2020 IEEE Intelligent Vehicles Symposium (IV)
October 20-23, 2020. Las Vegas, USA

Extraction and analysis of massive skeletal information from video data of crowded urban locations for understanding implicit gestures of road users

Andreas Keler[1], Patrick Malcolm[1], Georgios Grigoropoulos[1], and Niklas Grabbe[2]

Abstract— **This work explains the possible inferable information from a long-term video acquisition with cameras installed in close proximity to pedestrian movements with an unobstructed view of the entire intersection. The main goal is detecting implicit and explicit gestures and understanding communication and interactions between different types of road users. After explaining the designs of different gesture classification approaches, we relate the qualitative approach with our classification scheme for the extracted skeletons. To this end, a sequence with selected moving entities is selected and compared with the manually annotated video sequence. Results show the limitations of the automated approach and indicate a level of subjectivity in the manual annotation procedure. Subsequently, we discuss possibilities and restrictions of our approach and reflect on the importance of the specific conditions of video acquisitions. Depending on the field of view and distance between installed video cameras and moving vulnerable road users (VRUs), we are able to define the restrictions of our approach. As a result, we are able to define a selection of suitable applications for our approach.**

I. Introduction

Vulnerable road users such as pedestrians, bicyclists or e-scooter riders often come into interaction with multiple other types of road users in dense urban traffic situations. They also interact with each other as well as with the other road user types in areas that often lack markings or signage or where the different road user types are not segregated, such as designated pedestrian zones, sidewalks, shared surfaces, bicycle paths and road intersections. Due to the inherent flexibility of road user motion, there are a very high number of possible maneuvers and interactions with other road user types.

In road traffic, people communicate not only via prescribed signals such as indicators, brake lights and horns, but also through informal communication channels. According to Merten [1], various options are available for communication: schema formation, anticipatory behavior, non-verbal communication, facial expressions, eye contact, gestures and body movements. The automation must also be able to perceive all of these informal signs and interpret them in the environmental context in order to predict the behavior of others. Based on this, the automation can adopt an adapted behavior, which must also be understood by other road users. Thus, understanding the movements, interactions and intentions of observable vulnerable road users together with inferable local knowledge is important for partially and fully automated driving in urban areas to ensure safe and efficient traffic.

Therefore, section II of this work explains historical and state-of-the-art research into estimating human poses from video data and understanding the communication of vulnerable road users participating in urban traffic. We name the most significant applications of pose estimation methods, explain the motion correspondence problem, and deliver basic insights on interactions and communications between traffic participants.

This paper proposes an image-based method for extracting and classifying poses of pedestrians, cyclists, motorcyclists and e-scooter drivers from video data coming from a camera with a static position mounted at a high angle.

After applying the pre-trained convolutional neural network (CNN) PoseNet [2] on selected video data of a complex urban intersection in Munich, Germany, we gather a massive collection of skeletal information data sets. In section III, we propose a methodology for tracking detected pedestrians between video frames as well as a projection approach for translating pixel coordinates into rectified geographic coordinates. Besides the tracking, the estimated poses of every detected individual can be classified into gestures and we introduce novel ideas of simple and efficient implementations. Additionally, we explain a qualitative manual annotation procedure as a suggestion for evaluating the detection and pose estimation approach.

The outcomes of testing our approach on a select sample of the collected video data are presented in section IV.

The problems and challenges of the proposed methodology are discussed in section V and related to the resulting outcomes of our case study. We outline encountered and potential problems, restrictions of the approach and its general usefulness for transport-related applications.

Future steps and extensions of the presented applications are suggested in section VI.

II. State of the Art in Pose Estimation and Understanding VRU Communication

A. Pose estimation applications

Early research on pose estimation focused on part-based models, such as that by Fischler and Elschlager [3] and

[1] are with the Chair of Traffic Engineering and Control, Technical University of Munich, Munich, BY 80333, Germany;

[2] is with the Chair of Ergonomics, Technical University of Munich, Garching, BY 85747, Germany;

E-mail: andreas.keler@tum.de.

978-1-7281-6673-5/20/$31.00 ©2020 IEEE

Felzenszwalb, et al. [4]. The Deformable Part Model (DPM) by Fischler and Elschlager [3] was the base for numerous probabilistic graphical models for two-dimensional pose estimation. In such approaches, body parts are detected together with constraints and properties between the parts. On the other hand, there are single person pose estimation procedures which use deep Convolutional Neural Networks (CNN) achieving high performances.

Papandreou, et al. [2] divide multi-person pose estimation approaches into two groups, namely top-down and bottom-up detection procedures. The aim is to associate person part detections with person instances, where the top-down option implies the person detection first and subsequently the pose estimation In the bottom-up approaches, the first step corresponds to the detection of body parts and the second step to an association of the detected body parts to human instances [2].

The PoseNet model by Papandreou, et al. [2] can be used for typical state-of-the-art (2020) pose estimation of moving humans in image and video data. It consist of the two components: (1) person box detection, which is a Faster-RCNN system [5], and (2) person pose estimation, which predicts all 17 key points per person per bounding box (from the first component). The second component, the person pose estimation, consists of an (a) image cropping step, a (b) heat map and offset prediction with CNN step, a (c) model training step, a (d) pose rescoring step, and a (e) non-maximum suppression step based on object-key-point-similarity. Pose rescoring is estimated via a formula of the final instance-level pose detection based on the confidence of each key point by the maximization over locations and averaging key points. The last step in component 2e serves for excluding multiple detections in the person-detector stage.

In addition to robust methodologies, high quality training data sets are essential to guarantee efficient motion capture approaches. Therefore, training data is often multivariate with pairs of high-resolution stereo images and LiDAR data along with two-dimensional image labels and three-dimensional labels of pedestrians in a global coordinate frame [6]. Labelling the data can be performed by either manual or automated annotation approaches.

B. On the motion correspondence problem

First defined in the early 1990s, motion correspondence is described as a fundamental problem in computer vision and other disciplines [7] in which points measured at discrete time steps must be corresponded to points in previous and subsequent time steps in order to glean information about their motion. The problem originates from perception and vision research and focuses on extraction of motion and identifying physical directions from data often represented as arrays of dots [8]. There have been numerous suggested solutions for solving the motion correspondence problem, especially as extensions of the limitations of applying nearest neighbor algorithms [7, 9]. While in some cases, matching detected moving entities between the frames of a video appears trivial, there are instances where matching is challenging, such as in the case of numerous densely distributed moving objects. Therefore, Veenman, et al. [10] introduce existing greedy matching algorithms on individual, combined, and global motion models. These approaches take into account various constraints, such as trajectory smoothness and object speed limitations.

C. Interactions and communications between traffic participants

According to Fuest, et al. [11], there is explicit and implicit communication in road traffic. Examples given of explicit communication were light signals, the horn and the use of an indicator. The term implicit communication in traffic was used for instance when a car slows down and thus signals to the pedestrian that he or she can cross the road. It was also described that non-verbal communication (e.g. a waving hand) in traffic is considered as explicit communication. Rasouli, et al. [12] also describe hand movements in road traffic as an explicit form of communication.

In particular, Hürlimann and von Hebenstreit [13] distinguish between four forms of explicit and non-verbal communication of pedestrians. These are the gesture of grant, gesture of gratitude, gesture of intent and coercive gesture. The gesture of grant occurs most frequently, followed by the gesture of gratitude. The intentional gesture and the coercive gesture are used very rarely. If we look at the frequency distribution over the different age ranges, it is noticeable that the gesture of grant, gesture of gratitude and intentional gesture occurs in the middle age range with a maximum of 45 to 65 years of age and the coercive gesture is mainly used by seniors over 65. Furthermore, the hand signals are mainly used at pedestrian crossings without light signal system (LSS) or unprotected areas and very rarely at LSS.

In general, explicit gestures are rarely used, at only 3%, and even then, usually only as a reaction to driver actions. Instead, mostly implicit communication takes place. [12, 14]. In addition, Rasouli and Tsotsos [15] provide a comprehensive overview of 38 different factors involved in pedestrian decision-making process at the time of crossing when facing conventional cars or autonomous vehicles. Thereby, the factors that influence pedestrian behavior are divided into two groups: the ones that directly relate to pedestrians and environmental ones which relate to the surrounding context. Furthermore, Markkula, et al. [16] presented a conceptual framework for understanding interactive behavior in human and automated road traffic. The key contributions are a stringent definition of the term "interaction" and a taxonomy of the types of behaviors that road users exhibit in interactions which can be helpful for empirical methodologies in studying the interaction between road users.

There are several ways to analyze the interaction between road traffic users in a qualitative form. For example, Rasouli, et al. [12] create temporal interaction diagrams of actions and events between pedestrians and drivers to document the temporal execution of actions and their sequence and above all to discover behavioral patterns in interaction as pictured in Figure 1. After the sequences have been identified, the frequencies of the respective actions can be displayed in the form of so-called Sankey diagrams in order to discover frequent behavior patterns as pictured in Figure 2, which in turn can be used to predict behavior. To generate this data, ground based videos [12] or observation protocols [17-19] are required. For example, an observer at the roadside can log the interaction using an HTML app on the tablet with predefined buttons for the actions of the pedestrian and the driver and also general context information [20, 21] or a video recording can be edited afterwards using a video analysis software and temporal events and actions are logged. This procedure is very elaborate, subjectively and the observers need a high level of experience and knowledge of the domain. The reliability of the correct assignment of the actions, so that the protocol corresponds to the process in reality, could be increased by a preliminary observation in which three observers simultaneously record the same situations and their results are finally compared via the Cohens kappa to measure the interrater reliability. This ensures that the observer's assessments are objective.

Additionally, the data can be enhanced by thought processes of the road users using questionnaires answered by the corresponding road users [22].

Figure 1. Temporal interaction diagram of actions and events between pedestrian and driver, adapted from Rasouli, et al. [12]

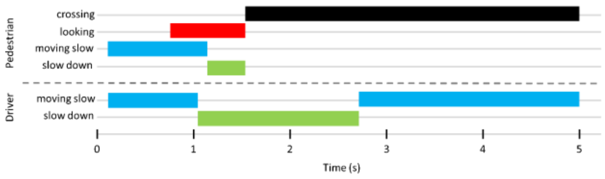
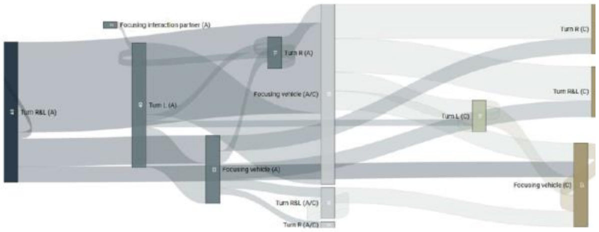


Figure 2. Example of a Sankey diagram for sequences and their frequencies of head movements of a pedestrian when crossing

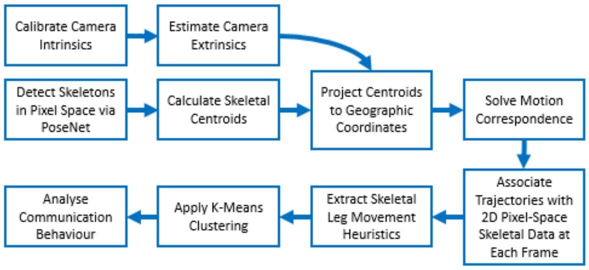


Numerous applications focus on detecting and understanding pedestrians from the ego perspective of a highly-automated vehicle. As Fang and López [23] point out, the decision to cross a pedestrian crossing might be estimated with an overall accuracy and therefore use a combination of CNN-based pedestrian detection, tracking and pose estimation. Due to the restrictions of mounting a number of cameras at selected street light poles of an intersection, often systems have numerous sensors attached to one box. Other implementations of video cameras employ four cameras covering together the whole surface area of one intersection.

III. Methodological Approach

Our approach takes the geographical relations of transport infrastructural design elements into account when defining functional regions. Based on the surface areas with assigned transport infrastructure for pedestrians, cyclists (together with e-scooter drivers) and motorists, we are able to estimate the heightened possibility of pedestrians to illegally cross the road assigned for motorists and cyclists. The components of our approach are presented by a flow chart in Figure 3 and will be explained in the following.

Figure 3. Flow chart of the presented approach for detecting skeletal information, solving the motion correspondence problem and analysing communication behaviour


Calibrate Camera Intrinsics
Estimate Camera Extrinsics
Detect Skeletons in Pixel Space via PoseNet
Calculate Skeletal Centroids
Project Centroids to Geographic Coordinates
Solve Motion Correspondence
Associate Trajectories with 2D Pixel-Space Skeletal Data at Each Frame
Extract Skeletal Leg Movement Heuristics
Apply K-Means Clustering
Analyse Communication Behaviour

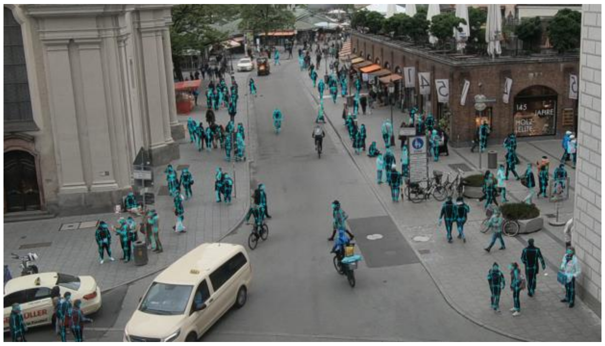
A. *Extracting skeletal information from pose estimation*

As already explained in subsection IIA, PoseNet is a vision model usable for human pose estimation in video data consisting of two successive stages. The first stage is the person box detection and cropping using a CNN backbone, which is pre-trained for image classification on ImageNet [24] and a Faster-RCNN system for person detection trained via the person category of the COCO person key points detection dataset [25] while ignoring the box annotations of the other 79 COCO categories. With its 2016 version of the COCO dataset, we have annotations of the key points, distinguishing body joints from face landmarks [2]. The object key point similarity, ranging from 0 to 1, serves for estimating the matching quality between ground truth and predicted poses, and based on these, we are able to estimate the overall matching confidence via an average precision metric.

We apply the pre-trained convolutional neural network (CNN) PoseNet [2] for multi-person detection and 2-D pose estimation on our video data sets showing a high angle view of the area of interest.

This paper proposes an image-based method for extracting and classifying poses of pedestrians, cyclists and e-scooter drivers from video data coming from a camera with a static position mounted at a high angle. An additional camera was installed showing toward the south at a lower angle shot. An example view of extracted skeletal information is pictured in Figure 4.

Figure 4. Resulting pose estimation via PoseNet of a video of an urban pedestrian area

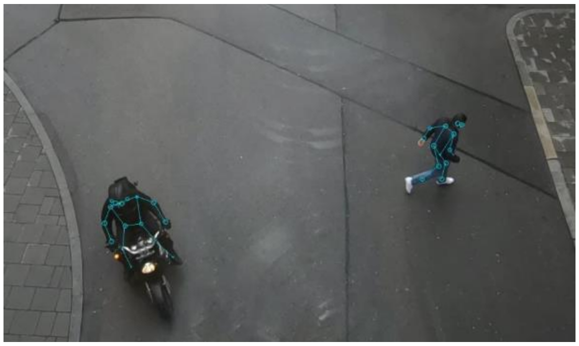


The subsequent procedure includes a transformation of pixel coordinates into geo-coordinates. By manually selecting distinguishing features which appear in both the captured image and in satellite imagery (and by either assuming a roughly flat ground plane or estimating point elevations), a list of point correspondences between 2D pixel space and 3D geo-coordinates can be generated. These, along with pre-calibrated intrinsic camera parameters, are then passed to a camera pose estimation algorithm, in our case OpenCV's SolvePnPRansac function, in order to determine the camera's position and orientation in 3D geo-coordinates. However, in order for this algorithm to function correctly, the input points in geo-space must first be normalized by subtracting out false northings and eastings so that the coordinate values are small. Using this information, we then deproject the skeletal centroids from 2D pixel space onto the 3D geo-space plane z=100 cm, a height which corresponds to an estimate of the average centroid height of all the observed skeletons.

B. *Distinguishing and tracking different traffic participants*

After the deprojection of skeletal centroids from pixel coordinates to geo-coordinates, we apply a motion correspondence algorithm to associate skeletons from each frame into coherent trajectories, which then allows us to calculate the instantaneous speed and acceleration at every frame and to identify patterns of implicit gestures of pedestrians and cyclists. The various greedy matching algorithms described by Veenman, et al. [10] were tested, but since these are unable to cope with more than one consecutive frame of missed detection, we found that for our data they did not provide adequate results. Therefore, we implemented a variation on the nearest neighbor matching approach in which we iterate through each frame and for each point in that frame we calculate a generalized cost for a connection from that point to each of the points in the following N frames (in our case N=4) equal to the Euclidean distance plus some penalty for points in frames beyond the next one. Already-connected points are given infinite cost, and virtual "no-connection" points are generated with a moderate cost to allow for the possibility of trajectory termination. Using these costs, we then apply a linear sum assignment algorithm to find the least cost assignment of connections from points in the current frame to the subsequent frames. In Figure 5 example skeletons are shown of a motorcyclist (left) and a pedestrian (right).

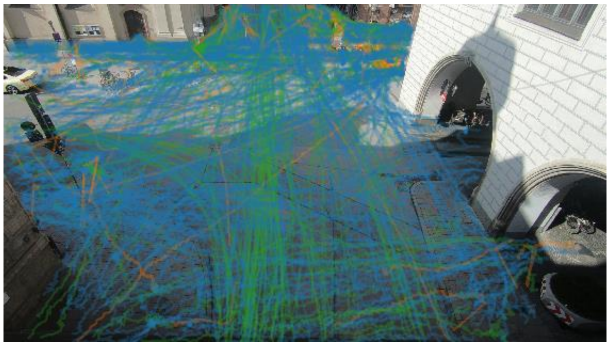
Figure 5. Example skeletons of a motorcyclist (left) and a pedestrian (right)



C. *Differentiation of pedestrians and cyclists*

In order to differentiate between pedestrians and cyclists, we applied a k-means clustering algorithm to the speeds of each trajectory. With this approach, we see what appears to be moderately good class separation, as shown in Figure 6 with cluster estimation for k=3 based on average skeletal trajectory speed, its standard deviation and the maximum observed speed. As this is an unsupervised clustering approach, it is difficult to precisely evaluate the accuracy of the technique, but one can easily observe that trajectories which follow the street, tend to be grouped into a different cluster than trajectories which originate from the sidewalk and cross the street.

Figure 6. Classified trajectories for k=3 based on aggregated skeletal point trajectory speed statistics



D. *Manual annotation of pedestrian gestures*

In addition to the conventional approaches of annotating raw data with additional attribute values for producing training data sets for motion capture, pose estimation and classification, there are approaches for observational methods for generating qualitative data from conducted analyses. Silva, et al. [26] introduce a method called (re)Action for coding interactions, events and consequences in three successive steps, namely open coding, axial coding and selective coding. It follows Straussian Grounded Theory [27]. In the first step, open coding, we define the different types of actors with the respective types of actions. It refers to a specific video data set without having prior knowledge on investigation area and typical maneuvers of road user types. Additionally, so called dimensional ranges describe the attribute formats of every actor-action, as temporal, integer or string. This step ends when no further actor-action category can be found. Subsequently, axial coding defines relationships between actor-actions and introduces them as interaction events in a conditional event matrix. In the last step, selective coding, consequences of actor-actions and events are identified through a quantitative analysis of the annotations of the previous steps. As Silva, et al. [26] have a case study on violating bicycle infrastructure as bike paths and lanes, there is a selection of different usable dimensional ranges of actor-actions including lateral positions, instantaneous speed, secondary activities and change of mode of transportation. For our cases study, we mainly focus on the identification of instantaneous speed and acceleration of detected road users, since these attributes indicate different road user types and implicit gestures as pedestrians slowing down their movement at specific locations. The latter is an example for an implicit gesture, where a pedestrian identifies a conflict with another road user. Whereas, accelerating pedestrians might indicate crossing a road segment assigned to vehicle drivers.

E. Designing and relating gesture classification approaches

In this paper we relate the quantitative and qualitative gesture classification approaches on one specific test data set. One focus is currently to distinguish pedestrians, cyclists, motorcyclists and e-scooter riders. Subsequently, the acceleration or instantaneous speed of generated trajectories (consisting of skeletal centroids) is used for detecting changes in movement that may relate to implicit gestures of road user. Other gesture classification approaches can include the clustering of repetitive skeletal movements based on an average pedestrian, cyclist, motorcyclist or e-scooter rider.

IV. Results

In our experiments, we found that the pre-trained PoseNet algorithm provided overall reliable detection of skeletons, albeit with occasional missed detections, usually due to visual obstruction or similarity in clothing color to that of the background. In order to assess the detection accuracy of the proposed methodology, a randomly selected sample video of 11.2 minutes consisting of 20208 frames was analyzed. In the video sample a total of 401513 skeletons were detected. In the trajectory generation step, each skeleton in each frame was reduced to a single representative point, each of which was then matched to a point in a future frame if possible. From these points, 23486 trajectories were generated, of which 19280 had an endpoint not near the edge of the image. This implies that 95.2% of the skeletons were either matched to a trajectory in a future frame or ended a trajectory near the edge of the image. Still, 82.1% of all trajectories ended inside the image rather than at the image edge, indicating problems with occlusion and dropout. Figure 7 presents the heatmap of trajectory endpoints not near the image edge. As expected, the majority of endpoints are found around fixed obstacles, while relatively few are located in the inner parts of the image not near obstructions. Therefore, it is clear that the number and density of static and dynamic obstacles has a severe effect on the detection accuracy.

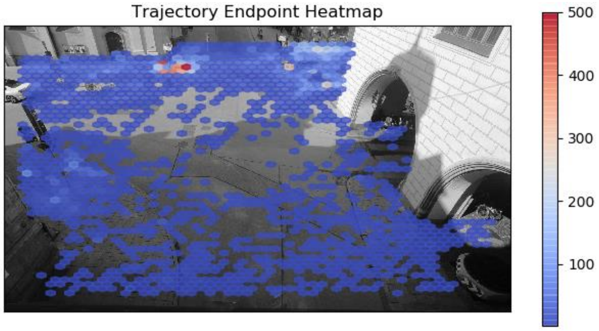


Figure 7: Trajectory endpoint heatmap

Disregarding trajectories ending near static obstacles (buildings, traffic signs, etc.), 50.1% of all matched trajectories ended unexpectedly, i.e. due to either dynamic obstacles such as vehicles, pedestrians, and cyclists or failed detections on the part of PoseNet due to measurement noise, lack of contrast between the subjects and the background, and random errors. Missed detections are also much more common for road users located farther from the camera who are thus represented by a smaller number of pixels. We also noted that the skeletal points detected in adjacent frames often exhibit a certain amount of noise, making tracking of fine movements challenging. Again, this is especially true of people who are situated far away from the camera.

Despite the limitations, larger bodily movements can be readily detected if the orientation of the person in relation to the camera is favorable. Therefore, the proposed methodology can detect skeletal points with a high degree of accuracy. However, as a result of static and dynamic obstructions, continuous trajectories are often interrupted, which remains an area for future improvement. Future work could focus on developing a methodology for estimating a full 3D representation of each skeleton as opposed to the simple 2D pixel-space representation used here, which would allow for much more reliable gesture detection.

In comparing several motion correspondence algorithms with our own modified nearest neighbor tracking approach, we found that while our approach is in some sense more naïve in that it does not consider trajectory smoothness, the ability to handle multiple consecutive missed point detections outweighs this disadvantage and provided better results on our data set as evaluated by manual verification. One reason for the success of our approach on this data set is the relatively moderate density of tracked points, which allows for less potential confusion between trajectories. In high-density data sets, more sophisticated approaches such as those described by Veenman, et al. [10] are likely to outperform ours. Future work could, expand our motion correspondence algorithm by adding cost penalties for unsmooth trajectories and compare the performance on higher-density data sets.

V. Discussion

We have shown that the use of simple k-means clustering to differentiate between pedestrians and bicyclists with moderate accuracy is possible. One major limitation is that the skeletal representations exist in 2D pixel space and are thus not invariant to changes in location or orientation relative to the camera. A methodology for estimating 3D geo-coordinates for each skeletal point as mentioned previously would likely greatly improve the ability to differentiate the classes. Furthermore, different heuristic metrics than the ones used here could be tested, and instead of simply extracting statistical measures for the metrics over time, a frequency analysis technique could be employed, by which the dominant frequency could be extracted and used as an input to the clustering algorithm.

VI. Outlook

With our approach, occupied space and how and when it is used can be examined further, as can the interaction of road users with their environment. The discussed methodology should be possible to be implemented also for real – time or near real – time applications considering the restrictions or the motion correspondence problem. In this context, the availability of key skeletal point positions of VRUs provides additional descriptive data for analyzing traffic situations and enables a broad selection of possible future applications. By including video data from additional nearby situated cameras or by using stereo cameras, it should be possible to more accurately detect road users and to deduce additional features such as three-dimensional skeletal information for more robust gesture detection and to allow for the estimation of the heights of every pedestrian, cyclist or e-scooter rider, which can lead to better road user detection and classification. Patterns involving groups of road users classified via relevant operational and communication features might be then used as information for infrastructure to vehicle (I2V) communications and for supporting automated driving technologies.

Also, typical maneuvers and typical interaction sequences can potentially be deduced from the skeletal point data enabling the analysis and data-driven modelling of traffic scenarios. For example, probabilities of crossing road segments by individual or groups of vulnerable road users might be useful for predicting behavioral patterns such as the violation of traffic regulations. These situations should be then defined as scenes of a fully defined scenario (for a detailed representation of the static model of the respective intersection). In addition to gesture classifications, it might be possible to apply various clustering techniques to find similar poses or sequences of detected gestures. Through the skeletal point data, typical traffic scenarios can then be identified and clustered enabling solutions in the area of automatic incident detection, traffic management and traffic safety.

Acknowledgment

This work is a result of the research project @CITY – Automated Cars and Intelligent Traffic in the City. The project is supported by the Federal Ministry for Economic Affairs and Energy (BMWi), based on a decision taken by the German Bundestag. The author is solely responsible for the content of this publication.

References

[1] K. Merten, "Kommunikationsprozesse im Straßenverkehr," *Symposion,* vol. 77, 1977.

[2] G. Papandreou *et al.*, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4903-4911.

[3] M. A. Fischler and R. A. Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Transactions on Computers,* vol. C-22, no. 1, pp. 67-92, 1973, doi: 10.1109/T-C.1973.223602.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 June 2008 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587597.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.

[6] W. Kim *et al.*, "PedX: Benchmark Dataset for Metric 3-D Pose Estimation of Pedestrians in Complex Urban Intersections," *IEEE Robotics and Automation Letters,* vol. 4, no. 2, pp. 1940-1947, 2019, doi: 10.1109/LRA.2019.2896705.

[7] I. J. Cox, "A review of statistical data association techniques for motion correspondence," *International Journal of Computer Vision,* vol. 10, no. 1, pp. 53-66, 1993/02/01 1993, doi: 10.1007/BF01440847.

[8] T. Watanabe and R. Cole, "Propagation of local motion correspondence," *Vision Research,* vol. 35, no. 20, pp. 2853-2861, 1995/10/01/ 1995, doi: https://doi.org/10.1016/0042-6989(95)00064-7.

[9] M. R. Dawson, "The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem," *Psychological Review,* vol. 98, no. 4, pp. 569-603, 1991, doi: 10.1037/0033-295X.98.4.569.

[10] C. J. Veenman, M. J. T. Reinders, and E. Backer, "Resolving motion correspondence for densely moving points," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 1, pp. 54-72, 2001, doi: 10.1109/34.899946.

[11] T. Fuest, L. Sorokin, H. Bellem, and K. Bengler, "Taxonomy of Traffic Situations for the Interaction between Automated Vehicles and Human Road Users," in *Advances in Human Aspects of Transportation*, Cham, N. A. Stanton, Ed., 2018// 2018: Springer International Publishing, pp. 708-719.

[12] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017: IEEE, pp. 264-269.

[13] F. W. Hürlimann and B. von Hebenstreit, *Verkehrssicherheit in der Praxis*. Bern: Verlag Hans Huber, 1987.

[14] D. Dey and J. Terken, "Pedestrian Interaction with Vehicles: Roles of Explicit and Implicit Communication," presented at the Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Oldenburg, Germany, 2017. [Online]. Available: https://doi.org/10.1145/3122986.3123009.

[15] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE transactions on intelligent transportation systems,* 2019.

[16] G. Markkula *et al.*, "Defining interactions: A conceptual framework for understanding interactive behaviour in human and automated road traffic," *PsyArXiv,* 2020.

[17] J. Imbsweiler, R. Palyafári, F. Puente León, and B. Deml, "Untersuchung des Entscheidungsverhaltens in kooperativen Verkehrssituationen am Beispiel einer Engstelle," in *at - Automatisierungstechnik* vol. 65, ed, 2017, p. 477.

[18] M. Šucha, "Road users' strategies and communication: driver-pedestrian interaction," *Transport Research Arena (TRA),* 2014.

[19] M. Vollrath, A. K. Huemer, C. Teller, A. Likhacheva, and J. Fricke, "Do German drivers use their smartphones safely?—Not really!," *Accident Analysis & Prevention,* vol. 96, pp. 29-38, 2016/11/01/ 2016, doi: https://doi.org/10.1016/j.aap.2016.06.003.

[20] A. Dietrich, "Interaction in Urban Traffic – Insights into an Observation of Pedestrian-Vehicle Encounters.," presented at the interACT Webinar, 2018. [Online]. Available: https://www.interact-roadautomation.eu/wp-content/uploads/interACT_Webinar_Andre-Dietrich_180509_final-1.pdf.

[21] A. Dietrich and J. Ruenz, "Observing Traffic – Utilizing a Ground Based LiDAR and Observation Protocols at a T-Junction in Germany," in *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Cham, S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, and Y. Fujita, Eds., 2019// 2019: Springer International Publishing, pp. 537-542.

[22] N. Merat, T. Louw, R. Madigan, M. Wilbrink, and A. Schieben, "What externally presented information do VRUs require when interacting with fully Automated Road Transport Systems in shared space?," *Accident Analysis & Prevention,* vol. 118, pp. 244-252, 2018/09/01/ 2018, doi: https://doi.org/10.1016/j.aap.2018.03.018.

[23] Z. Fang and A. M. López, "Is the Pedestrian going to Cross? Answering by 2D Pose Estimation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 26-30 June 2018 2018, pp. 1271-1276, doi: 10.1109/IVS.2018.8500413.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009: Ieee, pp. 248-255.

[25] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014: Springer, pp. 740-755.

[26] C. Silva, K. Clifton, and R. Moeckel, "Observational method and coding framework for analyzing the functionality of unprotected bicycle lanes," *Transportation Research Procedia,* vol. 41, pp. 559-571, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.trpro.2019.09.100.

[27] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.