Detection and Tracking of Accessibility Challenges

Camden Cummings[1], Anurag Ghosh[2], and Christoph Mertz[2]

***Abstract*—Buses can be difficult for people with accessibility challenges to use. By finding clips of people using buses, these clips can be analyzed, and common issues with bus design can be identified. The work describes such a model, used to find images of people with accessibility challenges. Using tracking algorithm DeepSORT, we tracked them through a series of images to create video clips. The model was trained using MMDetection, and was based on Mask2Former. Because of the lack of real world accessibility challenge datasets, it was trained on a combination of real data, and simulated data. The work presents a new method of evaluating public transportation for people with accessibility challenges.**

***Index Terms*—Machine learning, Computer vision, Accessibility**

I. Introduction

FROM parents with strollers to people using canes, buses pose a challenge for many of those who use them. At the same time, using public transportation is largely better for the environment and cheaper than using a private vehicle [1]. At the same time, people with disabilities are more reliant on public transportation [2]. Modifying vehicles for accessibility can be prohibitively expensive, and many ride-sharing vehicles are not equipped. However, the needs of people with disabilities are not met by current public transportation systems [2]. In one study, 26% of the participants reported that "routes to stops and stations were inaccessible" [2]. For example, 53% of people with visual impairments reported that drivers did not call out stops [2]. These challenges can discourage use of public transportation - directly because of public transportation barriers, some people with disabilities do not leave their homes [3]. Anecdotally, the force that it takes to move bus seats down has been mentioned as an issue for some people with issues with grip strength. Making buses that are designed for people with accessibility challenges in mind removes barriers to their use, and in so doing, makes buses that are more likely to be used.

Prior work has largely focused on interviewing people with accessibility challenges. While user studies can provide some information about people's challenges with buses, this self-reported data is reliant on what participants notice, or think to mention. By studying people in the environment using buses, an observer can look for patterns in the challenges people face. Over the past several years, the NavLab has gathered video data of people using Freedom Transit, a bus company in Pittsburgh. The bus data has seven different camera angles provided; two inside of the bus, and five outside of the bus.

VideoBecause of the amount of data, this poses an additional problem - how do you sort through the data to find video clips of people with accessibility challenges? To automatically find these clips, and analyze them, a model was created to look for images of people with accessibility challenges. Subsequently, we applied tracking techniques to track them through the videos, and create video clips that could be sorted out. The work provides a novel way to analyze the effectiveness of bus engineering for people with accessibility challenges, and is applicable for analyzing other public designs, in particular other forms of public transportation, which typically already use interior cameras.

Camden Cummings is with the Department of Robotics Engineering, Worcester Polytechnic Institute of Worcester, MA, 01609
cecummings@wpi.edu
Dr. Christoph Mertz and Anurag Ghosh are with the Robotics Institute, Carnagie Mellon University, Pittsburgh, PA 15213
{cmertz,anuraggh}@andrew.cmu.edu

II. Methods

A. Datasets

Identification of wheelchairs, and similar devices, is difficult because of the lack of large real world datasets for training machine learning models. Because of this difficulty, we supplemented our real world data with simulated data gathered from X-World.

1) Real Data: We used real world data of wheelchairs from OpenImages version 6. However, the dataset separately identified people and wheelchairs, not specifically wheelchair users - one group we were specifically trying to identify. To solve this, we assumed that the person with the closest bounding box to the wheelchair's bounding box was the person using the wheelchair. We were also interested in canes, for this classification, we used OpenImages' crutch dataset, using the same method to identify which person was using the crutch.

2) Simulated Data: X-World, by Zhang et al., is a simulated platform with image instances of wheelchairs and canes, as well as wheelchair and cane users [4]. We used the set of categories given by Zhang et al., where our identifications were nondisabled pedestrian, pedestrian using a wheelchair, pedestrian with visual impairment, person riding a vehicle, four wheel vehicle, two wheel vehicle (motorcycle, bike), wheelchair, and cane.

In addition to this data, we also had access to a set of bus data. The data was taken from cameras outside of the bus, and we used it to evaluate how well our model was working. Because of how the data was annotated, we were unable to train on this data.

In total, we there were 20928 simulated images and 961 real images in our training dataset. There were 4829 simulated

images and 50 real images in our validation dataset. Overall, our combination of simulated and real world data performed well in identification of accessibility challenges.

B. Model

We trained an instance segmentation model. Instance segmentation is when a model identifies individual instances of some type of object, rather than just identifying whether something is that type or is not. To train it, we used first used Detectron2. However, the codebase was unstable and training was inefficient. To resolve this, we switched to MMDetection, a platform that focuses on a quality codebase and efficiency in training [5]. After initially attempting training in Detectron2, we switched to MMDetection, as it trains faster than Detectron2, and has a more consistent codebase. We started out with a Mask2Former model, which performs well being trained on many datasets, and which is trained on the COCO dataset [6]. Mask2Former is largely used for providing an architecture across panoptic, instance and semantic segmentations. They evolve MaskFormer by adding a Transformer decoder [6]. COCO is a popular format and has large starting dataset for models. [7] By starting with a model that was already trained on COCO, we had to spend less time training, as the model already had some idea of objects to detect. We ended up using the 9th epoch of training, as it had the best results when run on images taken from inside and outside of the bus.

C. Tracking

Tracking was tested on two different methods, the Hungarian algorithm, and DeepSORT [8] [9] [10].

1) Hungarian Algorithm: The Hungarian algorithm is a simple method to match a set of options, in this case detections and detections from previous images. By matching them, you can track one person through a series of images. The algorithm relies on a metric to determine how objects are matched - which can be anything that informs on how to match them. By minimizing the metric for every object as much as possible, and ideal matching is established [10]. For our metric, we chose to use intersection over union, which is simple and fast to calculate, while reliably predicting bounding box closeness. By taking two bounding boxes, and dividing their intersection area by their union area, you can establish how close together they are. One limitation with the Hungarian algorithm is that, as shown in Figure 1 (top row), if a detection is made multiple times for the same person, the bounding boxes will still show up, but as two different tracks.

2) DeepSORT: To address the limitations within the Hungarian algorithm, we used to track was DeepSORT. DeepSORT is an evolution of Simple Online Realtime Tracking (SORT). SORT also relies on the techniques of the Hungarian algorithm and intersection over union, however it introduces velocity [11]. SORT does this, while also remaining real time, however, it does not handle occlusion well. DeepSORT incorporates image information to reduce this kind of mistake, as well as neighbor confusion. In DeepSORT, in contrast to the Hungarian algorithm, the bounding boxes are assumed to be referring to the same object, as shown in Figure 2 (bottom row). More importantly, DeepSORT decreases fragmentation of tracks, meaning that even when the model doesn't detect anything for several frames, the track will still be consistent [8] [9].

III. Results

The model detected people with accessibility challenges correctly 41% of the time. Additionally, we found that DeepSORT was more reliable than the Hungarian algorithm for tracking.

Fig. 1. Hungarian algorithm tracking person through the bus; in particular note that two detections are made of the same person, and the Hungarian algorithm does not recognize that they are the same person.

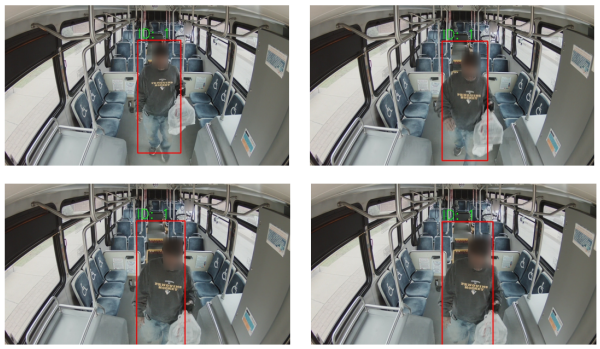

Fig. 2. DeepSORT tracking person on same set of images as above.

IV. Discussion

A. Limitations

Our model and tracking method was limited for several reasons. When it was tested on a dataset that contained people with accessibility challenges taken from cameras outside of the bus, it, as mentioned, correctly detected the person 41% of the time. However, when it was tested on the real dataset, on the same camera of this test batch, it failed to find any images. It was unclear if this was because the dataset did not contain any images of people with accessibility challenges using the bus, or if it was because the model was unable to detect the images that were in the dataset. The bus service the images

were taken from also provides a separate service specifically for people with disabilities, and as such, the number of people using the bus may be affected. Training the model also faced a number of limitations. As noted, large datasets of accessibility challenges are hard to find and vary in quality.

B. Future Work

In the future, this work could be expanded by finding instances of wheelchairs travelling from the outside cameras of the bus to the inside cameras of the bus. This is challenging because there's a space while people enter the bus where they are not visible to either camera, and continuing a correct identification through this period is difficult. Additionally, because there are two cameras inside of the bus, the results of these cameras could be compared to increase the chance of correct identification. Additional metrics could also be provided for how fast or slow people move through the bus, to see if it would be helpful for bus drivers to give more time for passengers with accessibility challenges to sit down. Another use for this data would be to study the accessibility of bus stops, and note trends in where people with accessibility challenges tend to travel. Additionally, route specific information can be gathered. For instance, when bus users with accessibility challenges tend not to use a specific bus stop that has been identified as being accessible, that bus stop can be reevaluated.

There are also several ways that in future, the model itself could be improved. Better results could be gained by annotating and training on data from the bus. Specifically, data taken from the inside of the bus could be helpful, as a smaller environment will increase occlusion, and as such, environment specific data will be helpful. Additionally, the model could be broadened by accepting lower threshold values for identification within the priority seat section. Tracking could also be improved by making sure that when people are separated from the object that they are using, they are still identified as someone with an accessibility challenge, and tracked as the same person. For example, when someone using a wheelchair moves off of their wheelchair to sit, they are still identified as the same person.

V. Conclusion

The work focuses on finding video clips of people using buses in a large dataset. The model was developed in MMDetection, based on Mask2Former. It was trained on a combination of real and simulated data, and was used in collaboration with DeepSORT. Traditionally, in researching accessibility tasks, interviews and surveys have been used to find issues in public transportation. When these methods are used, they rely on memory of what issues are present. By directly observing behavior, observations can directly be made about tasks that are difficult, but have been adapted to. Overall, the work extends prior research about people with accessibility challenges using public transport, and provides a novel way of studying difficulties.

Acknowledgment

This work was made possible by Anurag Ghosh and Dr. Christoph Mertz, as well as the rest of the NavLab. Additional thanks to Rachel Burcin, Dr. John Dolan, and the rest of RISS program for making the program possible. Special thanks to all of my fellow scholars.

References

[1] T. Xia, Y. Zhang, S. Crabb, and P. Shah, "Cobenefits of replacing car trips with alternative transportation: a review of evidence and methodological issues," Journal of environmental and public health, 2013. https://doi.org/10.1155/2013/797312

[2] J. L. Bezyak, S. A. Sabella, and R. H. Gattis, "Public Transportation: An Investigation of Barriers for People With Disabilities," Journal of Disability Policy Studies, 2017, 28(1), 52–60. https://doi.org/10.1177/1044207317702070

[3] "Transportation difficulties keep over half a million disabled at home," Bureau of Transportation Statistics and U.S. Department of Transportation, 2003. https://www.bts.gov/archive/publications/special_reports_and_issue_briefs/issue_briefs/number_03/entire

[4] J. Zhang, M. Zheng, and M. Boyd Eshed Ohn-Bar. "X-World: Accessibility, Vision, and Autonomy Meet" 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9742-9751, doi: 10.1109/ICCV48922.2021.00962

[5] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. Change Loy, and D. Lin. "MMDetection: Open MMLab Detection Toolbox and Benchmark," Computing Research Repository, 2019, http://arxiv.org/abs/1906.07155

[6] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1280-1289, doi: 10.1109/CVPR52688.2022.00135.

[7] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision, 2014. https://doi.org/10.1007/978-3-319-10602-1_48

[8] N. Wojke and A. Bewley, "Deep Cosine Metric Learning for Person Re-identification," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018 pp. 748-756. doi: 10.1109/WACV.2018.00087

[9] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 3645-3649, doi: 10.1109/ICIP.2017.8296962.

[10] H. W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955. https://doi.org/10.1002/nav.3800020109

[11] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 3464-3468, doi: 10.1109/ICIP.2016.7533003.