AUTO-G: GESTURE RECOGNITION IN THE CROWD FOR AUTONOMOUS VEHICLES

Pavani Tripathi, Rohit Keshari, Soumyadeep Ghosh, Mayank Vatsa, and Richa Singh

IIIT-Delhi, India

ABSTRACT

Autonomous driving is an active area of research. An important aspect of this problem is recognizing the gestures made by humans, both inside and outside the vehicle. In this paper, we present the *Auto-G* database that comprises different hand gestures for autonomous driving. The database encompasses several challenges such as occlusion, low resolution, motion blur, illumination variation, extreme pose variations, along with the presence of multiple gestures within a frame. We also propose an end-to-end pipeline for hand detection and gesture recognition. The proposed pipeline achieves a frame gesture recognition accuracy of 90.23% on the proposed Auto-G database.

***Index Terms*—** Autonomous Driving, Gesture Recognition, In-the-crowd.

1. INTRODUCTION

Autonomous driving [1] has been a topic of interest for both the automobile industry and the academic community for the last two decades. Since human error is the cause of 93% of road accidents [2], autonomous driving is expected to make roads safer worldwide. Recent technological progress in machine learning and human-computer interaction have enabled the introduction of hand gestures for vehicular control [2]. In order to make autonomous driving more convenient and user-driven, some studies [3, 4, 5] have been carried out for incorporating hand gestures in the autonomous driving framework by using gesture-based commands such as pointing to a parking lot, or taking a particular highway exit while the vehicle is in motion. To receive and understand gesture-based inputs from users sitting in a car (or on the road as shown in Fig. 1), the gesture detection and recognition system should be robust to unconstrained scenarios.

Although extensive research has been performed on recognizing human gestures in a constrained environment [6], performing the same for autonomous driving is a challenging task. Hand gestures for autonomous driving may encompass several challenges. As seen in Fig. 1, hand gestures in the crowd may be occluded by other people/objects in the foreground. This may result in a different semantic mapping of the occluded gesture, which leads to incorrect classification. The background may contain a person's face or a body part which can make hand segmentation challenging. Due to the rapid movement of body parts and/or the camera, motion blur is also introduced into the images. In addition to that, heavy pose and illumination variations are also introduced in such an unconstrained scenario. These variations make both detection and recognition for such hand gesture images a challenging problem.

Fig. 1. Illustrating various challenges, viz., occlusion, low resolution, pose variation, varied background, and illumination variation among a sample gesture present in the proposed database.

Manawadu *et al.* [2] have proposed an interface for driver to control the vehicle's lateral and longitudinal motions. Ohn-Bar *et al.* [7] have employed RGBD camera to capture gestures having a uniform background and single gesture in one sample. They have utilized hand-crafted features for classification of gestures. Ohn-Bar *et al.* [8] next studied the behavior of human surrounding vehicles. They have studied human agent interaction when humans are: 1) inside the vehicle cabin, 2) around the vehicle, and 3) inside surrounding vehicles. However, the databases available for gesture in autonomous vehicles have been captured in a constrained environment.

In order to promote and aid research in gesture detection and recognition for autonomous vehicles, this paper presents a novel database of human hand gestures which have been captured in unconstrained environments depicting an internal configuration of a car. To the best of our knowledge, this paper contains the first crowd based gesture recognition database and also an end-to-end pipeline for the same. We also perform extensive evaluations for both hand detection and gesture recognition using existing methods [9, 10]. These

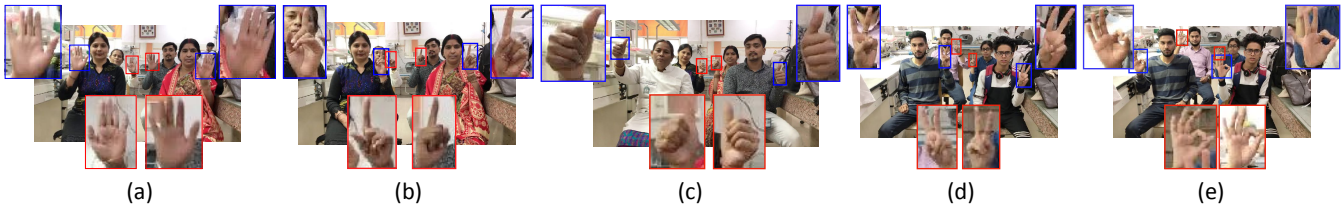

Fig. 2. Sample images from the proposed database. Images with blue border refer to the gestures shown by the subjects sitting in the front and images with red border represent gestures shown by the subjects sitting at the back. Each set from (a)-(e) presents a different gesture.

Fig. 3. Illustrates the various challenges: occlusion, motion blur and varied resolution, pose and illumination among the gestures (a)-(e) present in the proposed database.

evaluations reveal the challenging nature of the problem and help to develop useful insights into the problem of hand detection and gesture recognition.

2. AUTO-G: GESTURE IN CROWD DATABASE

Several databases have been proposed for gesture recognition [11, 12, 13], however, they have been acquired in a constrained environment containing gestures exhibited by a single subject. In this paper, a novel video database, Auto-G, for gesture recognition in autonomous driving vehicles is proposed. The database contains videos of subjects showing 5 different gestures. The videos have been created by randomly selecting 4 individuals from a group of 60 subjects and arranging them in a car-like seating to simulate the interior of a vehicle. In total, there are 60 videos (126,179 frames) in the proposed database. Further, bounding-box annotations have been manually marked using the Vatic tool [14] for $14,751$ video frames. A video can contain a maximum of 8 hands since a total of 4 subjects are present in it. All the hands present in a frame are classified into one of the 5 categories: ok, stop, single-finger, peace and great. Any intermediate gesture or hands at rest are annotated as belonging to the 'others' category. Thus, the proposed gesture recognition in the crowd is modeled as a 6-class classification problem.

Each video has been recorded for about 60 to 90 seconds, thus containing 2000 to 3000 frames per video. Once the video starts, the four subjects are asked to show one gesture for 5 to 10 seconds followed by showing the same gesture with the opposite hand for the next 5 to 10 seconds. Fig. 2 presents some sample video frames from the proposed Auto-G database with gestures zoomed out on the sides.

It is observed that left and right hand gestures differ significantly in pose and orientation even after being exhibited by the same subject. Further, no limitation is imposed on the gesture orientation and pose. Therefore, the proposed database contains unconstrained pose, orientations, and positions. Also, since no such constraints are applied, some of the gestures are visible in some frames and occluded in the others. Thus, showcasing the most challenging covariate of gesture recognition in the crowd that is occlusion. In autonomous driving, primarily three types of occlusions occur: hand-over-hand occlusion, face-over-hand occlusion and, occlusion caused by the body of another subject.

Another challenge posed by the proposed Auto-G database is illumination variation due to which the perceived skin tone of the subject changes. Motion blur and off-angle gestures are also present due to the transitions between gestures. This database contains several cases where either of these challenges is present thus making it an ideal database for developing systems designed for gesture recognition in the crowd. Fig. 3 illustrates all the challenges present in the proposed database. Each row corresponds to a different gesture. The database and the ground-truth bounding-box annotations will be available at http://iab-rubric.org/resources/autoG.html.

Recognizing multiple gestures in one frame is another challenging factor present in the database. Since four subjects are present in each video and no constraints are enforced on them, there is a difference in response time of the subjects. This results in cases where more than one gesture is present in a given frame, hence requiring the development of a scheme to predict a gesture for the frame.

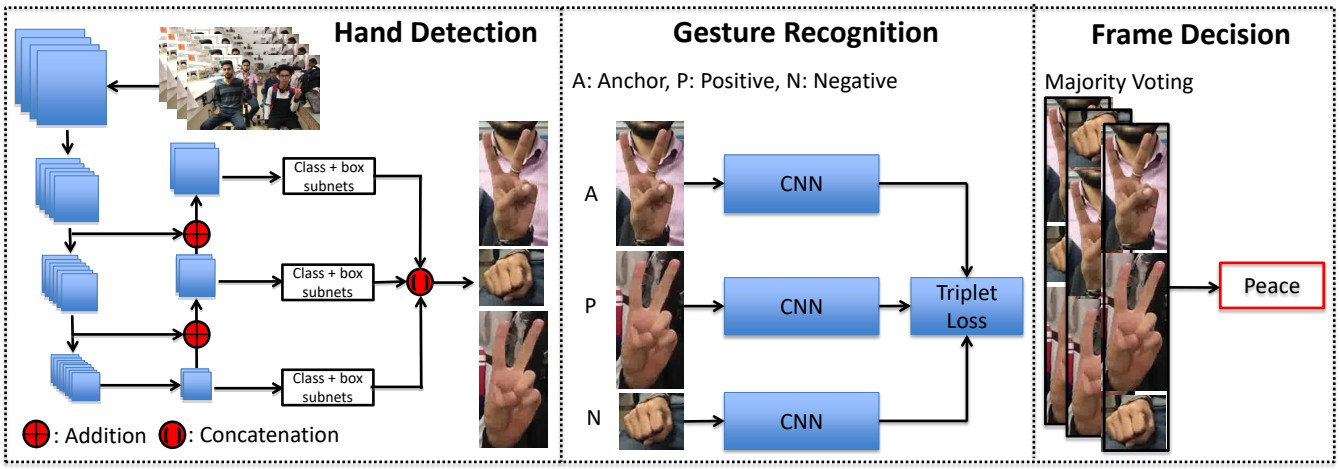

Fig. 4. Illustration of the proposed pipeline. There are three blocks in the proposed pipeline: 1) Hand detection using RetinaNet [10], 2) Gesture recognition, and 3) Frame decision.

3. PROPOSED PIPELINE

In this section, we propose a pipeline for gesture detection and recognition for autonomous driving. The pipeline consists of three blocks: 1) hand detection in the crowd using RetinaNet [10], 2) gesture recognition using triplet loss based discriminative model, and 3) majority voting for frame gesture recognition.

3.1. Hand Detection in the Crowd

Hand detection in the crowd is a challenging problem because of covariates such as motion blur, occlusion, varied background, illumination, and pose. RetinaNet [10] is a single-stage object detector and its loss function is so designed that it is robust to such hard samples. Hence, we propose to use RetinaNet [10] for hand detection in the crowd. It uses a multi-scale region proposal network with ResNet [15] as the backbone architecture. To train the model focal loss is used which is defined as:

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t) \tag{1}$$

where γ is the modulating factor and p_t is defined as,

$$p_t = \begin{cases} p & if \;\; y = 1 \\ 1-p & otherwise \end{cases} \tag{2}$$

p is the class probability of the object. This loss function allows the network to focus more on hard samples [10].

3.2. Gesture Recognition in the crowd

Khan *et al.* [16] analysed the performance of existing techniques on hand detection in the crowd. They reported that existing classifiers failed to detect hands separately in cases where there were hand-to-hand occlusions or the hands were extremely small in resolution.

To overcome these challenges we propose to use a triplet loss [17] based gesture recognition system. A triplet is a 3-tuple which can be represented as $(\vec{Z}_a, \vec{Z}_p, \vec{Z}_n)$ where, $\vec{Z}_a$ is the anchor image, $\vec{Z}_p$ the positive image, and $\vec{Z}_n$ as the negative image. The anchor is the image of class i, the positive image is another sample of the same class i, and the negative image is an image of another class k, where $i \neq k$. For an image, $g(.)$ gives the embedding of the image x. The loss function for training a model g using the triplet loss can be expressed as follows:

$$\left\| g(\vec{Z}_a) - g(\vec{Z}_p) \right\|_2^2 - \left\| g(\vec{Z}_a) - g(\vec{Z}_n) \right\|_2^2 + \alpha \tag{3}$$

$$\forall (\vec{Z}_a, \vec{Z}_p, \vec{Z}_n) \in \tau$$

where, τ is the set of all triplets generated from the training set of images and α is the margin coefficient which enforces a higher separation between the different classes in the output embedding space of the model $g(.)$. This model can also be utilized to recognize/match gesture classes which are different than those on which training is performed.

3.3. Frame Decision

Gesture recognition in autonomous driving vehicles requires the system to output a gesture decision for each frame. For predicting the gesture for each frame majority voting is performed. The voting is based on the number of same class gestures predicted per frame. In predicting the result for the frame priority is given to the recognised gestures over the 'others' category. Hence, even if one gesture is recognised, priority is given to it and the frame decision is given in favour of the recognised gesture.

4. EXPERIMENTAL RESULTS

Since there is no precedence to studying gesture recognition in the crowd, an exclusive set of experiments is performed on the proposed Auto-G database. The database contains 14,751 annotated frames and is split into subsets of 60% and 40% for training and testing respectively.

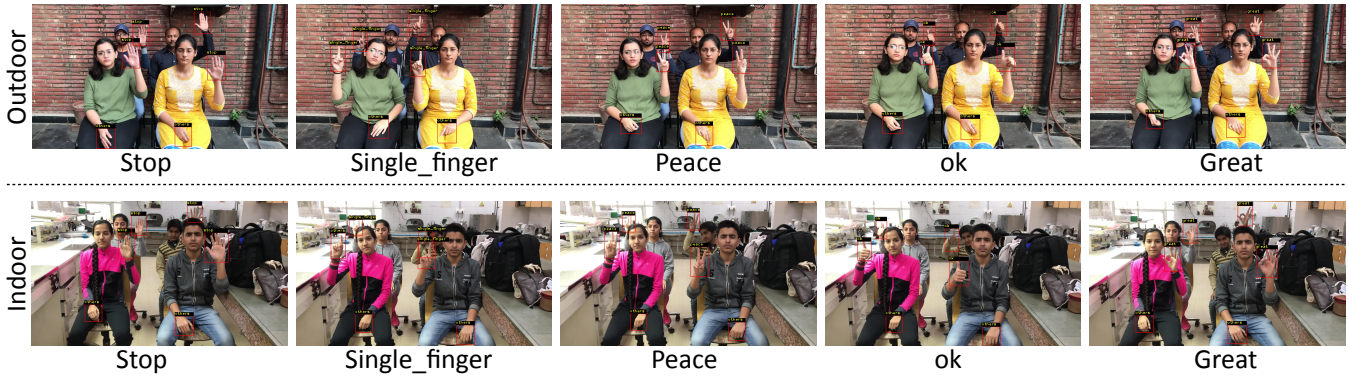

Fig. 5. Visualization of results on unlabeled part of the Auto-G database.

Table 1. Frame decision accuracy on the proposed Auto-G database (all values are in %).

Method	Detect	mAP	Identification	Rank-1
Faster-RCNN	Gesture	65.79	44.34	-
RetinaNet	Gesture	66.33	45.56	-
RetinaNet + DenseNet	Hand	95.04	52.38	15.78
Proposed	**Hand**	**95.04**	-	**90.23**

Hand detection in the crowd: To compare the detection performance, mAP measure is used. It is the metric to measure the accuracy of object detection and is defined as the average of the maximum precision at different recall values. As is evident from Table 1, hand detection achieves mAP of 95.04% when RetinaNet [10] is trained to detect hands. An additional experiment is performed to directly detect gestures using RetinaNet [10] and Faster-RCNN [9] (a two-stage object detector). However, the mAP value falls by 28.71% and 29.25% when the same model is trained to detect gestures. Thus implying that it is imperative to design a recognition model on-top of the hand detection model to perform gesture recognition in the crowd.

Gesture Recognition: The proposed dataset, Auto-G contains five gestures, thus allowing control of five different actions using the gestures. As shown by Zhang et al. [18], a verification model performs better in-case of zero-shot learning, hence, a triplet-loss based model is preferred. The results indicate the same since when DenseNet model is used in verification mode the accuracy is 15.78%, whereas, 90.23% is achieved using triplet-loss based discriminative model trained in verification mode.

Challenging cases: Autonomous driving requires accurate gesture recognition, however, while transitioning from one gesture to another there are a lot of intermediate gestures which might be misclassified. Fig. 6 presents some such cases from the Auto-G database and illustrates how half gestures or extremely blurry gestures were incorrectly recognised by the model.

The results suggest that gesture recognition in the crowd is a challenging problem but can be solved by creating an efficient end-to-end pipeline. To further strengthen the experimental evaluation, results on the unlabeled set of the database, Auto-G, are visualised in Fig. 5. It shows the results on both, indoor as well as outdoor scenarios. It can be seen from the images that the proposed pipeline is able to detect and recognise gestures which are occluded, have low or high resolution or have varying intensities due to illumination variation.

Fig. 6. Illustrates the challenging cases present in the database where the proposed pipeline failed to correctly recognise the gesture for the frame. The ground truth frame decision is in red and the predicted decision is in green (images are cropped to emphasise on the hands).

5. CONCLUSION

This research presents one of a kind database, *Auto-G*, for gesture detection and recognition related to autonomous driving. The database contains several challenging covariates for unconstrained gesture recognition such as pose, occlusion, illumination, motion blur, and low resolution. The performance on the database is benchmarked in terms of both hand detection and gesture recognition using state-of-the-art deep learning algorithms. The results obtained from the proposed Auto-G database should encourage researchers to further explore this interesting research problem.

6. ACKNOWLEDGEMENTS

M.Vatsa and R.Singh are partly supported through the Infosys Center for AI, IIIT Delhi. R. Keshari is supported through the Visvesvaraya Ph.D. Fellowship. S. Ghosh is supported through the TCS Research Fellowship.

7. REFERENCES

[1] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al., "Towards fully autonomous driving: Systems and algorithms," in *Elsevier IVS*, pp. 163–168. 2011.

[2] Udara E Manawadu, Mitsuhiro Kamezaki, Masaaki Ishikawa, Takahiro Kawano, and Shigeki Sugano, "A hand gesture based driver-vehicle interface to control lateral and longitudinal motions of an autonomous vehicle," in *IEEE ICSMC*, 2016, pp. 001785–001790.

[3] Jacques Terken, Pierre Levy, Chao Wang, Juffrizal Karjanto, Nidzamuddin Md Yusof, Felix Ros, and Sergej Zwaan, "Gesture-based and haptic interfaces for connected and autonomous driving," in *AHFSI*, pp. 107–115. 2017.

[4] Nicholas Kenneth Hobbs and Liang-yu Tom Chi, "Gesture-based automotive controls," 2015, US Patent 8,942,881.

[5] Zhongnan Shen, Fuliang Weng, and Benno Albrecht, "System and method for using gestures in autonomous parking," 2017, US Patent 9,656,690.

[6] Siddharth S Rautaray and Anupam Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[7] Eshed Ohn-Bar and Mohan Manubhai Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE TITS*, vol. 15, no. 6, pp. 2368–2377, 2014.

[8] Eshed Ohn-Bar and Mohan Manubhai Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE TIV*, vol. 1, no. 1, pp. 90–104, 2016.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, pp. 91–99. 2015.

[10] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *IEEE TPAMI*, pp. 1–1, 2018.

[11] Pichao Wang, Wanqing Li, Song Liu, Zhimin Gao, Chang Tang, and Philip Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *IAPR ICPR*, pp. 7–12. 2016.

[12] Isabelle Guyon, Vassilis Athitsos, Pat Jangyodsuk, and Hugo Jair Escalante, "The chalearn gesture dataset (cgd 2011)," *Springer MVA*, vol. 25, no. 8, pp. 1929–1951, 2014.

[13] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM TG*, vol. 33, 2014.

[14] Carl Vondrick, Donald Patterson, and Deva Ramanan, "Efficiently scaling up crowdsourced video annotation," *Springer IJCV*, vol. 101, no. 1, pp. 184–204, 2013.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, pp. 770–778. 2016.

[16] Aisha Urooj and Ali Borji, "Analysis of hand segmentation in the wild," in *IEEE CVPR*, pp. 4710–4719. 2018.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE CVPR*, pp. 815–823. 2015.

[18] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao, "Triple verification network for generalized zero-shot learning," *IEEE TIP*, vol. 28, no. 1, pp. 506–517, 2019.