1st stage of the algorithm, body joints, head orientation, etc are considered for creating the skeleton of that image. After that, the skeleton is fed to the next stage where it detects whether the person is moving from left to right or right to left.[13] GPDM, the algorithm is trained for the stages such as walking, starting, stopping, and starting. [27][28]

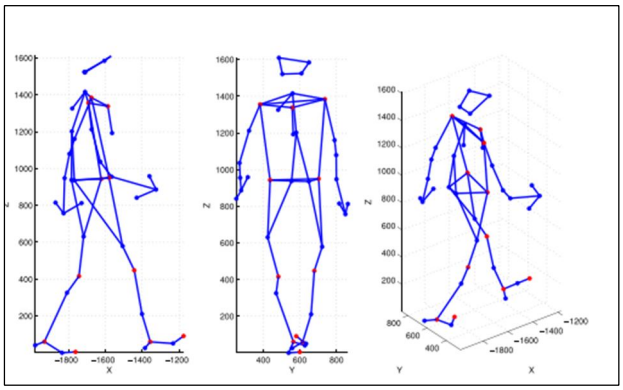


Fig. 5: Formation of Skeleton in GPDM

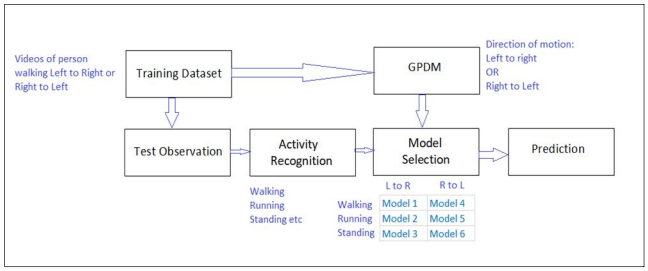


Fig.6: GPDM Model.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

E. YOLO [11]

YOLO is the algorithm built using the concept of [5]. YOLO is a multi-layer CNN model which can detect multiple objects at time. The objective of [2], which is the important part of our project, i.e., Action anticipation, can also be done using YOLOv3 by training the model using Darknet Master. The procedure of training is explained in [33].

YOLO is You Only Look Once is one of the real time object recognition algorithms. Using YOLO object detection, objects such as human beings, motor bikes, bicycles, cars, dogs, cats and many such objects can be detected. YOLO can be very much useful in Pedestrian's Intent Detection as it can detect many objects along with the detection of human beings. So, for example, some person is walking on the street with his pet dog, or some other person is about to ride his motorbike, then such scenarios can be considered while predicting the Intent of that human being. YOLO works directly on Videos, so the input to the algorithm is the video and the output that we get after detection will also be in the form of video. If the output is required in the form of frames, that is also possible by doing small modification in the algorithm. In Fig. 6 we can see person, cars, bicycle, handbag. Traffic signals, etc detected using YOLOv3.

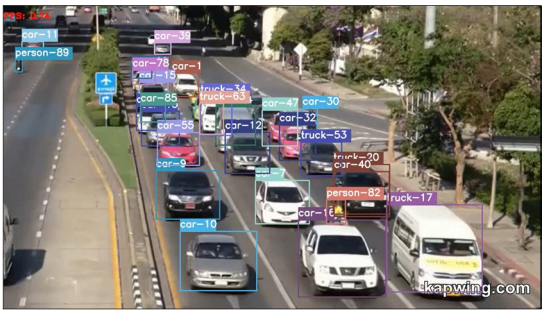


Fig. 6: Object Detection using YOLOv3.

III. INTENT DETECTION

For intent detection, there are different approaches used by researchers. Different features such as head orientation, pose estimation, path estimation etc are considered for training the algorithms. These algorithms have their superiority in different aspects such as complexity, ease of execution, accuracy etc. This study will help the designer choose the appropriate algorithm suitable for his application.

A. Trajectory Based

The trajectory-based algorithms depend on certain observations of the pedestrians. These are more inclined towards relying on the past motion and then predicting the location of the pedestrians in the future using contextual information such as 3D depths, road structure, head orientation and scene dynamics.[19]

1) *PLF [20]:* PLF stands for Pedestrian Locomotion Forecasting. PLF can be used to reason about pedestrian behaviour and path planning. The human dynamics of locomotion is nothing but the joint spatial movement of several key points on the human body. It is an outcome of the interaction between large scale trajectorial motion and finer body limb movements. This ability to predict human dynamics can assist in making proper decisions.

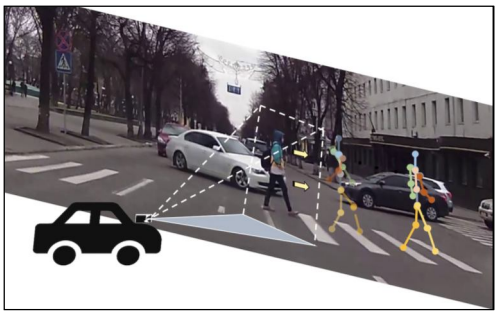


Fig.7: Pedestrian Locomotion Forecasting.

©IJRASET: All Rights are Reserved

2) *Constant Velocity:* This approach is based on the relative performance measure of the model with some baseline results. The baseline results are generated with two simple methods, pedestrian's constant velocity and constant acceleration for all the trajectories. In such cases, the predicted trajectory must follow the trajectory generated by keeping the last velocity constant throughout the detection.[22][23] It is difficult to always have a constant velocity if the pedestrian is in partial view. The motivation behind using this is if the velocity is increased too much, is it easy to compare with constant velocity according to the baseline result.

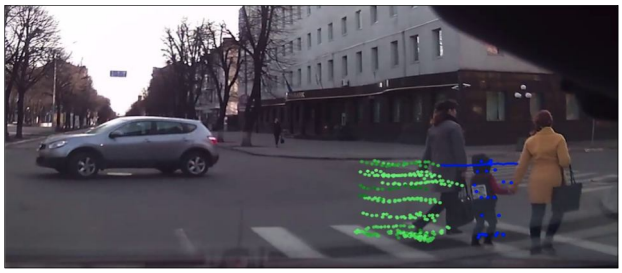


Fig.8: Tracing constant velocity of pedestrians

B. Pose Based [24]

Developing some computational methods for modelling human dynamics and forecasting how a pose might change later is a quite challenging task. In real scenarios, the pedestrians often obstruct each other and with the other objects in the scene. Moreover, obtaining full annotations of these human dynamics as well as the pose is an intensive task. Although, the Pedestrian Locomotion Forecasting method can provide a feasible solution to it.

1) *PLF [20]:* As previously mentioned, it is an egocentric setting for detection. PLF in pose-based intent detection includes generating frame-level supervision for human poses. The pose completion modules suppress the noise and fill missing joints of the human body. Later, these are split into global and local streams. The forecasting of human locomotion is feasible by combining the global and local stream along with the feature operations.

C. Action Anticipation

Action anticipation methods are relevant methodologies of intent understanding. These models learn to anticipate the next action by examining earlier actions i.e.: before it occurs. The future action is based on a combination of past visual inputs and past action recognition results. The following approaches discuss the topic more broadly.

1) *ID CNN [5]:* ID CNN stands for Intelligent Detection using Convolutional Neural Network. In these techniques there is no need to feed huge data every time, instead, the convolution operation is done only once per image and then features are extracted out of it. The network is much faster, improved, accurate, and optimized when compared with traditional CNN which is why it is called ID CNN. While using a multi-layer strategy along with introducing a contextual learning scheme the task of intent detection as well as prediction turns out to be easy.[32]
2) *TRN [21]*

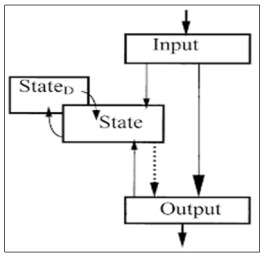


Fig.9: Flowchart of TRN

TRN stands for Temporal Recurrent Network. Sometimes, it is difficult to predict whether a pedestrian is running into the road or just walking toward the sidewalk. These actions can occur at any time, and we need to refine our inference, hypotheses every moment as we get evidence over time. With the TRN approach, a more discriminative representation of the frame can be done by jointly optimizing current and future action recognition. This incorporates the predicted future information to improve the performance of action anticipation in the present. TRN is nothing but explicitly predicting the future which helps to better classify actions in present.

©IJRASET: All Rights are Reserved

D. Graph Convolution Method [8]

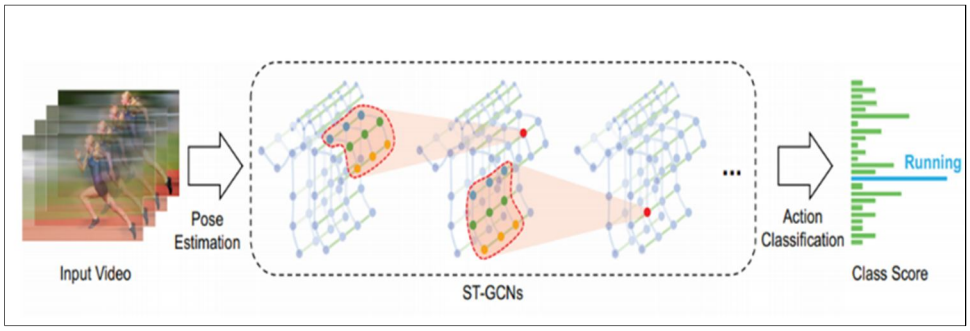


Fig. 10: Pose estimation on videos and construct spatial temporal graph on skeleton sequences. Multiple layers of spatial-temporal graph applied and gradually generate higher-level feature maps on the graph.

Graph convolution network is a type that work directly on graphs and take advantage of their structural information. Object recognition and image analysis can be viewed as problems structured along two-dimensional, so a graph can represent a general mathematical structure onto which numerous problems have been mapped and analysed. For pedestrian intent detection the human skeleton-based action recognition is useful.[31] Skeleton and joint trajectories of human bodies are robust to illumination change and scene variation for creating a GCN for intent detection there are two approaches 1) skeleton graph construction and 2) applying spatial graph convolutional neural network. A skeleton sequence is represented by 2D co-ordinates of each human joint in each frame, G= (V, E) on a skeleton sequence with N joints and T frames. Then applying graph CNN model on frames as in (Kipf and Welling 2017) [put reference no of the paper]: the intra body connections of joints are represented by adjacency matrix A and an identity matrix I representing self-connections. For a single frame GCN can be implemented with following formula:

$$fout = \Lambda^{-1/2} (A + I)\Lambda^{-1/2} finW,$$

where $\Lambda^{ii} = \Sigma\ (A^{ij} + I^{ij})$. Here the weight vectors of multiple output channels are stacked to form the weight matrix W. [10]

IV. DATASETS

A. JAAD [9]

JAAD stands for Joint Attention in Autonomous Driving. This dataset was designed in the year 2017 for studying the behaviour of traffic elements. The dataset consists of 346 video clips of 5 to 15 seconds each. In those clips, there are one or more human beings either walking by the road or crossing the road. For creation of this database, only one camera is used. All the clips are recorded from the moving vehicle (front view) using once camera. [29][30]

Table 1: Variations available in JAAD dataset

Categorial Variable	Value
Time of Day	Day/Night
Weather	Clear/Snow/Rain/Cloudy
Location	Street/Indoor/Parking Lot
Designated Crossing	Yes/No
Age Gender	Child/Young/Adult/Senior

The dataset has its most of the videos recorded form the urban areas, only few clips are from the rural areas. The dataset has videos from Canada, Ukraine, Germany, and USA.

©IJRASET: All Rights are Reserved

B. STIP [1]

STIP stands for Stanford TRI Intent Prediction. The dataset is developed by Stanford University and Toyota Research Institute collaboratively. The dataset consists of videos recorded from the moving vehicle. The total length of this dataset is 923 minutes. The videos in this dataset are recorded using 3 cameras, one placed on the centre of the vehicle and remaining 2 on the left and right sides of the vehicle.[25][26]

Fig. 11: Camera View of STIP dataset using 3 cams.

Table 2: Features available in JAAD and STIP datasets.

Features	JAAD	STIP
Pedestrian Crossing the road or not	Yes	Yes
Pedestrian's Intent of crossing the road (Prediction)	No	Yes
Pedestrian's Location on frame	No	Partially
Trajectory/Direction of motion of the Pedestrian	No	Yes

Table 3: Comparison between JAAD and STIP datasets.

Parameters	JAAD	STIP
Year	2017	2020
Length (Minutes)	46	923
No. of Frames	82,000	11,08,176
No. of Pedestrians	3,37,000	35,00,000
No. of Cameras used	1	3

©IJRASET: All Rights are Reserved

V. CONCLUSION

In this paper, all the different approaches and aspects of Pedestrian's Intent Detection have been covered. Pedestrian's Intent Detection is a new concept and is yet to be designed as an application. So, it is important for a developer to go through all the research work done on this topic and then design the application accordingly by selecting the appropriate method suitable for their application. These are some of our findings after studying this Pedestrian's Intent Detection in detail:

1) The Pedestrian's Intent Detection is a Deep Learning algorithm used to predict the intent of the pedestrians.

2) The functioning of this algorithm involves 2 stages that are:

a) Object Detection (Human Beings and other objects such as bicycle, pet dog, etc which will be helpful for detecting their Intent.)

b) Intent Prediction (Based on their trajectory, speed, objects associated with them, etc)

3) CNN gives you a very good accuracy as compared to HOG or LBP when it comes to object detection.

4) YOLO is again a CNN based algorithm which can provide the object detection output with maximum accuracy, but the limitation of YOLO is its bulkiness and processing time which is slightly higher than other methods.

5) For the Intent Prediction part multiple approaches were used by different researchers. Few of them are Head pose; Trajectory; Speed; etc. Each one of them has their own pros and cons. The good thing though is that they can be used together for providing more accurate predictions. For example, the prediction done by using trajectory and head pose together will be more accurate than the prediction done by using only trajectory or only head pose.

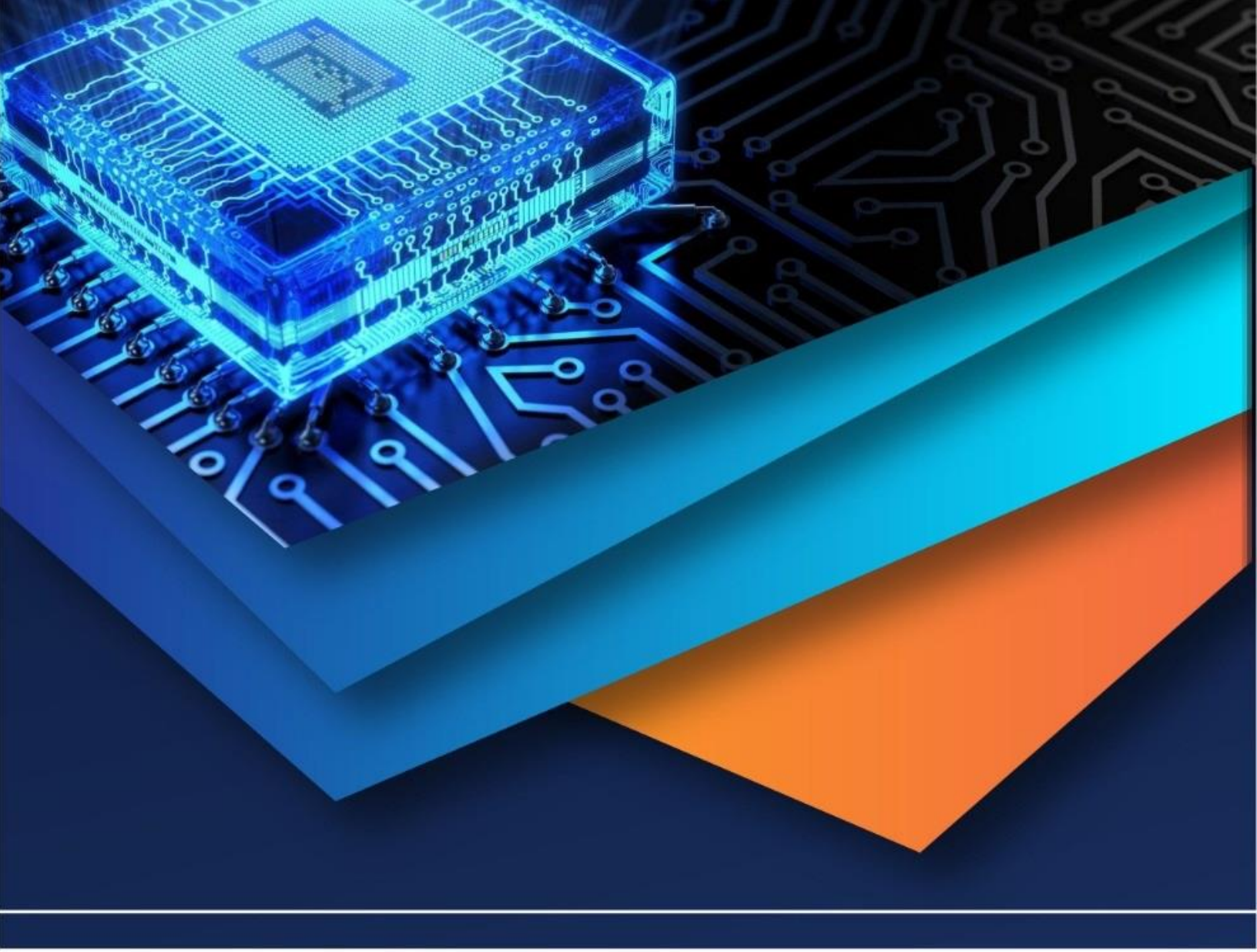
6) STIP and JAAD datasets are available with us which has videos recorded from a vehicle. The recorded videos contain pedestrians walking on the streets and crossing the road along with other objects such as other vehicles, bicycles, etc.

7) STIP is the bigger dataset than JAAD and it also makes use of 3 cameras (JAAD uses only one camera) for recording 3 different views (Left-side, centre and right-side view) from the moving vehicle.

REFERENCES

[1] Bingbin Liu, Ehsan Adeli, +5, "Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction "
[2] Dimitrios Varytimidis, Fernando Alonso-Fernandez, Cristofer Englund, "Action and intention recognition of pedestrians in urban traffic"
[3] Daniela Ridel, Eike Rehder, +3, "A Literature Review on the Prediction of Pedestrian Behaviour in Urban Scenarios"
[4] Raúl Quintero Mínguez, Ignacio Parra Alonso, +2, "Pedestrian Path, Pose, and Intention Prediction Through Gaussian Process Dynamical Models and Pedestrian Activity Recognition".
[5] Amish Jahangir Kapoor, Hong Fan and Muhammad Sohail Sardar, "Intelligent Detection Using Convolutional Neural Network (ID-CNN)"
[6] D. M. Gavrila, J. Giebel, and S. Munder. "Vision-based pedestrian detection: the protector+ system." Proc. of the IEEE Intelligent Vehicles Symposium, Parma, Italy, 2004.
[7] K. Mikolajczyk and C. Schmid. "A performance evaluation of local descriptors." PAMI, 2004. Accepted.
[8] Sijie Yan, Yuanjun Xiong, Dahua Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition"
[9] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," arXiv preprint arXiv:1609.04741, 2016.
[10] Denis Tome, Chris Russell, Lourdes Agapito, "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image."
[11] Joseph Redmon, "YOLO: Real-Time Object Detection"
[12] Sarfraz Ahmed, M. Nazmul Huda, +4, "Pedestrian and Cyclist Detection and Intent Estimation for Autonomous Vehicles: A Survey"
[13] Raquel Urtasun, David J. Fleet, and Pascal Fua, "3D People Tracking with Gaussian Process Dynamical Models."
[14] Frederik Diederichs Tobias Schüttke, and Dieter Spath, "Driver Intention Algorithm for Pedestrian Protection and Automated Emergency Braking Systems."
[15] Michael Goldhammer, Sebastian Köhler, +4, "Intentions of Vulnerable Road Users—Detection and Forecasting by Means of Machine Learning."
[16] Navneet Dalal, and Bill Triggs, "Histograms of Oriented Gradients for Human Detection."
[17] Ke-Chen Song, Yun-Hui YAN, +2, "Research and Perspective on Local Binary Pattern."
[18] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, "Understanding of a convolutional neural network."
[19] Ganglei He; Xin Li; +3, "Probabilistic intention prediction and trajectory generation based on dynamic bayesian networks."
[20] Karttikeya Mangalam, Ehsan Adeli, +3 "Disentangling Human Dynamics for Pedestrian Locomotion Forecasting with Noisy Supervision."
[21] Mingze Xu1, Mingfei Gao, +3, "Temporal Recurrent Networks for Online Action Detection."
[22] Stephanie Lefevre, Dizan Vasquez, and Christian Laugier," A survey on motion prediction and risk assessment for intelligent vehicles."
[23] Ivar Salomonson, and Karthik Murali Madhavan Ratha, "Mixed Driver Intention Estimation and Path Prediction Using Vehicle Motion and Road Structure Information."
[24] Chunyu Wang, Yizhou Wang, and Alan L. Yuille, "An approach to pose-based action recognition."
[25] D. Geronimo and A. M. L ´ opez, "Vision-based pedestrian protection systems for intelligent vehicles."
[26] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection."
[27] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric."
[28] D. Vazquez, A. M. Lopez, J. Marin, D. Ponsa, and D. Geronimo, "Virtual and real-world adaptation for pedestrian detection."
[29] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction."
[30] N. Japuria, G. Habibi, and J. P. How, "CASNSC: A context-based approach for accurate pedestrian motion prediction at intersections."
[31] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs."
[32] Ben-Yue Su, Jie Wang, +4, "A CNN-based Method for Intent Recognition Using Inertial Measurement Units and Intelligent Lower Limb Prosthesis."
[33] Joseph Redmon,"YOLO: Real-Time Object Detection."

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)