Human Gesture Recognition for Autonomous Vehicle through Skeleton Estimation

Hao Chen, Da Li, Thomas Tian

***Abstract*—we proposed a vision-based explicit method that calculated the probability distribution of human intentions through gesture recognition for autonomous vehicles. The method utilizes a human skeleton estimation model to find the locations of the human skeleton key points, and constructs a feature vector to represent the human gesture based on the estimated human skeleton key points. A gesture recognition model is further deployed to obtain the human intention probability distribution with feature vectors as the inputs.**

I. Background and Impact

The smooth operation of the intelligent transportation system is based on the interaction between the autonomous vehicles and humans, while the interpretation of human intentions is still one of the biggest challenges in deploying the intelligent transportation system. Researchers have proposed many strategies for human intention recognition and analysis ([1]). Verbal commutation and gesture communication are the most promising methods ([2], [3]). Human-vehicle interaction through verbal commutation shows the versatility in signal transportation, but still faces the challenges in speech recognition as well as the diversity and accent of the languages. Compared with the verbal language, human gesture is an universal language that is much easier to be recognized and interpreted, without being disturbed by the environment. However, how to accurately capture the human gesture and analyze the human intention timely still hinder the development of the intelligent transportation system.

In this project, we addressed the issue of capturing and interpreting the human intentions by proposing a computer vision based method. The proposed method detects and estimates the human skeleton from an image using a bottom-up approach. Then the estimated human skeleton is converted to a feature vector that represents the gesture of the human. With the feature vector, the method employs convolutional neural networks (CNNs) to learn and interpret the gesture. We released our code for inspection.

We envision that our method can be used by autonomous vehicle companies for scene understanding and pedestrian intention recognition (actually, there is already a start up doing similar work [4]) and we believe our method can be generalized and used in many other industries like security and surveillance.

II. Method

We proposed to use a vision-based algorithm for interpreting human intentions. The algorithm is based on estimating the human skeleton from an image, constructing a feature vector that represents the human gesture, and recognizing the human intentions from the feature vector, see Figure 1 for the pipeline of the proposed algorithm.

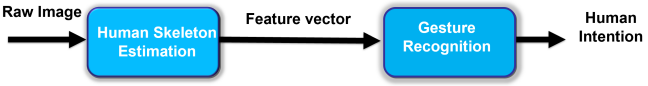


Fig. 1: The pipeline of the proposed algorithm

A. Human Skeleton Estimation

The major task of the proposed method is to estimate the human skeleton from an image. A common approach is to perform a single-person pose estimation on the basis of a person detector, which can be referred as a top-down approach. Although these top-down approaches can directly use the existing techniques, they suffer from the reliability of the person detector. The existing person detection techniques are not reliable when people are in close proximity, which is typically happened in traffic environment. Moreover, such top-down approaches usually need high computational load thus are not feasible for using in real time autonomous systems. On the other hand, bottom-up approaches, approaches that based on detecting skeleton key-points and then associating the detected key-points to form the correct skeleton, do not need to use global contextual cues from the entire body and are much computational inexpensive [2], [3].

In this project, we employed a bottom-up approach, following [5], to estimate human skeleton from an image. More specifically, we implemented CNN to build a skeleton estimation model that jointly learns the possible locations of human skeleton key-points as well as the associations of the detected human body parts. In this paper, the human skeleton key-points are the joints of human skeleton such as elbow, wrist, shoulder, *etc.*. Section III-A illustrates more details about the proposed method.

B. Gesture Extraction

Based on the estimated human skeleton, we need to extract the features that represent the gesture. The features should be invariant to the lengths of skeleton. We note that we did not consider the rotation invariance of the features premised on the assumption that most of the camera systems on

autonomous vehicle are fixed in similar locations and most human are walking or standing near autonomous vehicle, thus human captured by autonomous vehicle camera have similar orientations.

In this paper, we used a sequence of normalized vectors that encodes the orientation and association of some human skeleton key-points that are in a specific order to represent the gesture of the human, see Section III-B for more details.

C. Gesture Recognition

Recognizing and interpreting human intentions are the most challenging tasks for developing autonomous vehicles. Human cognitive model based approach and data-driven based approach are two common approaches that can be used to address this issue. In this paper, we employed CNN (different from the one we used to estimate human skeleton) to learn for interpreting a set of gestures with respect to different intentions such as a biker indicating lane changing, a pedestrian stopping the vehicles, see Section III-C for more details.

III. Prototype

A. Human Skeleton Estimation

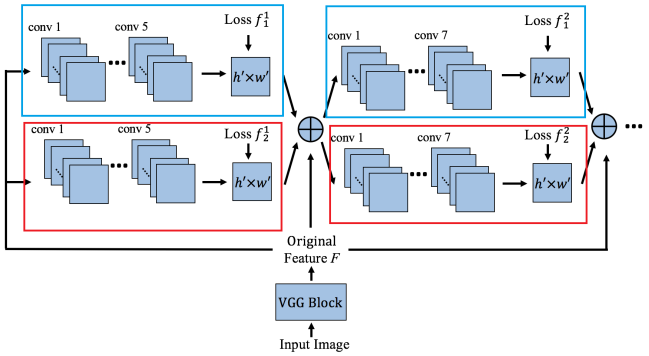
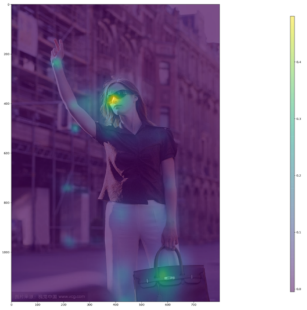


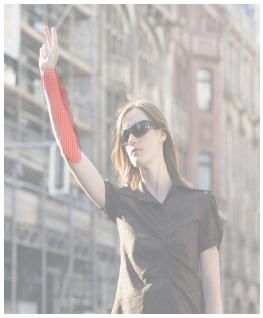
Fig. 2: The architecture of two-branch multi-stages Convolutional Neural Networks (CNNs); "conv" represents convolutional layer, each convolutional layer is followed by a Rule layer except the last convolutional layer at each branch.

We employed CNN to jointly learn the possible locations of human skeleton key-points as well as the associations of the detected key-points from an image. The image is firstly processed by the VGG-19 CNN model [6] to generate a feature map $\mathbf{F}$, then the feature map $\mathbf{F}$ is fed into a two-branch multi-stages CNN model shown in Figure 2.

The first branch (in blue box) is responsible for learning to detect human skeleton key-points, the output from the first branch at each stage is a set $\mathbb{S}^{\mathbf{i}} = (\mathbf{s_1^i}, \ldots, \mathbf{s_j^i}, \ldots \mathbf{s_n^i})$, where $\mathbb{S}^{\mathbf{i}}$ is a set of confidence maps for all the skeleton key-points at stage i and $\mathbf{s_j^i} \in \mathbb{R}^{w \times h}$ (w and h are width and height of the input image) is the confidence map for the jth human skeleton key-point (18 skeleton key points are specified, including: nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle, left eye, right eye, left ear and right ear). More specifically, at pixel location (x, y), $\mathbf{s_j}[x, y] = \rho_j$ is real number ($0 \leq \rho_j \leq 1$) which quantifies the likelihood of that location represents the jth human skeleton key-point. Figure 3(a) shows an overlay of the confidence maps (we noted that we did not apply the threshold in Figure 3(a) and scaled the likelihood to show all the possible locations including the false positives).



(a) Confidence maps for skeleton key-points detection, high temperature indicates large confidence.



(b) Parts affinity field that represents the association between the right hand and right elbow

Fig. 3: Confidence maps and Parts affinity field.

The second branch (in red box) is responsible for learning the parts affinity field that encodes the associations of the human skeleton key-points. The output from the second branch at each stage is a set $\mathbb{L}^{\mathbf{i}} = (\mathbf{L_1^i}, \ldots, \mathbf{L_k^i}, \ldots, \mathbf{L_C^i})$, where $\mathbf{L_k^i} \in \mathbb{R}^{w \times h \times 2}$ is a vector field and $\mathbf{L_k}[x, y]$ is a 2-D normal vector that indicates the orientation of the kth limp. Following [5], we refer to part pairs as limbs for clarity, despite the fact that some pairs are not human limbs (*e.g.*, the face). Figure 3(b) shows the parts affinity field that represents the association between the right hand and right elbow.

The outputs from the two branches ($\mathbb{S}^{\mathbf{i}}$, $\mathbb{L}^{\mathbf{i}}$) at each stage along with the feature map ($\mathbf{F}$) from the VGG-19 are concatenated to serve as the input for the next stage, i.e.,

$$\mathbb{S}^i = \rho^i(\mathbf{F}, \mathbb{S}^{i-1}, \mathbb{L}^{i-1}), \forall i \geq 2, \tag{1a}$$

$$\mathbb{L}^i = \phi^i(\mathbf{F}, \mathbb{S}^{i-1}, \mathbb{L}^{i-1}), \forall i \geq 2, \tag{1b}$$

where $i = 1, 2, ..., T$ is the stage index, ρ^i represents the branch 1 CNN model for inference stage i and ϕ^i represents the branch 2 CNN model for inference stage i.

In order to iteratively learn the confidence maps of skeleton key-points and the parts affinity fields, following [5], we used two loss functions for each branch at the end of each state. The loss function for the first branch (confidence maps prediction) represents the accumulative error between the predicted confidence maps and the ground truth confidence

maps, and defined as

$$f_{\mathbf{S}}^{i} = \sum_{j=1}^{J} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot ||\mathbf{S}_{j}^{t}(\mathbf{p}) - \mathbf{S}_{j}^{*}(\mathbf{p})||_{2}^{2}, \tag{2}$$

where $J = 18$ is the total number of the skeleton key-points, $\mathbf{p} = [x, y]$ is the pixel location and $\mathbf{W}(\mathbf{p})$ is an indicator function that $\mathbf{W}(\mathbf{p}) = 0$ when no valid annotation found at image location $\mathbf{p}$. Similarly, the loss function for the second branch (parts affinity fields detection) represents the accumulative error between the parts affinity fields and the ground truth parts affinity fields, and defined as

$$f_{\mathbf{L}}^{i} = \sum_{c=1}^{C} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot ||\mathbf{L}_{c}^{i}(\mathbf{p}) - \mathbf{L}_{c}^{*}(\mathbf{p})||_{2}^{2}, \tag{3}$$

where C is the total number of valid pairs of skeleton key-points (e.g., right hand and right elbow is a valid pair while right hand and left elbow is not). To measure the confidence in the association of a pair of detected skeleton key-points, following [5], we integrated each parts affinity field $\mathbf{L}_k$ along the line segment connecting two skeleton key-points locations (see the red line in Figure 4) to obtain the confidence value.

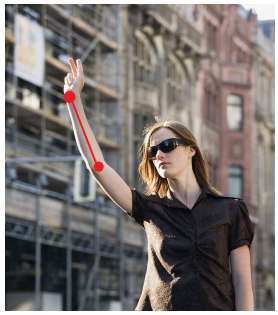


Fig. 4: Integral path for evaluating the confidence in the association of a pair of detected skeleton key points.

B. Gesture Extraction

We need to extract the features that are invariant to the lengths of skeleton. To do that, we select the locations ($[x_i, y_i]$) of 7 human skeleton key-points: left wrist, left elbow, left shoulder, neck, right shoulder, right elbow and right wrist. From these seven key points, we can obtain 6 directed lines $\mathbf{l_j}$ ($j = 1, 2, ..., 6$) following the specific order: neck to left shoulder, left shoulder to left elbow, left elbow to left hand, neck to right shoulder, right shoulder to right elbow, right elbow to right hand. Then, the vector of these 6 directed lines $\mathbf{L} = [\mathbf{l_1}, \ldots, \mathbf{l_6}]$ can be used to represent the gesture as $\mathbf{L}$ encodes the associations and orientations of the selected 7 human skeleton key-points. We further normalize $\mathbf{L}$ to get the feature vector $\hat{\mathbf{L}}$ that is invariant to skeleton length following Equation 4. Figure 5 shows visualization of human gesture and the corresponding feature vectors.

$$\hat{\mathbf{L}} = [\hat{\mathbf{l}}_\mathbf{1}, \ldots, \hat{\mathbf{l}}_\mathbf{6}], \tag{4a}$$

$$\hat{\mathbf{l}}_\mathbf{i} = \frac{1}{\sqrt{||\hat{\mathbf{l}}_\mathbf{i}||_2}}. \tag{4b}$$

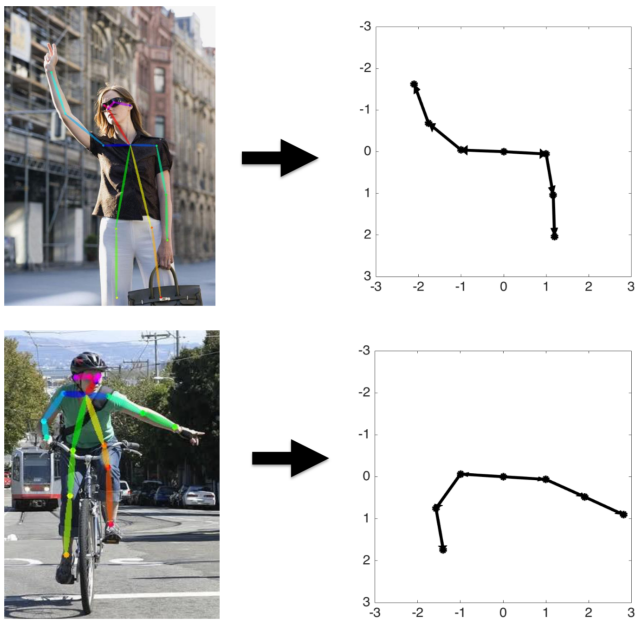


Fig. 5: Skeleton key points and the corresponding uniformed feature vectors. Top: A pedestrian requesting a ride; bottom: a biker making lane change.

C. Gesture Recognition Implementation

We note that for our best knowledge, there is no publicly realized data set for human intentions related to autonomous vehicle development (However, we did found a commercial data set from *humanising autonomy* [4] that could be used in our work). Since our focus in this work is the human skeleton estimation, we manually labeled 1000 images that have three most common intentions: pedestrian stopping the vehicle, pedestrian requesting a ride and biker indicating the lane changing, to test the gesture recognition concept. We used CNN, as shown in Fig. 6, to train a model to classify the human intentions using the feature vectors as input. The performance of the whole pipeline is presented and evaluated in Section IV.

IV. Result

The proposed method has two important steps, skeleton estimation and gesture recognition, and the performance depends on the accuracy of both steps. We implemented the skeleton estimation model following the instructions in [5]. We trained the model using COCO data set with a GeForce GTX 1080 GPU for 15 hours and the model shows satisfying

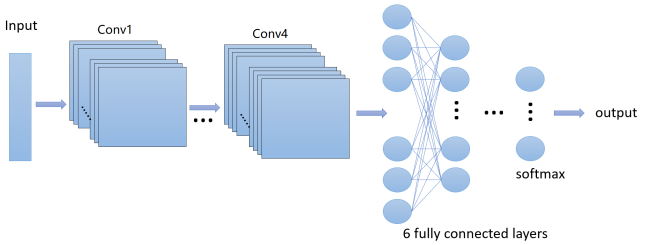


Fig. 6: The architecture of the Convolutional Neural Networks for gesture recognition.

performance. The model is able to detect the human skeletons (color lines in Figure 7). For the human gesture recognition model, we manually labeled 1000 images and the overall accuracy is 86% for interpreting the intentions correctly.

We tested the complete pipe-line on the images obtained from the internet and the images taken by our team. The testing results are shown in Figure 7. Figure 7(a) shows a scenario where a biker is using gesture to show a potential lane changing, and our method successfully interprets the gesture as biker indicating lane changing based on the probability distribution of different intentions (59.6% probability for biker indicating lane changing, 23.1% for pedestrian stopping the vehicle and 17.2% for passenger requesting a ride). Similarly, Figure 7(b) and Figure 7(c) show two scenarios where a pedestrian is using gesture to stop a vehicle when walking on the sidewalk and a passenger requesting a ride, respectively, and our method interprets the gestures correctly.

Figure 7(d) shows a case where the gesture recognition model may fail under the wrong estimated human skeleton. In Figure 7(d), a pedestrian is using gesture to stop a vehicle when walking on the sidewalk. However, our method interprets the gesture as biker indicating lane changing. The reason for the wrong interpretation is due to the skeleton estimation model falsely estimated the location of the left arm.

(a) A biker indicates lane changing.

(b) A pedestrian stops the car.

(c) A pedestrian requests a ride.

(d) Incorrectly classified example; the blue box shows the true location of the left arm and the red box shows the estimated location of the left arm.

Fig. 7: Human gesture interpretation results; the sub-captions are the ground truth intentions and the red texts indicate the intentions interpreted by the proposed method.

V. Conclusion

In this project, we proposed a vision-based method for interpreting human intentions through recognizing the human gesture. We first implemented a bottom-up approach to find the locations of the human skeleton key-points from an image and associated the skeleton key-points to form a human skeleton, then the estimated human skeleton is used to build a feature vector that represents the gesture of the human. Finally, we implemented a gesture recognition model to interpret the feature vector. Our method could identify three common human intentions related to autonomous vehicle scene understanding with relative high accuracy. We envision that our method can be used by autonomous vehicle companies for scene understanding and pedestrian intention recognition. We remark that our main focus in this project is to implement and test the human skeleton estimation method while the gesture recognition is an application of the human skeleton estimation method and needs to be further investigated.

References

[1] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE MultiMedia*, vol. 13, no. 2, pp. 18–31, 2006.

[2] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[3] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[4] "Humanising autonomy," https://www.humanisingautonomy.com/, accessed: 2018-12-12.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.