IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 6, 2019, accepted December 2, 2019, date of publication December 6, 2019,
date of current version December 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2958255

Practical Automated Video Analytics for Crowd Monitoring and Counting

KANG HAO CHEONG[1,2], (Member, IEEE), SANDRA POESCHMANN[3], JOEL WEIJIA LAI[1,2], JIN MING KOH[1], U. RAJENDRA ACHARYA[4], SIMON CHING MAN YU[5], AND KENNETH JIAN WEI TANG[1]

[1]Science and Math Cluster, Singapore University of Technology and Design (SUTD), Singapore 487372
[2]SUTD-MIT International Design Centre, Singapore 487372
[3]Engineering Cluster, Singapore Institute of Technology, Singapore 138683
[4]Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Singapore 599489
[5]Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Kang Hao Cheong (kanghao_cheong@sutd.edu.sg)

This work was supported by the SUTD-MIT International Design Centre (IDC) under Grant IDG21900101 and Grant IDIN19001.

ABSTRACT Video surveillance is gaining popularity in numerous applications, including facility management, traffic monitoring, crowd analysis, and urban security. Despite the increasing demand for closed-circuit television (CCTV) and related infrastructure in public spaces, there remains a notable lack of readily-deployable automated surveillance systems. In this study, we present a low-cost and efficient approach that integrates the use of computational object recognition to perform fully-automated identification, tracking, and counting of human traffic on camera video streams. Two software implementations are explored and the performance of these schemes is compared. Validation against controlled and non-controlled real-world environments is also demonstrated. The implementation provides automated video analytics for medium crowd density monitoring and tracking, eliminating labor-intensive tasks traditionally requiring human operation, with results indicating great reliability in real-life scenarios.

INDEX TERMS Crowd monitoring, counting, traffic monitoring, data analytics, background subtraction, security.

I. INTRODUCTION

Video surveillance is an integral component of modern urban security, and when coupled with computational analytics, can have greatly expanded functionality including facial recognition, motion detection, traffic and crowd monitoring, and automated hazard alarms [1]–[7]. The continued advancement in computational tools and machine learning has in principle enabled automation of a wide variety of practical analyses on image and video inputs [8]–[14]; more advanced machine intelligence systems are also increasingly capable of fulfilling traditionally human-controlled tasks that require real-time complex decisions, for instance initiating mitigation measures for severe traffic congestion or the dispatching of emergency services [15]–[18]. In general, automated surveillance eliminates the need for round-the-clock manual monitoring, thereby reducing manpower requirements [19]. This stands to yield operational cost reductions and productivity improvements.

Nonetheless, amidst the progressing state-of-the-art, integration of automated analytics in commercial video surveillance for crowd monitoring and counting is an area that can be further explored [19]; and there is at present limited literature on demonstrated effective low-cost systems for deployment. In security and management sectors, there remains a great reliance on traditional manual monitoring of CCTV footage [20], [21], and human patrols to conduct crowd monitoring and tracking. Utilizing computer vision and real-time automated analytics in replacement of manual labour not only reduces operational costs but also eliminates human errors and lapses [22]–[24] —we seek to develop a viable deployment-ready implementation in this study.

In this paper, we examine several viable approaches to automated crowd monitoring and tracking in indoor and outdoor scenarios, ultimately selecting a statistical background subtraction (BGS) scheme and a convolutional neural

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

network-based single shot detector (SSD). These methods are easily deployable in the real world. A software solution is developed for use in general public spaces, and we validate the performance of the platform through indoor controlled tests in a shopping mall, and outdoor non-controlled tests in a public transport hub with considerable human traffic. It is noted that the implementation of this video analytics system can also be applied to a broader range of scenarios—for instance, in factories to detect personnel in restricted places or in dangerous proximity to machinery, or in high-rise buildings to detect crowd densities that exceed safe thresholds for timely evacuation in case of emergencies. Time-oriented data collected from such deployments can also be logged and transmitted to a dashboard for data-informed planning and predictive analytics [25].

The structure of the paper is as follows—a technical review is first provided (Section II), followed by a discussion on the methodology employed and the development of the software solution (Section III), and finally validation test results (Section IV) and concluding remarks (Section V).

II. TECHNICAL REVIEW

We first provide an overview of object recognition frameworks, the challenges associated with achieving satisfactory performance, and the application of these frameworks in automated video analytics. A fundamental operational requirement of automated video surveillance analytics is the ability to identify and track different objects within the recorded footage, hence the need for object recognition; in the current context of crowd analysis, recognition of human subjects is critically relevant.

While visual recognition and classification of objects is intuitive to human perception, robust computational implementation is challenging [26]. Varying exposure to outdoor conditions, changing illumination levels and direction, intermittent and sustained visual obstruction, and unpredictable movement of tracked subjects must all be overcome for reliable operation of recognition systems, oftentimes with limitations on available computational power [27]–[30]. The resolution and clarity of available video footage is also typically non-ideal, limiting the effectiveness of pre-processing techniques aimed at compensating for variance in image conditions. In our context of recognition and tracking of human subjects, additional complexities arise from the wide range of possible dynamical behaviour—for instance, two persons in physical contact may be detected as a single entity, and the shape profile of a person may change drastically because of carried items or differing attire.

Advanced machine vision systems are already being developed for security- and safety-critical applications, such as driverless vehicles and autonomous drones [31]–[35]. These systems typically employ convolutional neural network (CNN)-based solutions that are trained on massive datasets of numerous modalities, including infrared and visible video input, lidar data, sound pick-ups from microphones, and navigational data from GPS or inertial guidance. Existing studies have shown excellent performance in the identification of key markers, such as lane boundaries, traffic signs, and pedestrians on systems intended for driverless vehicles [36]–[39] under a wide range of lighting and driving conditions. CNN-based image recognition has also been applied very successfully to facial identification tasks [40]–[44], achieving large reductions in error rate when compared to non-CNN methods. These types of CNN-based methods are presently employed for automated user identification and tagging systems in prominent social media platforms [45].

In general, CNN-based systems are hugely robust to changing background conditions and object appearances, but are typically computational expensive to train and run. In comparison to explicit rule-based or statistical methods, the employed CNN architectures are also more akin to black boxes and offer limited tractability—troubleshooting and tuning the systems for specific environments can therefore be challenging. There is also recent evidence attributing the efficacy of neural network deep-learning solutions to fine-tuning rather than a fundamental architectural advantage, suggesting that a properly tuned classical method may be able to achieve similar performance in certain scenarios [46]–[50].

Indeed, non-CNN methods have been deployed to perform similar tasks. A real-time system for pedestrian tracking using gray-scale images from stationary cameras has been demonstrated [51], with satisfactory robustness to visual occlusions and ambiguities in perceived subject shape profiles. The implementation relied on Gaussian-mixture foreground masking followed by contour detection through a principle component analysis (PCA) model. Numerous studies on pedestrian and traffic tracking, motion detection and analysis, and object classification using non-CNN methods have also been presented to date [30], [52]–[55], suggesting good viability in these approaches. Non-CNN methods may hence be preferred in some scenarios.

III. METHODOLOGY

The software package developed in this study comprises a video processing back-end encompassing human subject recognition and tracking, and a front-end graphical interface for operators. The software implementation is broadly discussed in Section III-A, with object recognition methods in Sections III-B1–III-B2, and lastly tracking and counting techniques in Sections III-C–III-D. A block diagram summarizing the video tracking and counting process is given in Figure 1.

A. SOFTWARE IMPLEMENTATION

Our software package is implemented on *Python* with the Open Source Computer Vision (*OpenCV*) library. *OpenCV* supports machine deep-learning frameworks, and provides image manipulation, object identification, and motion tracking tools that are greatly relevant for the development of software in our context [56], [57]. Our specific implementation assumes a pre-existing video surveillance system that writes to a centralized storage pool, from which footage may be

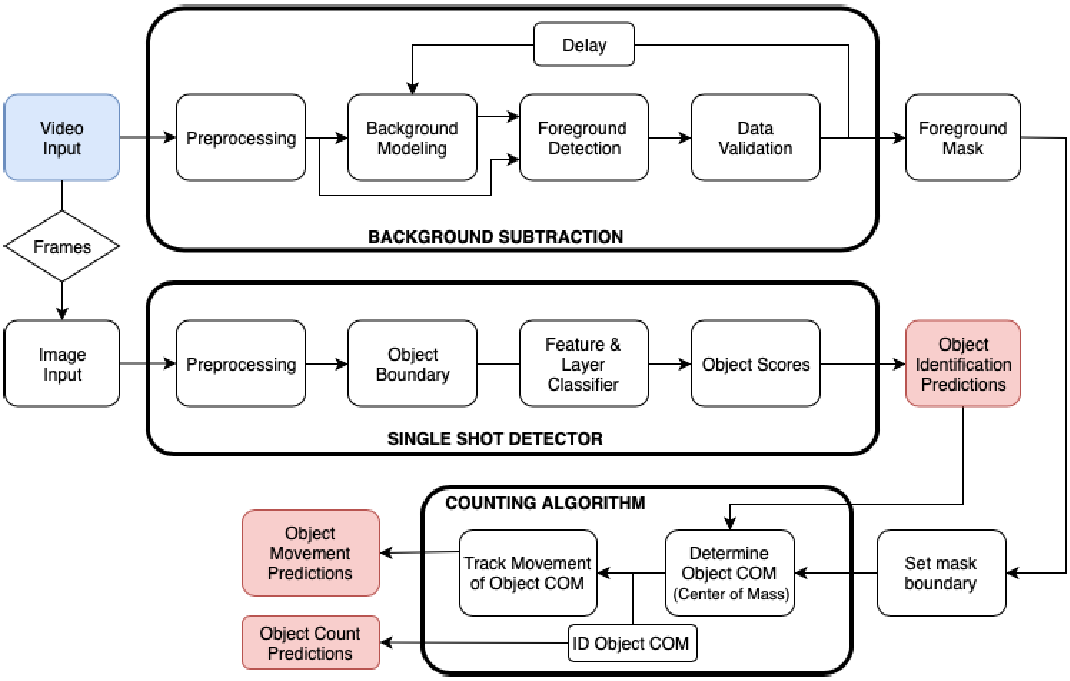


FIGURE 1. **Block diagram depicting data flow in the adopted video analytics pipeline. The input undergoes either BGS or SSD, yielding human personnel identification and count.**

pulled in real-time for analysis; as such all functionalities are developed and tested in a stream-based format. The software implementation accommodates video streams of general frame size and rate, but we use footage of 720p at 30 fps for illustrative purposes in this paper, unless otherwise stated.

B. OBJECT RECOGNITION

We examine two categories of object recognition methods in detail—background subtraction and CNN-based image classifiers. A comparison of the performance between chosen variants of these two methods in controlled and non-controlled test environments is later presented in Section IV.

1) BACKGROUND SUBTRACTION

A widely used method for detecting moving objects from a stationary camera placement is background subtraction (BGS) [58]. In general, the operation of such a method relies on a known background frame with no present objects. This background reference is then subtracted from each frame of the video footage, or subset of frames to reduce computational cost, therefore yielding frames containing only foreground objects. Appropriate contour detection or region segmentation models can then be applied to isolate distinct objects in these frames. An example illustrating the operation of BGS is presented in Figure 2. BGS-based methods are presently applied in commercial video surveillance systems for malls and public spaces.

Simplistic implementations of BGS typically suffer from limited reliability, due mostly to changing background conditions. In an outdoor environment, volatile weather, illumination changes, and reflections from surfaces on moving objects can all diminish the ability of the reference frame subtraction to separate background and foreground elements. A number of methods to overcome these problems have been utilized to date. Pre-processing of footage frames to remove glare and illumination changes can be utilized; major changes in background and ambient conditions can be detected through regression against a history of frames, and the background reference either adjusted or recaptured at opportune times; a comprehensive set of background references can also be captured *a priori* against possible ambient conditions, selections of which are subtracted from footage frames on a trial basis until a sufficiently clean output is produced. Movement patterns of detected objects across numerous frames can also be used as an additional filter against false positives—for instance, human subjects must realistically be in contact, or otherwise close proximity, with the ground at all times.

An early variant of a BGS-based object recognition framework is the Mixture of Gaussians (MOG) method introduced in 2001, utilizing a Gaussian mixture background/foreground segmentation algorithm [59], [60]. An improved version, named MOG2, was later presented, with a significant improvement being an automatic selection scheme for the number of Gaussian kernels used for each pixel, in place of the constant number of distribution kernels in the original MOG [61], [62]. As a result, MOG2 provides better adaptability to changing illumination conditions in scenes. A more recent algorithm is the GMG [63], named after its founders, which combines statistical background image estimation and per-pixel Bayesian segmentation. GMG uses the first few hundred frames of the input footage to construct

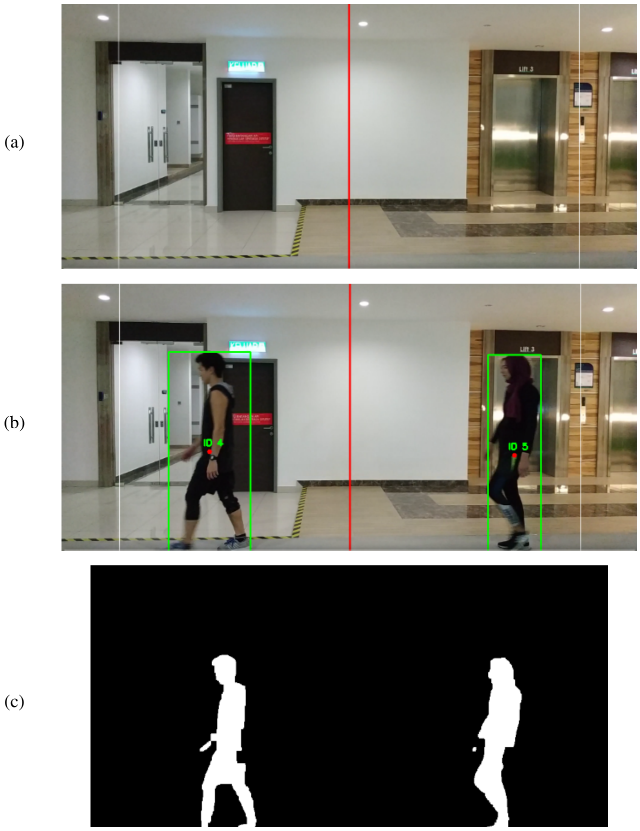


FIGURE 2. Illustration of the functioning of BGS. (a) An empty scene as the background mask; (b) a scene with human subjects in the foreground; and (c) the scene after background subtraction, separating the human subjects.

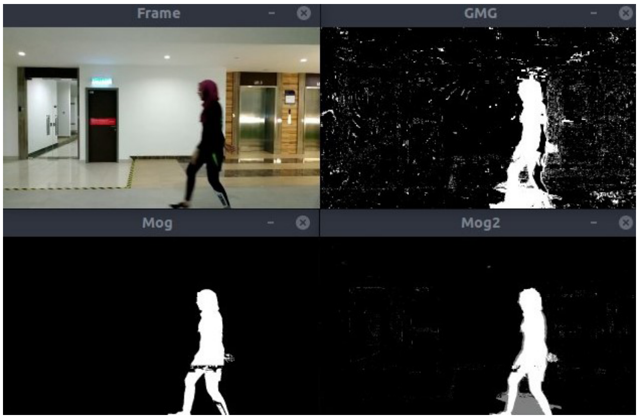


FIGURE 3. Frame comparison of the MOG, MOG2 and GMG background subtraction schemes, showing a cleaner result from MOG2.

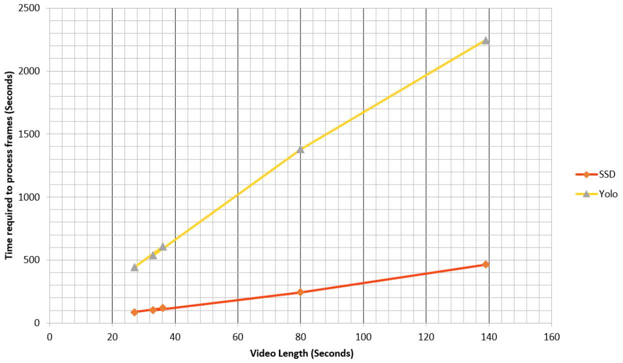


FIGURE 4. Pilot study results comparing the performance of YOLO and SSD object recognition schemes, suggesting vastly cheaper computation using SSD.

a background model, which is then used for backgound subtraction.

We note that GMG is limited in suitability for outdoor scenes with constant motion of objects, as there is often no dedicated time periods available for background capture. Deploying GMG in the intended use case of automated crowd analysis in busy public spaces is therefore not viable. On the other hand, there is evidence in existing literature that MOG2 provides good results in practice [64], [65], and is also sufficiently computationally cheap to deploy in large-scale video analysis applications. Our own pilot programme also suggests that MOG2 consistently produces better human subject identification and segmentation results than MOG and GMG in indoor use cases (example in Figure 3). We therefore employ MOG2 in our BGS-based implementation.

A three-step processing pipeline is utilized for incoming video frames. First, a brightness and colour-correction filter is applied to adjust for under- or over-exposure, or changing daylight conditions. Noise reduction is also used to improve image quality. Next, a background subtraction mask is computed and applied through MOG2. Lastly, contours on the masked image are identified through contrast segmentation, to extract the bounding boxes and positions of the human subjects. This process is computationally cheap, but face potential limitations—detection accuracy is compromised if BGS is not completely successful due to background fluctuations, and differentiation between human and non-human subjects may not be ideal.

2) CONVOLUTIONAL NEURAL NETWORKS

We consider CNN-based recognition frameworks You-Only-Look-Once (YOLO) [66], a state of the art real-time object detection system shown to be capable of identifying and classifying objects effectively, and the Single Shot Detector (SSD) [67], also a highly-established method. Both of these frameworks run the full incoming frames through a CNN in a region-wise manner to yield bounding boxes and class probabilities on identified objects. These CNNs are pre-trained on large imagery datasets. A pilot programme had been carried out to compare the computational running cost of YOLO and SSD on preliminary hardware (Figure 4), revealing that SSD is considerably faster in processing incoming frames, and is therefore more viable in achieving real-time stream-based automated video analysis with. We choose SSD as the preferred CNN-based solution. We utilize *MobileNet* supported on the deep neural network (DNN) module of *OpenCV* for SSD implementation. *MobileNet* provides pre-trained CNNs for image classification, and can robustly handle frames of different aspect ratios and sizes.

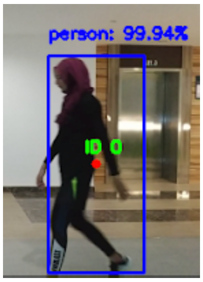


FIGURE 5. Recognition of a moving human subject through SSD.

Specifically, the model utilized in this paper is an implementation of Google's *MobileNet SSD* [68], which was initially trained on the Common Objects in Context (COCO) dataset [69]. The model was further refined on PASCAL VOC0712 [68]. A sample snapshot of the execution of this implementation on real-world footage is presented in Figure 5, on which a moving human subject with headwear is identified with $>$ 99% confidence.

C. TEMPORAL TRACKING

Analyzing incoming video streams frame-wise does not guarantee temporal continuity in the identified subjects, and therefore cannot immediately support counting functions. In order to support reliable counting of moving subjects, a temporally-consistent labelling of subjects must be achieved between frames, such that distinct objects are not misidentified as being identical (leading to under-counting), and identical objects are not misidentified as being distinct (leading to over-counting). In essence, objects undergoing movement has to be continuously tracked across all frames in which they appear.

We implement this by comparing the centroids of identified bounding boxes for each frame against those of the previous, and labelling pairs as identical on a nearest-distance basis. Centroids in the previous frame that have no matching counterpart in the current are deemed to have left the scene, and centroids in the current frame that have no match in the previous are deemed as new subjects that have entered. This is illustrated in Figure 6. To cope with subject occlusions and intermittent image quality issues, a loss-of-visibility threshold of $n_{lv} = 18$ frames is set, such that if a subject disappears from view within the scene and reappears within n_{lv} frames in a location deemed to be matching by the nearest-distance scheme, a new subject identification will not be assigned and the reappeared subject will be considered identical to the previous. A movement rate threshold can also be set (say, to the typical human running speed), such that subjects that move faster than expected between frames are identified as distinct.

D. SUBJECT COUNTING

The counting of identified subjects can be set to be scene-wise—subjects are counted the moment they enter the scene captured by the video camera, and real-time on-scene subject counts are recorded together with cumulative counts (starting from a specified time, say, the start of each day). This mode is useful if on-scene subject counts are of interest, for instance, to monitor the number of people in an enclosed space, or congestion conditions along corridors and passageways. Alternatively, counting can be set to be portal-wise, that is, subjects are added to a cumulative count only when they cross a specified boundary in the captured scene. This mode is useful to monitor crowd influx or outflux through key doorways or area perimeters, for instance, in tracking the boarding of public buses or commuter movement through security checkpoints.

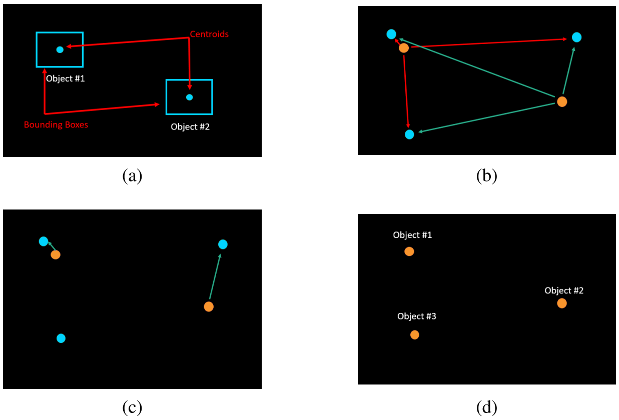
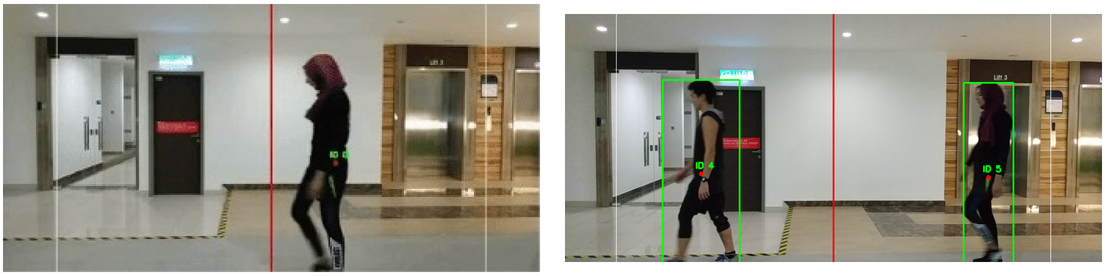


FIGURE 6. Illustration of the subject tracking mechanism. (a) Two subjects are recognized in the scene and labelled with distinct identification numbers, and the centroids of their bounding boxes are computed; (b) the subjects have moved in the next frame, and the new positions are compared against the previous; (c) identifications are made on a nearest-distance basis; (d) identification numbers are consistently applied, with a new subject deduced to have entered the scene. Images adapted from [70].

In the scene-wise mode, extremal boundaries can be set by the user on the video scene, only within which counting is active. Subject tracking is maintained throughout the entire scene regardless of boundary settings, however, so that identical subjects repeatedly crossing the set boundaries will not be misidentified as distinct instances, so long as they remain within the scene throughout. In the portal-wise mode, the user may select from a default preset of 10 regularly spaced boundary lines, or freely draw a desired boundary. Movement direction filtering can also be set, such that only subjects moving to the left or right are counted.

E. DATA PROCESSING

It is important to keep the analysis low-cost, as practical deployment hardware may be limited in speed, especially when there are multiple video streams sharing compute time. To reduce computational load, the program can be configured to perform BGS or SSD object identification only every N_0 frames; in our implementation, subject tracking and counting is set to occur every $N_0 = 6$ frames, corresponding to a $\sim$ 200 ms refresh rate. These limits were found to be sufficient in providing good tracking results for typical pedestrian traffic encountered in our test environments (Section IV), and can



(a) (b)

FIGURE 7. Video snapshots of the controlled environment tests, for (a) a constrained case of a single subject in-scene at any point in time, and (b) multiple subjects in-scene simultaneously.

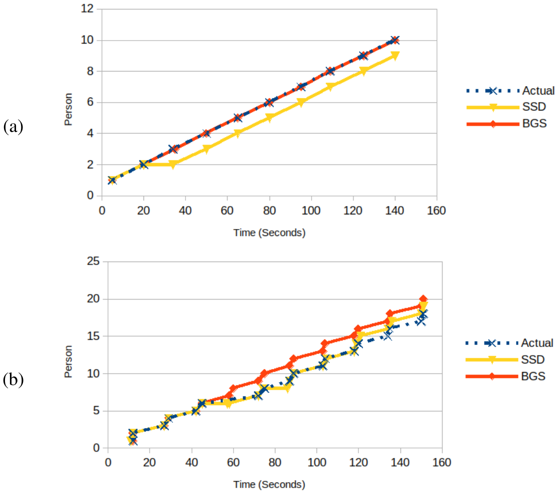


FIGURE 8. Subject counting results in the controlled environment comparing BGS and SSD methods, in (a) the constrained case of a single subject in-scene at any point in time, and (b) with multiple subjects in-scene simultaneously. The actual counts were obtained through manual counting by watching the same video footage, matched against an on-site surveyor for consistency.

be adjusted for different hardware capabilities. While such an approach effectively reduces the imposed computational load per video stream, the trade-off between practicality and accuracy stands to be further characterized; an important line of development for future work is also in studying alternative approaches that does not impact analysis frame rate.

F. COMPUTATIONAL RESOURCE

Reasonable computational cost is a requisite for viable deployability in the real-world—a considerably modest workstation was hence used for testing purposes. The workstation was a laptop equipped with an Intel I5 8250U quad-core processor at 1.6 GHz (base), 8 GB of RAM, and an Intel UHD Graphics 620 graphics processor, running the Ubuntu (Linux) operating system. On this platform, the SSD method runs in real-time on the CPU.

IV. RESULTS & DISCUSSION

The developed video analysis software was tested in two types of environments—first a controlled environment (Section IV-A), and then a non-controlled environment (Section IV-B) for validation. In these validation tests, portal-wise subject counting mode was utilized.

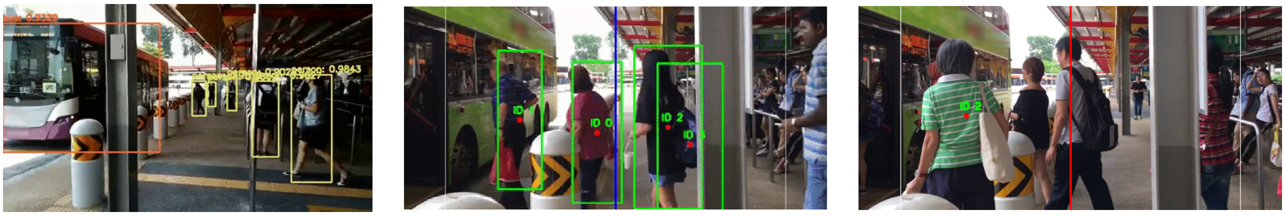
A. CONTROLLED ENVIRONMENT

An indoor area in a shopping mall, in proximity to a lift lobby, was used as an indoor controlled environment. A number of subjects, dressed in varying attire, was sent to walk across the captured space at varying speeds, ranging from a slow walk typical of the elderly to fast jogs. These subjects are of mixed genders and of varied heights. The controlled studies were conducted with steady artificial lighting and primarily constant environmental parameters; each test lasted a duration of 150 seconds with both the BGS and SSD methods, and a manual on-site count was performed simultaneously to match the results against. Snapshots of the controlled validation tests are shown in Figure 7, and subject counting results are presented in Figure 8. These results indicate satisfactory counting accuracy for both BGS and SSD methods, with BGS notably achieving perfect accuracy in the idealized single-subject scenario, but is ultimately outperformed by SSD in more realistic multiple-subject scenarios. The SSD method yielded a maximum of a single miscount in these controlled tests, suggesting good deployment viability in the significantly more demanding non-controlled outdoor environments.

B. NON-CONTROLLED ENVIRONMENT

Non-controlled validation tests were performed at a public transport hub with considerable human traffic. The camera placements were chosen for the purpose of tracking commuter volume boarding public buses at various terminals, suitable for assessing the ability of the system to cope with massive crowd surges. The tests were conducted over typical bus-boarding durations of approximately 20 seconds, with both BGS and SSD used, and a simultaneous manual count to match results against. Sample snapshots showing the recorded environments are presented in Figure 9, and a comparison of subject counting results is given in Figure 10.

It was observed that SSD yields $> 92\%$ accuracy (mean square error), illustrating the robustness of SSD in handling large crowd densities and volatile outdoor illumination conditions. The obvious failure of BGS in comparison to the controlled tests can be attributed to its inability to handle rapidly



(a) (b) (c)

FIGURE 9. Video snapshots of the non-controlled environment tests. (a) Full scene captured by the camera; (b) zoomed snapshot of boarding queue onto a public bus; and (c) zoomed snapshot of the queue at a later point in time.

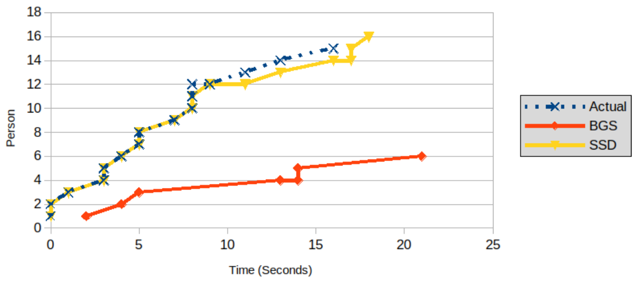


FIGURE 10. Subject counting results in the non-controlled environment, comparing BGS and SSD methods. The actual counts were obtained through manual counting by watching the same video footage, matched against an on-site surveyor for consistency.

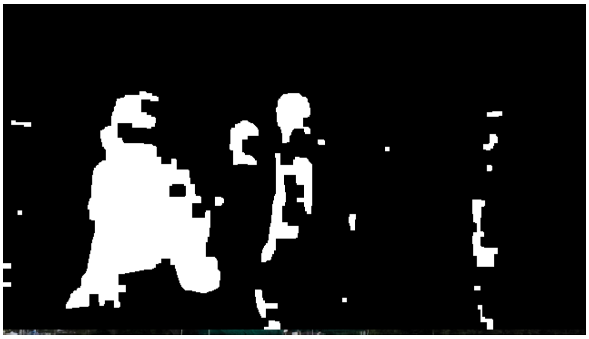


FIGURE 11. Separated foreground elements by BGS, corresponding to the input frame shown in Figure 9(c). The fragmented masking of human subjects is clearly observed.

varying backgrounds. We illustrate this in Figure 11, in which the imprecise and fragmented masking of overlapping subjects by BGS can be seen. The MOG2 implementation of BGS statistically constructs a background mask from a subsample of video frames, and is thus theoretically able to readjust for changing background conditions; but in this real-world deployment in a transport hub, there is insufficient time for such a mechanism to work as intended. With background masks of inadequate quality, multiple subjects in close proximity are frequently misidentified as a single subject, hence resulting in the severe under-counting. It is thus obvious that the SSD implementation is greatly more suitable for use, especially in places of significant human traffic.

V. CONCLUSION

In this study, we have considered a number of classical and CNN-based object recognition techniques for real-time video analytics, and have developed a software platform implementing BGS and SSD methods suitable for deployment for crowd monitoring in public spaces. Real-world validation of our solution has been carried out with both controlled and non-controlled tests, and the results strongly indicate good accuracy of the SSD system, even in outdoor conditions. This yields great confidence in expanding the deployment of the developed system into other venues.

Our proposed automated video analytics for crowd monitoring and tracking will enable significant manpower savings, especially in key security-sensitive installations such as public transport facilities and protected areas, where CCTV monitoring is oftentimes performed by human operators. Data collection of crowd density and movement can be performed more consistently and with better accuracy than otherwise achievable with manual monitoring. It is noted that the software solution developed here can accept multiple video streams from a centralized storage location, suitable for operation in facilities management or public spaces with multiple installed security cameras. Other useful applications of the current framework may include utilization in factories to detect personnel in restricted places or unsafe proximity to equipment.

Further extensions of the current framework may include layering facial identification on top of the current object recognition and tracking for enhanced surveillance capabilities, or to configure recognition for potentially dangerous items such as knives or firearms in the fight against terrorism, to enhance commuter safety and security.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

[1] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking," *IEEE Signal Process. Mag.*, vol. 22, no. 2, pp. 38–51, Mar. 2005.

[2] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," in *Proc. 37th IEEE Appl. Imag. Pattern Recognit. Workshop*, Oct. 2008, pp. 1–8.

[3] A. C. Caputo, *Digital Video Surveillance and Security*. Oxford, U.K.: Butterworth-Heinemann, 2014.

[4] P. Remagnino, G. A. Jones, N. Paragios, and C. S. Regazzoni, *Video-Based Surveillance Systems: Computer Vision and Distributed Processing*. Boston, MA, USA: Springer, 2002.

[5] T. P. Chen, H. Haussecker, A. Bovyrin, R. Belenov, K. Rodyushkin, A. Kuranoc, and V. Eruhimov, "Computer vision workload analysis: Case study of video surveillance systems," *Int. Technol. J.*, vol. 9, no. 2, pp. 109–118, 2005.

[6] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Eng. Appl. Artif. Intell.*, vol. 41, pp. 103–114, May 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197615000081

[7] L. Zhang, M. Shi, and Q. Chen, "Crowd counting via scale-adaptive convolutional neural network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1113–1121.

[8] H. Liu, S. Chen, and N. Kubota, "Intelligent video systems and analytics: A survey," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1222–1233, Aug. 2013.

[9] G. Ananthanarayanan, P. Bahl, P. Bodík, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha, "Real-time video analytics: The killer app for edge computing," *Computer*, vol. 50, no. 10, pp. 58–67, 2017.

[10] A. A. Adams and J. M. Ferryman, "The future of video analytics for surveillance and its ethical implications," *Secur. J.*, vol. 28, no. 3, pp. 272–289, 2015.

[11] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Computer Vision*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Germany: Springe, 2002, pp. 343–357.

[12] J. Barthélemy, N. Verstaevel, H. Forehead, and P. Perez, "Edge-computing video analytics for real-time traffic monitoring in a smart city," *Sensors*, vol. 19, no. 9, p. 2048, 2019.

[13] M. A. Uddin, A. Alam, N. A. Tu, M. S. Islam, and Y.-K. Lee, "SIAT: A distributed video analytics framework for intelligent video surveillance," *Symmetry*, vol. 11, no. 7, p. 911, 2019.

[14] X. Kang, B. Song, and F. Sun, "A deep similarity metric method based on incomplete data for traffic anomaly detection in IoT," *Appl. Sci.*, vol. 9, no. 1, p. 135, 2019.

[15] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges," *Computer*, vol. 44, no. 6, pp. 32–39, Jun. 2011.

[16] R. Sundar, S. Hebbar, and V. Golla, "Implementing intelligent traffic control system for congestion control, ambulance clearance, and stolen vehicle detection," *IEEE Sensors J.*, vol. 15, no. 2, pp. 1109–1113, Feb. 2015.

[17] S. Y. Kim, Y. Jang, A. Mellema, D. S. Ebert, and T. Collinss, "Visual analytics on mobile devices for emergency response," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, Oct./Nov. 2007, pp. 35–42.

[18] S. Kim, R. Maciejewski, K. Ostmo, E. J. Delp, T. F. Collins, and D. S. Ebert, "Mobile analytics for emergency response and training," *Inf. Vis.*, vol. 7, no. 1, pp. 77–88, 2008.

[19] M. Zabocki, K. Gościewska, D. Frejlichowski, and R. Hofman, "Intelligent video surveillance systems for public spaces—A survey," *J. Theor. Appl. Comput. Sci.*, vol. 8, no. 4, pp. 13–27, 2014.

[20] H. Kruegle, *CCTV Surveillance: Video Practices and Technology*. Amsterdam, The Netherlands: Elsevier, 2011.

[21] H. Keval and M. Sasse, "Man or gorilla? Performance issues with CCTV technology in security control rooms," in *Proc. 16th World Congr. Ergonom. Conf., Int. Ergonom. Assoc.*, 2006, pp. 10–14.

[22] A. Şentaş, I. Tashiev, F. Küçükayvaz, S. Kul, S. Eken, A. Sayar, and Y. Becerikli, "Performance evaluation of support vector machine and convolutional neural network algorithms in real-time vehicle type classification," in *Proc. Int. Conf. Emerg. Internetworking, Data Web Technol.* Cham, Switzerland: Springer, 2018, pp. 934–943.

[23] S. Kul, S. Eken, and A. Sayar, "Evaluation of real-time performance for BGSLibrary algorithms: A case study on traffic surveillance video," in *Proc. 6th Int. Conf. IT Converg. Secur. (ICITCS)*, 2016, pp. 1–4.

[24] S. Kul, S. Eken, and A. Sayar, "Distributed and collaborative real-time vehicle detection and classification over the video streams," *Int. J. Adv. Robot. Syst.*, vol. 14, no. 4, pp. 1–12, 2017.

[25] M. Ulvi, S. Eken, and A. Sayar, "Service oriented visual interpretation tool for time series data," *Anadolu Univ. J. Sci. Technol.-Appl. Sci. Eng.*, vol. 14, pp. 191–198, Jan. 2013.

[26] E. R. Davies, *Machine Vision: Theory, Algorithms, Practicalities*. Amsterdam, The Netherlands: Elsevier, 2004.

[27] L. Tyapi and K. Sowmya, "Real time human detection from video surveillance," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 5, pp. 4413–4417, 2015.

[28] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 66–77, Sep. 2010.

[29] C. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, and A. Hampapur, "Video analytics for surveillance: Theory and practice," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 16–17, Sep. 2010.

[30] S. A. Velastin, "CCTV video analytics: Recent advances and limitations," in *Proc. Int. Vis. Informat. Conf.* Berlin, Germany: Springer, 2009, pp. 22–34.

[31] E. D. Dickmanns, "The development of machine vision for road vehicles in the last decade," in *Proc. Intell. Vehicle Symp.*, vol. 1, Jun. 2002, pp. 268–281.

[32] B. Ranft and C. Stiller, "The role of machine vision for intelligent vehicles," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 8–19, Mar. 2016.

[33] M. Slade and F. Marmoiton, "Toward smart autonomous cars," in *Intelligent Transportation Systems*. Boca Raton, FL, USA: CRC Press, 2016, pp. 94–120.

[34] V. Moskalenko, A. Moskalenko, A. Korobov, O. Boiko, S. Martynenko, and O. Borovenskyi, "Model and training methods of autonomous navigation system for compact drones," in *Proc. IEEE 2nd Int. Conf. Data Stream Mining Process. (DSMP)*, Aug. 2018, pp. 503–508.

[35] A. A. Zhilenkov and I. R. Epifantsev, "The use of convolution artificial neural networks for drones autonomous trajectory planning," in *Proc. IEEE Conf. Russian Young Res. Elect. Electron. Eng. (EIConRus)*, Jan./Feb. 2018, pp. 1044–1047.

[36] Y. Satılmiş, F. Tufan, M. Şara, M. Karslı, S. Eken, and A. Sayar, "CNN based traffic sign recognition for mini autonomous vehicles," in *Proc. Int. Conf. Inf. Syst. Archit. Technol.*, J. Świątek, L. Borzemski, and Z. Wilimowska, Eds. Cham, Switzerland: Springer, 2019, pp. 85–94.

[37] M. Hirz and B. Walzel, "Sensor and object recognition technologies for self-driving cars," *Comput.-Aided Des. Appl.*, vol. 15, no. 4, pp. 501–508, 2018.

[38] A. Shustanov and P. Yakimov, "CNN design for real-time traffic sign recognition," *Procedia Eng.*, vol. 201, pp. 718–725, Dec. 2017.

[39] H. Wang, B. Wang, B. Liu, X. Meng, and G. Yang, "Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle," *Robot. Auton. Syst.*, vol. 88, pp. 71–78, Feb. 2017.

[40] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[41] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May/Jun. 2017, pp. 650–657.

[42] S. Guo, S. Chen, and Y. Li, "Face recognition based on convolutional neural network and support vector machine," in *Proc. IEEE Int. Conf. Inf. Automat. (ICIA)*, Aug. 2016, pp. 1787–1792.

[43] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing*, vol. 299, pp. 42–50, Mar. 2018, doi: 10.1016/j.neucom.2018.03.030.

[44] S. Balaban, "Deep learning and face recognition: The state of the art," *Proc. SPIE*, vol. 9457, May 2015, Art. no. 94570B, doi: 10.1117/12.2181526.

[45] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang, "Applied machine learning at facebook: A datacenter infrastructure perspective," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 620–629.

[46] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet," in *Proc. Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://openreview.net/forum?id=SkfMWhAqYQ

[47] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, 2015.

[48] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.

[49] R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, "Shot classification of field sports videos using alexnet convolutional neural network," *Appl. Sci.*, vol. 9, no. 3, p. 483, 2019.

[50] M. He, H. Luo, B. Hui, and Z. Chang, "Pedestrian flow tracking and statistics of monocular camera based on convolutional neural network and Kalman filter," *Appl. Sci.*, vol. 9, no. 8, p. 1624, 2019.

[51] V. Abrishami, A. Rezaee, H. Baherzadeh, and H. Abrishami, "Real-time pedestrian detecting and tracking in crowded and complicated scenario," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.

[52] A. Dore, M. Soto, and C. S. Regazzoni, "Bayesian tracking for video analytics," *IEEE Signal Process. Mag.*, vol. 27, no. 5, pp. 46–55, Sep. 2010.

[53] M. Bilal, A. Khan, M. U. K. Khan, and C.-M. Kyung, "A low-complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260–2273, Oct. 2017.

[54] R. Lucas, A. Rowlands, A. Brown, S. Keyworth, and P. Bunting, "Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 3, pp. 165–185, Aug. 2007.

[55] Z. Miao, S. Zou, Y. Li, X. Zhang, J. Wang, and M. He, "Intelligent video surveillance system based on moving object detection and tracking," in *Proc. Int. Conf. Inf. Eng. Commun. Technol.*, 2016, pp. 1–4.

[56] K. Pulli, A. Baksheev, K. Kornyakov, and V. Eruhimov, "Real-time computer vision with OpenCV," *Commun. ACM*, vol. 55, no. 6, pp. 61–69, Jun. 2012.

[57] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to OpenCV," in *Proc. 35th Int. Conv. MIPRO*, 2012, pp. 1725–1730.

[58] S. H. Shaikh, K. Saeed, and N. Chaki, *Moving Object Detection Using Background Subtraction*. Cham, Switzerland: Springer, 2014, pp. 15–23.

[59] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Boston, MA, USA: Springer, 2002, pp. 135–144.

[60] E. Grevelink, "A closer look at object detection, recognition and tracking," Intel, Santa Clara, CA, USA, Tech. Rep., 2017. [Online]. Available: https://software.intel.com/en-us/articles/a-closer-look-at-object-detection-recognition-and-tracking

[61] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. ICPR*, 2004, pp. 28–31.

[62] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 773–780, 2006.

[63] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *Proc. Amer. Control Conf. (ACC)*, 2012, pp. 4305–4312.

[64] T. Trnovszký, P. Sýkora, and R. Hudec, "Comparison of background subtraction methods on near infra-red spectrum video sequences," *Procedia Eng.*, vol. 192, pp. 887–892, Jun. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877705817327005, doi: 10.1016/j.proeng.2017.06.153.

[65] I. Benraya and N. Benblidia, "Comparison of background subtraction methods," in *Proc. Int. Conf. Appl. Smart Syst. (ICASS)*, 2018, pp. 1–5.

[66] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[67] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[68] *Chuanqi305 Mobilenet-SSD*. Accessed: Oct. 29, 2019. [Online]. Available: https://github.com/chuanqi305/MobileNet-SSD

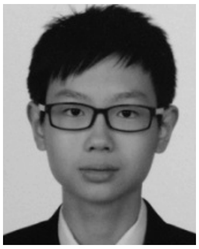
[69] *N/A, Common Objects in Context*. Accessed: Oct. 29, 2019. [Online]. Available: http://cocodataset.org/#home

[70] A. Rosebrock. (2018). *Simple Object Tracking With OpenCV*. Accessed: Aug. 8, 2019. [Online]. Available: https://www.pyimagesearch.com/2018/07/23/simple-object-tracking-with-opencv/

KANG HAO CHEONG (M'18) received the B.Sc. degree (Hons.) from the Department of Mathematics and University Scholars Programme, National University of Singapore (NUS), in 2007, the Postgraduate diploma in education from the National Institute of Education, Singapore, and the Ph.D. degree from the Department of Electrical and Computer Engineering, NUS, in 2015. He was an Assistant Professor with the Engineering Cluster, Singapore Institute of Technology, from 2016 to 2018. He is currently an Assistant Professor with the Science and Math Cluster, Singapore University of Technology and Design (SUTD). He is also affiliated with the SUTD-MIT International Design Centre.

SANDRA POESCHMANN received the B.Eng. degree (Hons.) in sustainable infrastructure engineering (building services) from the Singapore Institute of Technology, in 2018. She is currently an Associate Geospatial Specialist with the Government Technology Agency (GovTech), Singapore.

JOEL WEIJIA LAI received the B.Sc. degree (Hons.) in physics with a second major in mathematical sciences from Nanyang Technological University (NTU), Singapore, in 2017. He is currently pursuing the Ph.D. degree with the Singapore University of Technology and Design (SUTD), Singapore. He stayed on NTU to develop technology enhanced learning tools for the Division of Physics and Applied Physics. He went to become a Research Assistant with SUTD.



JIN MING KOH received the NUS High School Diploma (High Distinction), in 2016. Since 2017, he has been undertaking research projects offered by K. H. Cheong. He is currently with the California Institute of Technology.

U. RAJENDRA ACHARYA received the Ph.D. degree from the National Institute of Technology Karnataka, Surathkal, India, and the D.Eng. degree from Chiba University, Japan. He is currently a Senior Faculty Member with Ngee Ann Polytechnic, Singapore. He is also an Adjunct Professor with Taylor's University, Malaysia, an Adjunct Faculty with the Singapore Institute of Technology–University of Glasgow, Singapore, and an Associate Faculty with the Singapore University of Social Sciences, Singapore. He has published more than 400 articles in refereed international SCI-IF journals (345), international conference proceedings (42), books (17) with more than 20 000 citations in Google Scholar (with H-index of 75), and Research Gate (RG) score of 47.1. He holds three patents. His major academic interests include biomedical signal processing, biomedical imaging, data mining, visualization, and biophysics for better healthcare design, delivery, and therapy. He is ranked in the top 1% of the Highly Cited Researchers for the last four consecutive years (2016–2019) in computer science according to the Essential Science Indicators of Thomson.

SIMON CHING MAN YU received the B.Eng. degree (Hons.) in ACGI and the Ph.D. degree in DIC from the Department of Mechanical Engineering, Imperial College, London, in 1987 and 1991, respectively. He is currently a Professor with the Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. He joined Nanyang Technological University, Singapore, immediately after his graduation, as a Lecturer. He became a Senior Lecturer, in 1996, and an Associate Professor, in 2000. He has been involved in the University Administration, since 2000: a Principal Staff Officer (President's Office), from 2000 to 2003, a University Council Member, from 2001 to 2003, the Vice Dean, Admission Office, from 2003 to 2006, and the Head of the Division of Aerospace Engineering, School of Mechanical and Aerospace Engineering, from 2008 to 2013. He moved over to the Singapore Institute of Technology as a Professor and a Programme Director, in 2013, to establish one of the first engineering degree programmes offered solely by the University. He has published more than 200 research articles in archive journals and conferences. He managed to secure more than SGD 50 million external grant during his tenure in the Nanyang Technological University, especially during the period as the Head of the Division of Aerospace Engineering.

KENNETH JIAN WEI TANG received the B.Eng. degree (Hons.) in sustainable infrastructure engineering (building services) and the M.Eng.Tech. degree from the Singapore Institute of Technology (SIT), in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Science and Math Cluster, Singapore University of Technology and Design.