A Computer Vision Framework for Human User Sensing in Public Open Spaces

Peng Sun
patcivil@umich.edu
University of Michigan
Ann Arbor, Michigan

Rui Hou
rayhou@umich.edu
University of Michigan
Ann Arbor, Michigan

Jerome P. Lynch
jerlynch@umich.edu
University of Michigan
Ann Arbor, Michigan

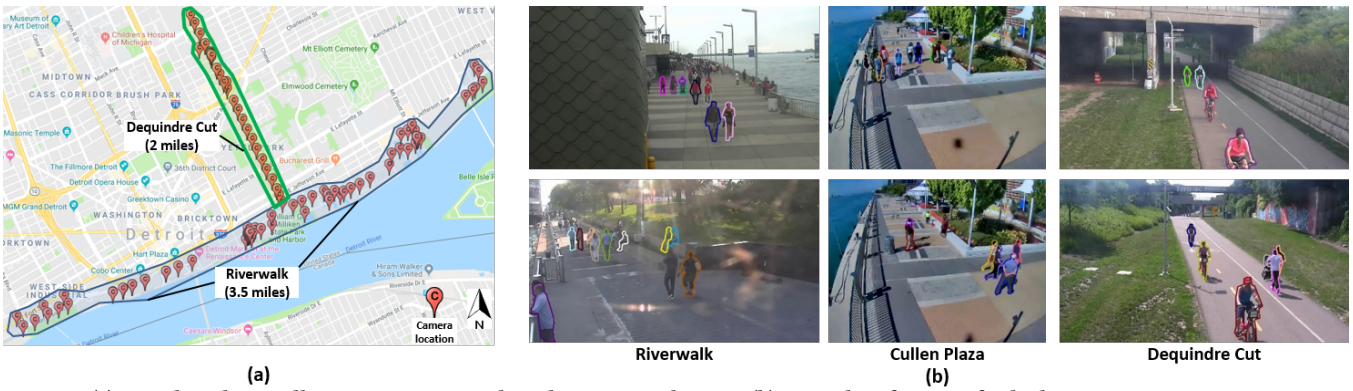


Figure 1: (a) Distributed surveillance camera network in the DRFC park space; (b) examples of camera feeds showing user activities at various park locations.

ABSTRACT

The field of Urban design considers how people utilize public open spaces (POS) when designing spaces such as parks, plazas, and streets. Current methods of observing public space use rely on visual observation which consumes much time and effort to detect users' physical activities in large POS; these methods also only provide qualitative observations of how patrons behave in these areas. Active sensors, such as wearable sensors and smart phones with GPS tracking capabilities, have high costs and cannot sense all users in a POS (namely, such sensors are "blind to" those without wearable sensors). Therefore, it is appealing to make use of video data from pre-installed surveillance cameras in POS to extract POS use information from video using computer vision methods. This paper proposes a sensing framework based on computer vision to measure human activities in POS. As part of the study, an extensively labeled datset of people and their activities in POS (termed OPOS) is used to train detectors. A case study of the proposed framework is presented using security camera feeds from a greenway at the Detroit Riverfront. The $AP^{0.50}$ results of the trained detector are 96.3% for *pedestrian* detection and 96.5% for *cyclist* detection, respectively. These results show such an approach can reliably track patrons in parks to ascertain their behavior and to inform future POS improvements.

DFHS'19, November 10, 2019, New York, NY, USA

ACM ISBN 978-1-4503-7007-3/19/11…$15.00
https://doi.org/10.1145/3360773.3360880

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **Object detection**.

KEYWORDS

Computer vision, deep learning, activity recognition, public open space, urban planning

ACM Reference Format:
Peng Sun, Rui Hou, and Jerome P. Lynch. 2019. A Computer Vision Framework for Human User Sensing in Public Open Spaces. In *The 1st ACM International Workshop on Device-Free Human Sensing (DFHS'19), November 10, 2019, New York, NY, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3360773.3360880

1 INTRODUCTION

Public open spaces are essential landscape elements used in the design of cities. POS provide venues for urban activities such as social and physical activities. Studies [3, 6, 27] show that physical activities promoted by the availability of inviting POS substantially reduce the risk of chronic diseases in communities. POS like parks support a healthy lifestyle including offering access to POS for exercise and for convening with nature. As a result, significant efforts are also placed on ensuring such spaces are well designed to offer an inviting atmosphere to encourage public use. Due to these benefits, POS like parks, green ways, and plazas are widely adopted by urban designers when efforts are aimed at driving the transformation of once economically depressed cities into lively urban environments.

In 2013, the Detroit Riverfront Conservancy (DRFC) was incorporated with the mission of restoring the international riverfront area in Detroit. The newly renovated green lands, plazas, and pavilions of the DRFC are connected along the Detroit River. The first

phase of the DRFC transformation project was the creation of a three-and-a-half mile walkway along the east end of the riverfront. This walkway connects Ralph C. Wilson Centennial Park in midtown to Gabriel Richard Park located to the east of midtown. An additional greenway was developed running north from the river to Eastern Market called the Dequindre Cut which is two miles long. The DRFC park area now attracts approximately three million visitors annually. The DRFC aims to perform post-occupancy evaluation (POE) [24] and to use POE insights to inform future rounds of renovation and development. However, the methods available to measure the patron usage of DRFC spaces are largely manual, such as visually counting and mapping users, and doing test walks [5]. Hence, an automatic method of people sensing is needed to study the utilization of POS. The extensive camera network installed in the DRFC park (Fig. 1) will be used.

In this study, an automatic user sensing framework is proposed by using surveillance cameras and deep learning. First, a brief review of the existing research on human sensing in POS is presented. Second, a cyber-physical-social sensing (CPSS) framework is proposed by using a surveillance camera network and convolutional neural networks (CNN) for automating the detection of human users. Third, a detection model is trained on a custom dataset with the detection performance evaluated. In the end, a case study is demonstrated using the framework to perform user detection and mapping in a number of POS in the DRFC region.

2 HUMAN SENSING IN PUBLIC OPEN SPACES

For many years, researchers have been working on human activity recognition (HAR) using different types of sensing methods. The sensors can be installed on humans (where sensors are mobile and follow subjects) or installed within POS where sensors collect data once subjects enter the POS (e.g. passive infrared sensors); such sensors are static and embedded within the POS environment. Within the studies using active sensors for human sensing, wearable sensors (e.g. watches, trackers) tend to dominate. Multiple types of wearable sensors have been investigated to detect human activities, to locate subjects, and to monitor subject health conditions. For example, researchers [15, 23, 32] have attached multiple inertial measurement sensors (e.g. accelerometers, gyroscopes, and magnetometers) to subjects to measure their motion attributes and to recognize their activities by processing motion data. However, wearable sensors are intrusive to users [14] and can only sense those wearing the sensors. Among the passive sensors used for human sensing, there are studies that use geophones [22], LiDAR [31], infrared sensors [13], and vision-based sensors [2] to track people. Traditional computer vision (CV) methods using cameras usually rely on a few visual features extracted from image making them difficult to achieve robust people detection. In the past few years, researchers [8, 18] have started to use deep-level features of images to extract a high-level representation of image features that allow for human activity recognition.

In urban planning and design, manual observation remains the primary approach to study POS (either directly or through video recordings [29]) with pedestrian movement usually mapped by hand. With the development of CV-based sensing technology, there are emerging studies [1, 11, 12] using CV in detecting humans in both open and enclosed spaces. However, few studies have been focused on measuring the usage of POS. A study reported in [30] employed a computer vision-based method to measure human activity in a POS. The people detection method was based on background subtraction and blob detection. The detection robustness and accuracy is presumed to suffer from adopting low-level features of images. In contrast, the goal of this paper is to perform automatic, robust people sensing in POS using deep learning methods.

3 HUMAN SENSING FRAMEWORK

3.1 Cyber-Physical-Social Systems (CPSS)

Cyber-physical systems (CPS) are an emerging class of systems consisting of a physical system coupled with sensing and/or actuation systems and computing [19]. CPS applications have ranged from autonomous cars to smart grids. While humans may be end-users of CPS, they are often not explicitly included in the design and operation of CPS platforms. A cyber-physical-social system (CPSS) is a CPS that considers human factors as part of the system [28]. CPSS consists of not only a CPS but also human interaction where human knowledge, mental capabilities, and sociocultural elements are all key features of the CPSS performance [20].

In this study, a CPSS architecture is proposed to measure POS utilization for improving the space as shown in Fig. 2. In the physical space, POS users (e.g. pedestrians, cyclists, skaters) are captured by pre-installed surveillance cameras in real time. Video streams of images are processed by a pre-trained detection model on an edge-computing device (e.g. CPU or GPU) for user sensing (e.g. detection, localization, and tracking); processed results and/or image-based data is shared with other cameras through a communication network. The processed data will also be transmitted to the cloud for access to computing (e.g. prediction of possible recurrence of one user in another camera view) or data management (e.g. dataset augmentation or video storage for special events). The processed data associated with POS use can be shared with the social system layer, which includes the community, urban designers and managers of the POS. The processed results representing the utilization of the POS can provide insightful information to POS designers to make informed decisions for future designs. The automatic sensing framework can provide a quick assessment method to quantify the use of a POS allowing for improved future investments.

A large surveillance camera network consisting of 100 cameras has been installed by the DRFC to monitor and ensure the safety of the entire POS the DRFC manages. The locations of the surveillance

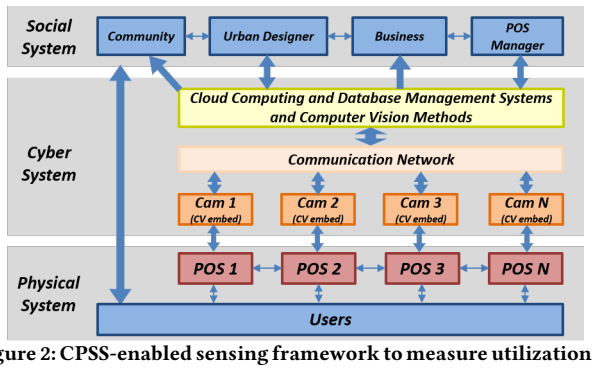


Figure 2: CPSS-enabled sensing framework to measure utilization of public open spaces (POS) with computer vision and deep learning.

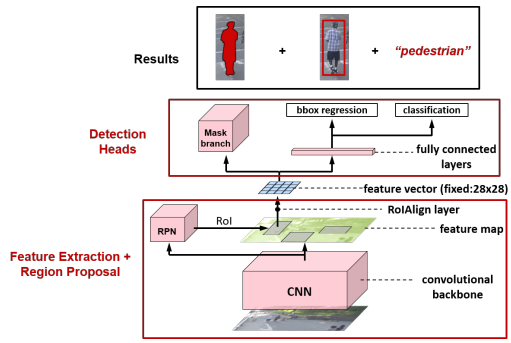


Figure 3: Schematics of the Mask R-CNN detection model for human user detection.

cameras are denoted by red markers in Fig. 1a. The surveillance cameras can capture different user activities at various locations as shown in Fig. 1b.

3.2 Deep Neural Networks for Human Detection

Deep Neural Networks (DNN) are utilized to identify and locate objects in images or video frames. There are two types of DNN-based detection methods: single-stage (i.e. non-region based) and two-stage (i.e. region based) methods. Region-based detection models (e.g. Faster R-CNN [26]) rely on region proposal networks (RPN) [26] to estimate bounding boxes (bbox) on the feature maps which are extracted by convolutional neural networks (CNN). Due to the region proposal stage, region-based detection methods consume more computational resources and take more time to execute. In contrast, non-region based models (e.g. YOLO [25], RetinaNet [17]) perform object detection without a separate region proposal step which reduces their computational demands. The past few years have led to significant improvements in region-based detectors (especially Mask R-CNN [9]) that they now often outperform single-stage detectors on the speed-vs-accuracy trade-off curve [4].

Mask R-CNN is a region-based detection method that provides a richer set of information of a detected object with an additional instance segmentation. Hence, a Mask R-CNN detector model (Fig. 3) is utilized in this study for user detection and activity recognition in POS. The CNN part of the model extracts multiple feature maps from an image while the RPN generates regions of interest by sliding over the processed feature maps. The feature maps can be generated from a feature pyramid network (FPN) [16]. In this study, ResNet50-FPN (50 layered ResNet[10] with FPN) is chosen as the CNN backbone to extract feature maps from an input image. The weights of the RPN will be updated in parallel with the CNN backbone during the model training process.

4 EXPERIMENTS AND RESULTS

4.1 OPOS Dataset and Model Training

A custom dataset for POS studies termed OPOS ("Objects in Public Open Spaces") is created and includes 7826 annotated images that are collected from DRFC surveillance cameras. The weights of the backbones were first pre-trained on the ImageNet-1K dataset and then on the coco_2017_train dataset (pre-trained weights are obtained from the Detectron website [7]). The training and test sets of the OPOS dataset are split by the ratio of 9:1. The pre-trained weights are fine-tuned on the OPOS dataset using the maskrcnn-benchmark platform [21] implemented on an NVIDIA 1070 GPU. The training schedule includes 90k iterations and the fine-tuning process on the OPOS dataset consumes 14.8 hrs.

4.2 Detection Performance Evaluation

The detection task requires detecting 11 object classes (including 6 classes of people: *pedestrian*, *cyclist*, *peopleother*, *sitter*, *scooterer*, *skater*) which are a subset of all 15 classes in the OPOS dataset. The trained detector is evaluated on the test dataset consisting of 783 images. In the rest of the study, AP is referred to as bbox AP because the segmentation AP values are very close to the bbox AP values. The detailed performance of the detection model is demonstrated in Table 1. The AP results at IoU=0.50 ($AP^{0.50}$, where IoU means intersection over union) for *pedestrian* and *cyclist* (the most two common classes) are as high as 96.3% and 96.5%, respectively. The mean average precision of overall objects ($mAP^{0.50}$) is 87.9% demonstrating a satisfactory detection performance.

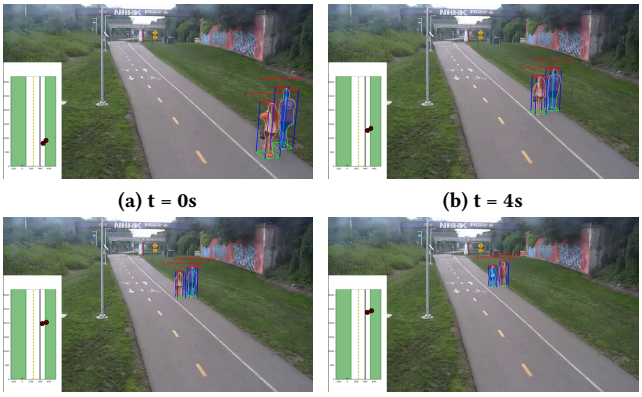
4.3 Human Detection on the Dequindre Cut

The mapping and 3D bbox estimation modules are built on the detection model to achieve detection and mapping tasks (i.e. placing people in a georeferenced system). Camera parameters are calibrated and the assumption of flat ground is adopted. The steps are as follows: (1) detecting users where users are located and segmented on images; (2) extracting pixel location with the location of the bottom pixel (corresponding to feet) and top pixel (corresponding to head) are retrieved (denoted as pink dots in Fig. 4); (3) mapping users using 2D image to place users in a 3D world space; (4) estimating 3D bbox where the horizontal sizes of a *pedestrian* are assumed fixed (i.e. w=60cm and d=50cm) and the height is optimized by a few trial calculations with geometry constraints (i.e. trial height varies from 1.5m to 2.0m with a step size of 0.05m).

The locations of human users at a section of the Dequindre Cut (a pedestrian path in the riverfront area as shown in Fig. 1) are projected to a 2-D road map (width=4.5m, length=32m). A camera is located 0.8 m from the left edge of the road. As shown in Fig. 4, the user mapping and 3d bbox estimation of each detected user can be obtained during a period of time. For example, the height of the *pedestrian* on the right side is estimated as 195cm at t = 0s (detected x=5.42m, y=8.90m, Fig. 4a), as 190cm at t = 2s (detected x=5.33m, y=11.15m), as 185cm at t = 4s (detected x=5.21m, y=13.61m, Fig. 4b), as 185cm at t = 6s (detected x=5.23m, y=15.70m), as 185cm at t = 8s (detected x=5.48m, y=17.78m), as 185cm at t = 10s (detected x=5.31m, y=20.00m, Fig. 4c), as 175cm at t = 12s (detected x=5.20m, y=21.77m), and as 180cm at t = 14s (detected x=5.35m, y=24.13m,

AP Metrics	Most common ppl. cls.			Ppl.-all	Overall
	ped.	*cycl.*	*ppl.other*		
$AP^{0.50}$	**96.3%**	**96.5%**	74.1%	89.2%	87.9%
$AP^{0.75}$	92.6%	95.4%	59.6%	78.5%	81.3%
$AP^{0.50:0.05:0.95}$	74.7%	81.2%	53.4%	69.4%	66.2%
AP^{sm}	55.2%	62.8%	24.1%	53.5%	57.6%
AP^{med}	75.3%	80.9%	52.8%	71.3%	67.8%
AP^{lg}	81.2%	86.7%	80.0%	73.0%	77.4%

Table 1: Details of people detection performance (bbox) of Mask R-CNN model.



(a) t = 0s (b) t = 4s

(c) t = 10s (d) t = 14s

Figure 4: Examples of user detection and 3D bbox estimation at the Dequindre Cut.

Fig. 4d), respectively. Hence, when a user is detected in near-field, mid-field and far-field (>20m), the estimations of the height by the proposed method are robust and the localization is very accurate. The processed spatio-temporal data of human detection are formatted and stored in GIS map layers so that the information can be used for other urban studies carried out using GIS analysis.

5 CONCLUSION AND FUTURE WORK

In this paper, a CPSS-enabled user sensing framework is presented along with a baseline DNN-based detection model. A case study of user detection and mapping on a park greenway in Detroit is demonstrated as well. The detection results of human activity in the POS reveals that the baseline detector has a bbox $mAP^{0.50}$ of 87.9% for overall objects, and a $mAP^{0.50}$ of 89.2% for the *"people"* super-category. The $AP^{0.50}$ for the most two common people classes (*pedestrian* and *cyclist*) that appear in the Detroit Riverfront are 96.3% and 96.5%, respectively. In the future, the study would serve as a stepping stone to build other sensing modules (e.g. counting and tracking) that are associated with urban planning studies.

ACKNOWLEDGMENTS

Thanks to Hao Zhou and Hsing-Yi Song for the help of camera calibration. This work is supported by the National Science Foundation (NSF) under grant #1831347. Additional support was provided by the Michigan Institute for Data Science (MIDAS).

REFERENCES

[1] Dan Barnes, Will Maddern, and Ingmar Posner. 2017. Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. In *2017 IEEE International Conference on Robotics and Automation*. IEEE, 203–210.
[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.
[3] Jacinta Francis, Lisa J Wood, Matthew Knuiman, and Billie Giles-Corti. 2012. Quality or quantity? Exploring the relationship between public open space attributes and mental health in Perth, western Australia. *Social Science & Medicine* 74, 10 (2012), 1570–1577.
[4] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. 2019. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353* (2019).
[5] Jan Gehl and Birgitte Svarre. 2013. *How to study public life*. Island press.
[6] Billie Giles-Corti, James F Sallis, Takemi Sugiyama, Lawrence D Frank, Melanie Lowe, and Neville Owen. 2015. Translating active living research into policy and practice: One important pathway to chronic disease prevention. *Journal of Public Health Policy* 36, 2 (2015), 231–243.
[7] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. https://github.com/facebookresearch/detectron.
[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8359–8367.
[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.
[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[11] Houman Hediyeh, Tarek Sayed, Mohamed H Zaki, and Greg Mori. 2014. Pedestrian gait analysis using automated computer vision techniques. *Transportmetrica A: Transport Science* 10, 3 (2014), 214–232.
[12] Stefan Hoermann, Martin Bach, and Klaus Dietmayer. 2018. Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling. In *2018 IEEE International Conference on Robotics and Automation*. IEEE, 2056–2063.
[13] Seiichi Honda, Ken-ichi Fukui, Koichi Moriyama, Satoshi Kurihara, and Masayuki Numao. 2007. Extracting human behaviors with infrared sensor network. In *2007 Fourth International Conference on Networked Sensing Systems*. IEEE, 122–125.
[14] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. 2017. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognition* 61 (2017), 295–308.
[15] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials* 15, 3 (2012), 1192–1209.
[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.
[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Cmputer Vision*. 2980–2988.
[18] Jiaying Liu, Yanghao Li, Sijie Song, Junliang Xing, Cuiling Lan, and Wenjun Zeng. 2018. Multi-modality multi-task recurrent neural network for online action detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2018).
[19] Yang Liu, Yu Peng, Bailing Wang, Sirui Yao, and Zihe Liu. 2017. Review on cyber-physical systems. *IEEE/CAA Journal of Automatica Sinica* 4, 1 (2017), 27–40.
[20] Zhong Liu, Dong-sheng Yang, Ding Wen, Wei-ming Zhang, and Wenji Mao. 2011. Cyber-physical-social systems for command and control. *IEEE Intelligent Systems* 26, 4 (2011), 92–96.
[21] Francisco Massa and Ross Girshick. 2018. maskrcnn-benchmark: Fast, modular reference implementation of instance segmentation and object detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark.
[22] Mostafa Mirshekari, Shijia Pan, Jonathon Fagert, Eve M Schooler, Pei Zhang, and Hae Young Noh. 2018. Occupant localization using footstep-induced structural vibration. *Mechanical Systems and Signal Processing* 112 (2018), 77–97.
[23] Lionel M Ni, Dian Zhang, and Michael R Souryal. 2011. RFID-based localization and tracking technologies. *IEEE Wireless Communications* 18, 2 (2011), 45–51.
[24] Wolfgang FE Preiser, Edward White, and Harvey Rabinowitz. 2015. *Post-occupancy evaluation*. Routledge.
[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.
[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.
[27] James F Sallis, Myron F Floyd, Daniel A Rodríguez, and Brian E Saelens. 2012. Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* 125, 5 (2012), 729–737.
[28] Fei-Yue Wang. 2010. The emergence of intelligent enterprises: From CPS to CPSS. *IEEE Intelligent Systems* 25, 4 (2010), 85–88.
[29] William Hollingsworth Whyte. 1980. *The social life of small urban spaces*.
[30] Wei Yan and David A Forsyth. 2005. Learning the behavior of users in a public space through video tracking. In *2005 Seventh IEEE Workshops on Applications of Computer Vision*, Vol. 1. IEEE, 370–377.
[31] Zhi Yan, Tom Duckett, and Nicola Bellotto. 2017. Online learning for human classification in 3d lidar-based tracking. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 864–871.
[32] Allen Y Yang, Sameer Iyengar, Shankar Sastry, Ruzena Bajcsy, Philip Kuryloski, and Roozbeh Jafari. 2008. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.