Head pose feature for prediction pedestrian intention to crossing the road using LSTM

1st Hanugra Aulia Sidharta
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hanugra.207022@mhs.its.ac.id

2nd Muhammad Ilham Perdana
Department of Informatics
Universitas Muhammadiyah Malang
Malang, Indonesia
ilhamperdana@umm.ac.id

3rd Eko Mulyanto Yuniarno
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

4th Berlian A1 Kindhi
Department of Electrical Automation Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
berlian@its.ac.id

5th Mauridhi Hery Purnomo
Department of Electrical Engineering
Department of Computer Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

***Abstract*—Understanding pedestrian behaviour when crossing the road is an important key to the development of autonomous vehicles. Because pedestrians are considered Vulnerable Road Users (VRUs), they are likely to be killed if they are involved in an accident. To ensure their safety, it is then necessary to predict the pedestrian's intention based on their behaviour. In this experiment, we propose head pose observation for predicting their intention, by observing pedestrians' head pose data, we can then predict their intention to cross the road. To achieve this purpose, we use human head detector and head pose extraction feature, and the resulting yaw, pitch and roll as three head pose features. To select the most optimal feature is important for predicting pedestrian intention, then we make 7 combination scenarios based on these three features and compare it with the same model. Based on this scenario, it is proved that all these three data are optimal to observe pedestrian intention. There are three behavioural annotation that have been used, there are crossing, not crossing and will crossing. We derive will crossing from the annotation looking and not crossing while waiting at the roadside. Prediction of pedestrian behaviour is done by using LSTM model, and the resulting precision on crossing and not crossing with 0.98, while will crossing is 0.94.**

***Index Terms*—Head pose features, predicting pedestrian intention, pedestrian behavior, feature selection, LSTM.**

I. Introduction

Pedestrian safety is one of important factor in autonomous vehicle development, due to pedestrian is unprotected by any safety feature, which makes pedestrian has high risk if they are involved in accident. Pedestrians are therefore classified as Vulnerable Road Users (VRU), along with cyclists and horse riders. This risk increases when pedestrians share the road facility with another road user, such as when they are crossing the road. Although pedestrian safety facilities are being built to reduce this problem, according to Insurance Institute for Highway Safety (IIHS) data there will be 7,388 cases in 2022, increasing by 13% from 2021 [1].

There are many approaches that aim to ensure pedestrian safety, from building more sophisticated sensing and perception infrastructure system [2], introducing new safety standard system [3], esigning system that can interact between vehicle, pedestrian and surrounding infrastructure [4], or by equipping vehicle with additional sensor. here are several devices that can be used to assist in reading the surrounding vehicles, such as LIDAR [5] [6], monocular camera [7] [8], and stereo camera [9] [10]. In our experiment we were focusing developingto predict pedestrian behavior using a monocular data.

Predicting pedestrian behaviour when crossing the road is a challenging task because each pedestrian has their own unique behaviour, resulting in a large variance of data. To overcome this problem, then prediction of pedestrian behaviour can be done by observing pedestrian joint movement, their pose pattern, trajectory movement and head pose. Joint movement, pose pattern and trajectory movement features are useful for predicting pedestrian walking direction, while head pose observation is useful for predicting pedestrian intention. Ahad proposes the observation by observing specific human joints, then perform the prediction by using machine learning with angular data from these joints [11]. Ding uses a combination of angular movement and distance measurement of human joint to construct human pose [12]. Cao observes pedestrian from human posture perspective by using skeleton orientation data [13]. Meanwhile, Yan is predicting pedestrian behaviour from their trajectory by optimizing global temporal attention [14].

Pedestrian head pose can be obtained by detecting the head of each pedestrian then extracting its feature. Vu build a detector based local, global and pairwise model with CNN framework that create context aware data [15]. Chi develops human head detector that works in parallel with human detection as a single stage detector [16]. While Chen builds a light-head detector by using a single-stage detector as an efficient object detector [17]. There are three features that can be extracted from the pedestrian, yaw, pitch and roll. These features are useful for observing the pedestrian's attention while crossing the road. The head pose feature can be obtained by using 2D face landmark annotation [18], and Ruiz to refine it by removing key point landmarks and using multi-loss neural networks [19].

The author would like to thank the Ministry of Finance of the Republic of Indonesia for the Beasiswa Pendidik 2020 Programme with LPDP Scholarship.

In this experiment, we propose to perform prediction of pedestrians intention to cross the road by using deep learning approach, due to pedestrians behaviour is very unique, thus generating data with large variance. To solve this problem, we propose to understand pedestrian behaviour through their head pose features (raw, pitch and roll). To choose the most optimal feature, we perform feature selection with 7 scenarios. While the prediction of pedestrian intention while crossing the road is done using LSTM architecture. Our approach is superior to the experiment conducted by Perdana [21] on the same dataset, which uses a distance-based KNN approach with a precision score of 0.97, while our approach is superior with a precision score of 0.98.

II. Proposed Method

Fig. 1 is a block diagram of our proposed method, which can be divided into four subsections. In this experiment, we use a monocular dataset consisting of video and pedestrian behaviour annotation. This video is divided into single frames, then on each frame we detect the head of each pedestrian, each detected head is marked with bounding box. This process is continued with head pose extraction to get yaw, pitch and roll as head pose feature. LSTM is used as a deep learning approach to predict the pedestrian's intention to cross the road.

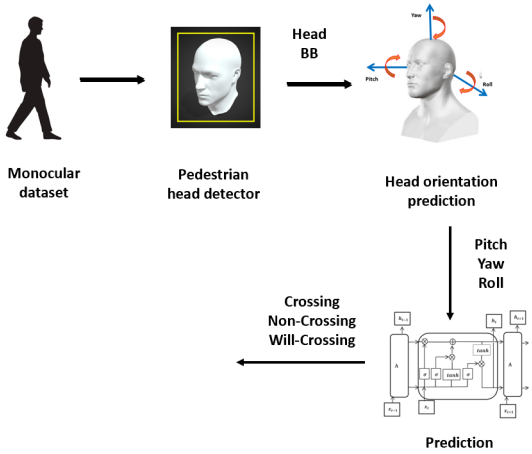

Fig. 1. Block diagram of proposed method.

A. Monocular dataset

There are several public datasets that can be used to observe pedestrian behaviour while crossing the road, recorded using a camera. In this experiment, we use a Joint Attention in Autonomous Driving (JAAD) dataset that was recorded using a single camera or monocular data [20]. This dataset is recorded using a camera placed on the windscreen of a car, this means to provide a natural visual perspective from the driver's point of view. This dataset was built to study behaviour patterns between pedestrians and car drivers for autonomous driving context. JAAD dataset is suitable for our purpose because it is equipped with behaviour annotation for each pedestrian. These behavioral annotations are very useful to understand pedestrians' behavior while crossing the road.

In this experiment we narrow down the observation for a single pedestrian involved in each video, our reason is due to this framework is on early stage of development. JAAD dataset is contain 346 video with 686 pedestrian has been annotated for their behaviour, while on each video is contain from a single pedestrian up to 6 pedestrian. Thus, we sample a total of 28 videos that have only a single pedestrian, this video then extracted into 4054 frames.

B. Pedestrian head detector

There are two approaches that are commonly used to create an object detector in deep learning domain, which is single stage detector and two stage detector. Two stage object detector is classified as multi stage object detector that construct by generating a region proposal, then continued by extracting features from each proposed region. Meanwhile, single stage object detector is constructed by skipping the generation of a region proposal as on two stage object detector. However, this model detects directly through a dense sampling of particular location. Thus resulting single stage object detector is superior in inference speed, while two stage detector is excellent for its high localisation also better accuracy.

In this experiment, we were make use of YOLO that utilizes concept of single stage dectector. YOLO is well known superior to detect various of object, from person, animal, transportation vehicle and daily home appliance. YOLO having capabilities to detect many object on single frame by dividing image into to SxS grid of region, then predict object bounding box on each region by probability weight. In this experiment we use YOLOv3 for detecting head of pedestrian on each provided frame. To achieve this goal, then the detector in this model is pre-trained by using two datasets, Crowd Human dataset[5] and Hollywood Head (HH) dataset[6].

C. Head pose estimation

The head pose of the human can be obtained by measuring the inclination of the human skull to a specific axis, this measurement can be divided into three different orientations, such as yaw, pitch and roll. Yaw is related to head rotation measured from the vertical axis, while pitch is related to head rotation measured from the transverse axis, and roll is related to head rotation measured from the sagittal axis. By combining these three features then we can obtain euler rotation angle composed with α, β, and γ. With this three data then we can then observe the pedestrian's head movements when they are turning right and left, tilting up and down, and also when they are swinging left and right. Human head pose observation movement can be seen in Fig. 2.

In this research head pose estimation is obtain by utilizing WHENet model architecture. This model is inspired from multi-loss framework develop by Ruiz[7] then refined it through wrapped loss, in this work Ruiz is estimating head pose from three different loses, and each loss model is specific for each angle (yaw, pitch, and roll). WHENet model has been pre-trained with several dataset, such as AFLW2000, BIWI, 300W-LP combined with CMU Panoptic Dataset. WHENet can estimate head pose with range value of yaw from -180 to 180, while pitch and roll range value is -99 t0 99.

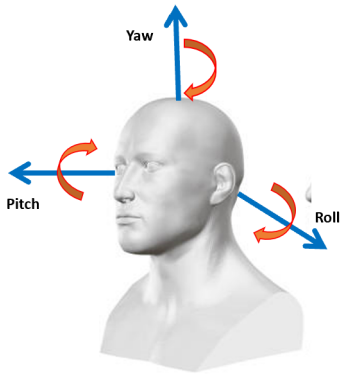

Fig. 2. Yaw, Pitch and Roll as extracted head pose feature.

D. Pedestrian intention prediction

There are yaw, pitch and roll as three features of head pose that have been used for predicting pedestrian intention to cross the road, this feature is paired with three pedestrian annotations. Based on the above data, then in this experiment, we propose to use deep learning approach for predicting pedestrian's intention to cross the road. We propose LSTM approaches for this purpose, because LSTM is suitable for predicting a serial data as our pedestrian feature. Then data segmentation is used to give LSTM more comprehensive data. This is obtained by using 10 frames to represent 0.6 second observation, that creating data with dimension 10 x 3 as input. Our LSTM model was constructed with two hidden layer that consisted of 64 and 32 unit, combine with softmax as their activation function, this model was optimized using adam optimizer.

III. Result and Discussion

In this experiment, we propose to predict pedestrian intention to cross the road by using head pose features. Our head pose features are composed of pitch (P), yaw (Y) and roll (R) which can be used for predicting pedestrian intention to cross the road. Fig. 3 is illustrate process on pedestrian when crossing the road, we are overlay all image with bounding box as result head detection, also three head pose feature as our proposed method. In Fig. 3 we can see pedestrian head pattern when crossing the road, in the left side pedestrian is walking forward while watching in her front, while the middle figure is captured when pedestrian is tilted her head toward her right side, and followed by looking traffic on her right side.

Fig. 3. Overlay pedestrian image with their head pose feature when crossing the road.

There are two annotations that can be used to understand the pedestrian's decision while crossing the road, the crossing annotation, which indicates that the pedestrian has decided to cross the road, and the not-crossing annotation as the opposite, by walking in the roadside or standing on the roadside. To enrich the observation, we then add another annotation to support the observation of the pedestrian's intention, we add the observation in the looking phase. We derive looking annotations with not-crossing annotations to create another annotation, which we call will-crossing annotation. The reason is that based on our visual observation, we found that pedestrian is waiting and standing waiting the road is clear was labelled as not-crossing annotation, while pedestrian in this situation is labelled as looking. Therefore, in this experiment we use three annotations, such as crossing (C), not crossing (NC) and will crossing (WC).

To provide independent data and proposed model can learn generally without reaching overfit, then we use data shuffle. Then our proposed model is trained along 50 epochs, and organise number of samples to be processed by 32 of batch size. In this experiment we create learning scenarios based on the number of data and feature used. There are 7 scenarios with different feature combination, and each combination will train on same model. This scenario was created as selecting the most important head pose feature for predicting pedestrian intention. The result of these scenarios can be seen in the TABLE I There are four model performance scores (accuracy, loss, validation accuracy and validation loss). To provide a concise comparison, we use variance, standard deviation and mean as a statistical approach. Based on the data in TABLE I we can see that a single feature is lacking of performance, while the use of two features is achieve better than one feature. However, their performance is poor compared to the use of three features, with only reaching a maximum accuracy of 0.9233 with the combination of pitch and roll. Thus with these data we can conclude that all features are important for predicting pedestrian intention.

TABLE I
SELECTION MOST OPTIMAL FEATURE FROM SEVEN SCENARIOS

Num of feature		**1**	**1**	**1**	**2**	**2**	**2**	**3**
Feature		**P**	**R**	**Y**	**PR**	**YP**	**YR**	**YPR**
acc	Var	0.02	0.02	0.01	0.01	0.01	0.01	0.01
	SD	0.15	0.13	0.08	0.11	0.10	0.09	0.08
	Mean	0.73	0.77	0.75	0.92	0.89	0.87	0.95
loss	Var	0.08	0.06	0.03	0.06	0.05	0.03	0.03
	SD	0.29	0.24	0.16	0.24	0.22	0.19	0.18
	Mean	0.58	0.52	0.60	0.19	0.27	0.30	0.13
val acc	Var	0.01	0.01	0.00	0.01	0.01	0.00	0.00
	SD	0.09	0.08	0.06	0.09	0.08	0.07	0.07
	Mean	0.65	0.68	0.70	0.88	0.84	0.86	0.93
val loss	Var	0.01	0.01	0.01	0.03	0.03	0.02	0.02
	SD	0.10	0.08	0.10	0.18	0.18	0.14	0.15
	Mean	0.83	0.73	0.71	0.32	0.39	0.34	0.19

Fig.4 shows the detailed performance of the model at all epochs, with the accuracy score at the top of the image and the loss score at the bottom. In this figure we can see that model is performing well data learning, and model is reaching stable performance since 20 epoch. And based on this figure we can conclude that model is not overfitting, also training and validation performance is good.

We use precision, recall and f1-score to measure the prediction performance of this model. From TABLE II we saw that precision of crossing and not crossing is reach 0.98

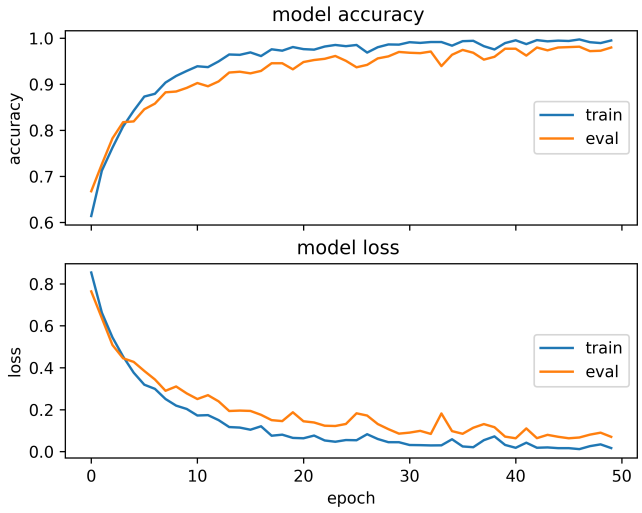

Fig. 4. Proposed LSTM Model accuracy and loss performance .

while will crossing is reach 0.94. Precision of will crossing is lower compared with other class, this is possible due to will crossing is actually derivation from combination looking and not crossing. While recall is behave opposite, with best performance reach by will crossing, this gained by classifier that can predict many positive sample. The F1-score of all classes reach the same value, around 0.97. All these data are supported by confusion matric in TABLE III.

TABLE II
PREDICTION METRIC OF PROPOSED MODEL

Class	precision	recall	f1-score
C	0.98	0.96	0.97
NC	0.98	0.95	0.96
WC	0.94	0.99	0.97

TABLE III
CONFUSION MATRIX OF PREDICTION PEDESTRIAN INTENTION

Class		T	F
C	**T**	557	10
	F	23	551
NC	**T**	918	4
	F	12	204
WC	**T**	770	23
	F	2	346

IV. CONCLUSION

This study explores head pose features for predicting pedestrian intention when crossing the road. We use a combination of three head pose features such as pitch, yaw and roll to observe pedestrian behaviour. The head pose feature is obtained by using YOLO as a human head detector, and the WHENet model is used to obtain these three features. In this experiment, pedestrian image and pedestrian behaviour are provided by JAAD dataset. We derive two pedestrian behaviour, which is looking and not crossing as will cross, that useful to observe pedestrian that will cross, then this data is learning by a LSTM model. We perform feature selection by testing combination of these features, and result that predicting pedestrian intention was required all head pose features. Our constructed LSTM model performance can be measured by its prediction evaluation, with precision on C and NC is 0.98, while WC is 0.94.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Finance of the Republic of Indonesia for the Beasiswa Pendidik 2020 Programme with LPDP Scholarship.

REFERENCES

[1] Insurance Institute for Highway Safety, Highway Loss Data Institute, 'Fatality Facts 2021 Pedestrians.' https://www.iihs.org/topics/fatality-statistics/detail/pedestrians:text=Posted May 2023.-,Trends,their lowest point in 2009.
[2] G. Dimitrakopoulos, L. Uden, and I. Varlamis, The Future of Intelligent Transport Systems. Elsevier, 2020.
[3] B. Williams, Intelligent Transport Systems Standards. 2008
[4] P. Sewalkar and J. Seitz, "Vehicle-to-pedestrian communication for vulnerable road users: Survey, design considerations, and challenges," Sensors (Switzerland), vol. 19, no. 2, 2019, doi: 10.3390/s19020358.
[5] M. G. Ocando and N. Certad, "Autonomous 3D mapping of an enviroment , while simultaneously making 2D SLAM , using a single 2D LIDAR and ROS *," pp. 2–7, 2017, doi: 10.1109/SBR-LARS-R.2017.8215333.
[6] J. E. Naranjo, J. G. Zato, and C. De Valencia, "Analysis of Lidar Sensors for New Adas Applications . Usability in Moving Obstacles," Test, no. January, 2009.
[7] H. Ci, X. Ma, C. Wang, and Y. Wang, "Locally Connected Network for Monocular 3D Human Pose Estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 14, no. 8, pp. 1–1, 2020, doi: 10.1109/tpami.2020.3019139.
[8] X. Ji, Q. Fang, J. Dong, Q. Shuai, W. Jiang, and X. Zhou, "A survey on monocular 3D human pose estimation," Virtual Real. Intell. Hardw., vol. 2, no. 6, pp. 471–500, 2020, doi: 10.1016/j.vrih.2020.04.005.
[9] N. G. S. Sai Srinath, A. Z. Joseph, S. Umamaheswaran, C. L. Priyanka, M. Malavika Nair, and P. Sankaran, "NITCAD - Developing an object detection, classification and stereo vision dataset for autonomous navigation in Indian roads," Procedia Comput. Sci., vol. 171, no. 2019, pp. 207–216, 2020, doi: 10.1016/j.procs.2020.04.022.
[10] F. Camara et al., "Pedestrian models for autonomous driving part ii: High level models of human behaviour," IEEE Trans. Intell. Transp. Syst., 2020, doi: 10.1109/tits.2020.3006767.
[11] M. A. R. Ahad, M. Ahmed, A. Das Antar, Y. Makihara, and Y. Yagi, "Action recognition using kinematics posture feature on 3D skeleton joint locations," Pattern Recognit. Lett., vol. 145, pp. 216–224, 2021, doi: 10.1016/j.patrec.2021.02.013.
[12] W. Ding, B. Hu, H. Liu, X. Wang, and X. Huang, "Human posture recognition based on multiple features and rule learning," Int. J. Mach. Learn. Cybern., vol. 11, no. 11, pp. 2529–2540, 2020, doi: 10.1007/s13042-020-01138-y.
[13] W. Cao, Y. Lu, and Z. He, "Geometric Algebra Representation and Ensemble Action Classification Method for 3D Skeleton Orientation Data," IEEE Access, vol. 7, pp. 132049–132056, 2019, doi: 10.1109/ACCESS.2019.2940291.
[14] X. Yan, X. Zhong, Z. Yang, R. Zhang, W. Huang, and Z. Wang, "Global Temporal Attention Optimization for Human Trajectory Prediction," Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern., vol. 2022-Octob, pp. 184–189, 2022, doi: 10.1109/SMC53654.2022.9945387.
[15] T. H. Vu, A. Osokin, and I. Laptev, "Context-aware CNNs for person head detection," Proc. IEEE Int. Conf. Comput. Vis., vol. 2015 Inter, pp. 2893–2901, 2015, doi: 10.1109/ICCV.2015.331.
[16] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational learning for joint head and human detection," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 10647–10654, 2020, doi: 10.1609/aaai.v34i07.6691.
[17] C. Chen, M. Liu, X. Meng, W. Xiao, and Q. Ju, "RefineDetLite: A lightweight one-stage object detection framework for CPU-only devices," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2020-June, pp. 2997–3007, 2020, doi: 10.1109/CVPRW50498.2020.00358.
[18] A. Bulat and G. Tzimiropoulos, "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)," Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-October, pp. 1021–1030, 2017, doi: 10.1109/ICCV.2017.116.

[19] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., vol. 2018-June, pp. 2155–2164, 2018, doi: 10.1109/CVPRW.2018.00281.
[20] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior,” Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017, vol. 2018-Janua, pp. 206–213, 2017, doi: 10.1109/ICCVW.2017.33.
[21] M. I. Perdana, W. Anggraeni, H. A. Sidharta, E. M. Yuniarno, and M. H. Purnomo, “Early Warning Pedestrian Crossing Intention from Its Head Gesture using Head Pose Estimation,” Proc. - 2021 Int. Semin. Intell. Technol. Its Appl. Intell. Syst. New Norm. Era, ISITIA 2021, pp. 402–407, 2021, doi: 10.1109/ISITIA52817.2021.9502231.