High-level feature extraction for Crowd Behaviour Analysis: a Computer Vision approach

Alessandro Bruno$^{1[0000-0003-0707-6131]}$, Marouane Ferjani^{1}, Zoheir Sabeur$^{1[0000-0003-4325-4871]}$, Banafshe Arbab-Zavar$^{1[0000-0001-8884-7588]}$, Deniz Cetinkaya$^{1[0000-0002-1047-0685]}$, Liam Johnstone^{1}, Muntadher Sallal$^{1[0000-0002-1755-3264]}$, and Djamel Benaouda^{1}

Department of Computing and Informatics at Bournemouth University, Poole , UK
abruno@bournemouth.ac.uk; mferjani@bournemouth.ac.uk;
zsabeur@bournemouth.ac.uk; barbabzavar@bournemouth.ac.uk;
dchetinkaya@bournemouth.ac.uk; ljohnstone@bournemouth.ac.uk;
msallal@bournemouth.ac.uk; dbenaouda@bournemouth.ac.uk
https://www.bournemouth.ac.uk/

Abstract. The advent of deep learning has brought in disruptive techniques with unprecedented accuracy rates in so many fields and scenarios. Tasks such as the detection of regions of interest and semantic features out of images and video sequences are quite effectively tackled because of the availability of publicly available and adequately annotated datasets. This paper describes a use case scenario with a deep learning models' stack being used for crowd behaviour analysis. It consists of two main modules preceded by a pre-processing step. The first deep learning module relies on the integration of YOLOv5 and DeepSORT to detect and track down pedestrians from CCTV cameras' video sequences. The second module ingests each pedestrian's spatial coordinates, velocity, and trajectories to cluster groups of people using the Coherent Neighbor Invariance technique. The method envisages the acquisition of video sequences from cameras overlooking pedestrian areas, such as public parks or squares, in order to check out any possible unusualness in crowd behaviour. Due to its design, the system first checks whether some anomalies are underway at the microscale level. Secondly, It returns clusters of people at the mesoscale level depending on velocity and trajectories. This work is part of the physical behaviour detection module developed for the S4AllCities H2020 project.

Keywords: Crowd Behaviour · Computer Vision · Artificial Intelligence · Deep Learning

1 Introduction

Over the last decade, the scientific community observed a lot of progress in Artificial Intelligence and Computer Vision. Consequently, several application domains spanning object modelling, detection, segmentation, healthcare, crowd dynamics are addressed using computer vision approaches [6] [22] [20] [5]. The advent of Deep Learning [16] prompted both academics and industry to push the bar on the proposed solutions for several scenarios and use-cases. Since the introduction of AlexNet in 2012 [15], much attention has been focused on Deep Neural Networks to achieve increasingly higher accuracy rates on the topics above and tasks. Some architectures represent milestones in the deep learning literature, namely GoogleNet [24], Inception-V4 and ResNet [23], GANs [10], YOLO [18]. As the literature review shows, AI allowed achieving unprecedented accuracy rates in so many research fields, albeit some paradigms exhibit drawbacks [29]. For instance, supervised learning relies on the availability of a great deal of manually annotated data. Big-sized datasets such as ImageNet[8] come along with millions of images and the corresponding annotations, making supervised learning a suitable paradigm to perform different tasks. Generally speaking, the hand-labelling of images and video sequences is labour intensive and time-consuming. That especially applies to all those domains such as biomedical imaging, behaviour understanding, visual perception, where in-depth knowledge and expertise are required. Some object detection and segmentation tasks are easily extended to video sequences by optimising the image-related version.

Research interest in crowd behaviour analysis has grown remarkably over the last decades. As a result, crowd behaviour analysis has become a multidisciplinary topic involving psychology, computer science, physics. A crowd can be thought of as a collection of individuals showing movements that might be temporarily coordinated upon a common goal or focus of attention [2]. That could apply to both spectators and moving people. Consequently, there are three main levels at which crowds can be described: microscale, mesoscale, macroscale. At the microscale level, pedestrians are identified individually. The state of each of such individuals is delivered by position and velocity. At the mesoscale level, the description of pedestrians is still identified by position and velocity, but it is represented statistically through a distribution function. At the macroscale level, The crowd is considered as a continuum body. Furthermore, it is described with average and observable quantities such as spatial density, momentum, kinetic energy and collectiveness. This paper describes a use case scenario for crowd behaviour analysis and provides an integrated solution. The proposed solution relies on both supervised and unsupervised learning paradigms depending on the task to work out. The proposed solution has been developed within the research activities for the European Research Project S4AllCities [1]. The experiments have been carried out on the publicly available UCSD Anomaly Detection Dataset [27].

2 Related Work

One of the main goals of crowd behaviour analysis is to predict whether some unusual phenomenon takes place to ensure peaceful event organizations and minimize the number of casualties in public areas. This section summarises the scientific literature on the topic by looking into approaches relying on different principles and methodologies. The more traditional methods of crowd behaviour analysis build on the extraction of handcrafted features either to set up expert systems or to feed neural networks and classification systems. For instance, texture analysis tackles the detection of regular and near regular patterns in images [3]. Saqib et al. [21] carried out crowd density estimation using texture descriptors while conversely, some methods address crowd analytics using physics concepts and fluid dynamics as in [9]. However, images and videos in real scenarios contain nonlinearities that have to be faced efficiently for gaining accuracies in the results. [25] Some computer vision-based methods face the challenging topic by checking groups of people exhibiting coherent movements [27]. Other techniques focus on path analysis using mathematical approaches while psychologists highlighted some aspects regarding emergency and situational awareness [19]. A shared line in the methods above is the increase in demand for security measures and monitoring of crowded environments. Therefore, by zooming in on the topic, one can unearth several applications that are closely related to crowd analysis: person tracking [19], anomaly detection [28], behaviour pattern analysis [7], and context-aware crowd counting [17]. As briefly mentioned in the previous section, despite the introduction of deep learning solutions being with high accuracy rates, some open issues related to density variation, irregular distribution of objects, occlusions, pose estimation remain open in the topic of crowd analysis [14]. The following section introduces the integrated solution developed for the S4AllCities project [1].

3 Proposed Method

In this section, the proposed method is thoroughly described by highlighting the role played by each module. The overall architecture for the integrated solution is depicted in figure 1 with three main blocks: homographic projection, supervised deep learning models, unsupervised learning module. The following subsections focus on each of the steps mentioned above.

3.1 Pre-processing

The first step of the proposed integrated solution consists of planar homography to project head-plane points onto the ground-plane. As widely described by Hartley and Zisserman [11], planar homography relates the transformation between two planes (up to a scale factor). The homography matrix H has 8 degrees of freedom. That means that four matches are enough to calculate the transformation. The main goal here is to remove or correct the perspective of the given view of the pedestrian-area-overlooking camera. In the use-case scenario, at least four coordinates of pedestrians are needed. They can be easily fetched by enacting YOLOv5 until the four pedestrians are detected. Then, the approach will generate an approximation on the plane-to-plane projection depending on the average height of pedestrians in the given camera's field of view.

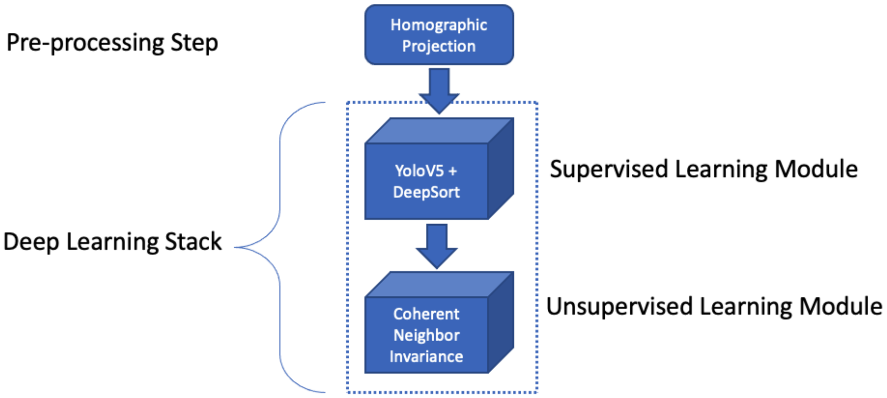
Pre-processing Step
Homographic Projection
Deep Learning Stack
YoloV5 + DeepSort
Supervised Learning Module
Coherent Neighbor Invariance
Unsupervised Learning Module

Fig. 1. Deep Learning Stack is depicted in the figure.

3.2 Supervised Deep Learning Module

Inspired by Hou et al.'s method [12] on vehicle tracking, the first of two deep learning modules sees the integration of two popular models such as YOLOv5 [13] and DeepSORT [26]. The former is one of the most accurate models for object detection. At the same time, the latter tracks down human crowd movements over video sequences, which is the extension of the popular YOLOv4 by Bochkovskiy et al. [4]. For a given frame having N pedestrians, $P(x, y)_{i=1,...,N}$ represents the i^th pedestrian' spatial coordinates. YOLOv5 is quite accurate in detecting pedestrians (see figure 2; it does not perform re-identification though. That is why it has been necessary to integrate DeepSORT, which is responsible for tracking down the pedestrians in a video sequence by assigning them a specific reference number. DeepSORT keeps trace of $P(x, y)_{i=1,...,N}$ across different times (t0, t1, ..., tn). In figures 3-4 an example referring to ID 1 pedestrian is shown. YOLOv5 returns all spatial coordinates of the pedestrians detected as a sequence of bounding boxes. They will be then ingested by DeepSORT, which runs measurement-to-track associations using nearest neighbour queries in visual appearance space (see figure 5). On top of both modules, the system is capable of retrieving the spatial coordinates, and the reference number of the pedestrians tracked across the area overlooked by a CCTV camera. The extraction of the details mentioned above is taken every second. Having timestamps, spatial coordinates and reference number allows extracting velocity and storing trajectories. A time frame Δt is taken as a reference to work out the detection of anomalies in the crowd behaviour at the microscale level. Being t_0 the initialisation time of the system, $t_0 + \Delta t$ is the earliest time where it is possible to detect any anomalies in crowds. Gaussian distributions are considered to analyse pedestrian velocity within the Δt time range. An example of trajectories out of video sequences is given in figure 6. The system evaluates anomalies as the samples that deviate from the normal distribution. The more a sample is distant from the distribution, the more likely an anomaly is within the crowd behaviour.

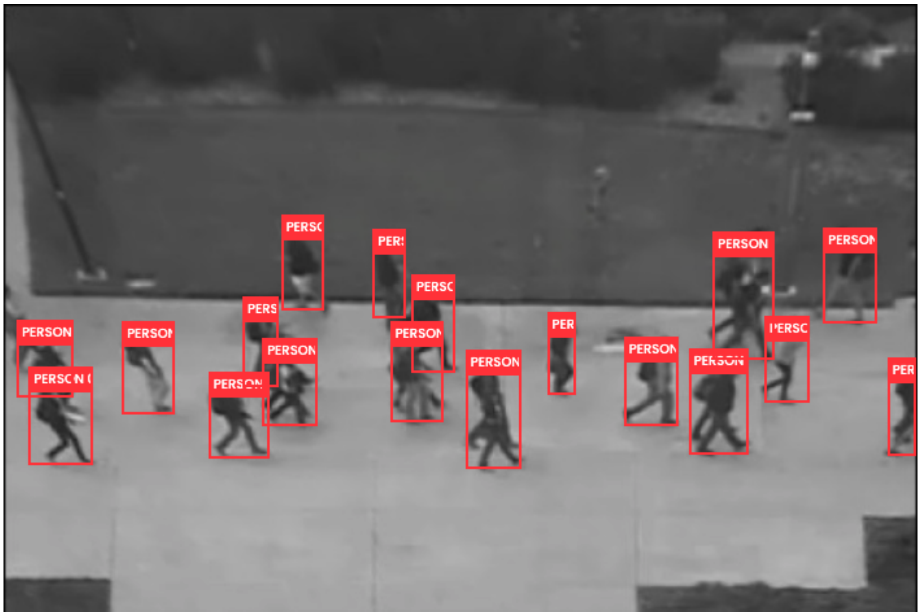

Fig. 2. An example of pedestrian detection from video frames is given above.

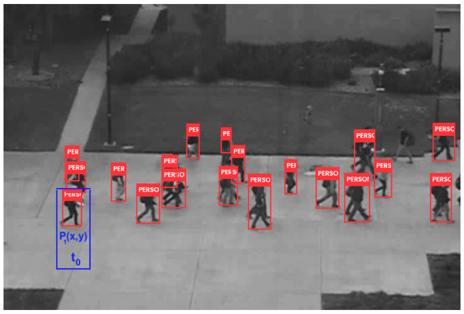

Fig. 3. Pedestrian detection at time t_0

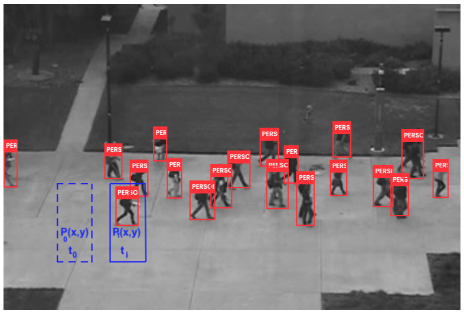

Fig. 4. Pedestrian detection at time t_1

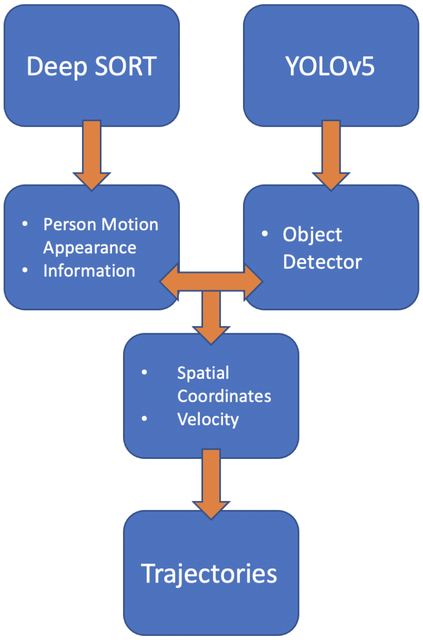

Fig. 5. The first deep learning module consists of the integration of DeepSORT and YOLOv5

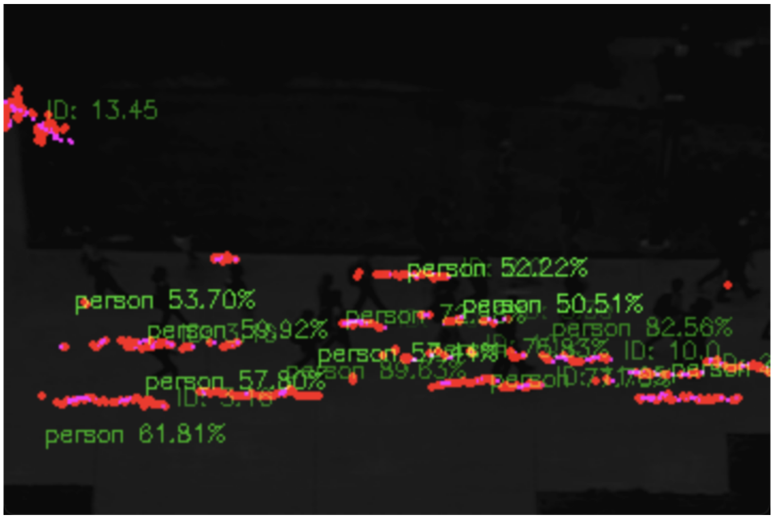

Fig. 6. The first deep learning module consists of the integration of DeepSORT and YOLOv5

3.3 Unsupervised Learning for Trajectory Clustering

Due to the advances in detection and tracking techniques, the ability to extract high-quality features of moving objects such as trajectories and velocities is now possible. These features can be critical in understanding and detecting coherent motions in various physical and biological systems. Furthermore, the extraction of these motions enables a deeper understanding of self-organized biological systems. For instance, in surveillance videos, capturing coherent movements exhibited by moving pedestrians permits acquiring a high-level representation of crowd dynamics. These representations can be utilized for a plethora of applications such as object counting, crowd segmentation, action recognition and scene understanding, etc.

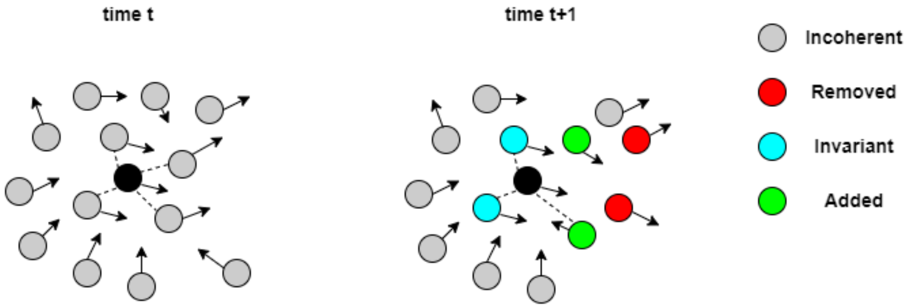

Fig. 7. An exhibition of coherent neighbour invariance. the green dots are viewed as invariant neighbors of the centered black dot (for K = 4).

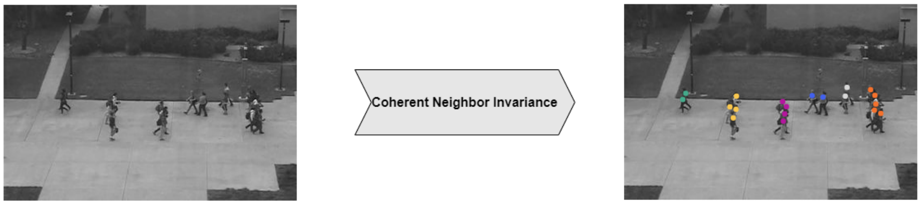

Fig. 8. coherent motion detection in action

Whilst coherent motions are regarded as macroscopic observations of pedestrians' congregational activities, these motions can be distinguished through the interaction among individuals in local neighbourhoods. Inspired from Zhou [30], the Coherent Neighbor Invariance technique is deployed to capture the coherent motion of crowd clutters. The key characteristics that establish the difference between cohesive and arbitrary movements are listed below:

- **Neighborship Invariance**: the spatial-temporal relationship among individuals is inclined to prevail overtime.
- **Velocity Correlations Invariance**: neighboring individuals exhibiting coherent movement showcase high velocity correlations.

Conversely, incoherent individuals that showcase relative independence tend to lack the mentioned properties. To illustrate the Neighborship Invariance property, Figure 7 displays the use of K nearest neighbour to highlight the emergence of global coherence in local neighborships. The equation below quantifies the velocity correlations between neighbouring individuals, which allows discerning coherent motions.

$$g = \frac{1}{d+1} \sum_{\lambda=t}^{t+d} \frac{v_\lambda^i \cdot v_\lambda^{i_k}}{\|v_\lambda^i\|^2 \cdot \|v_\lambda^{i_k}\|^2} \tag{1}$$

Where:

- g : velocity correlation between i and i_k
- v_λ^i : velocity of individual i at time λ
- $v_\lambda^{i_k}$: velocity of individual i_k at time λ
- d : duration of the experiment

4 Experimental Results

An experimental campaign has been carried out over the publicly available UCSD Anomaly Detection Dataset [27]. The dataset consists of video sequences acquired with a stationary camera overlooking pedestrian areas. The dataset offers videos with variable conditions of crowd density, and cameras' field of view. Most of videos contains only pedestrians, still anomalies are represented by bikers, skaters, small carts, pedestrian entities crossing a walkway or walking in the grass that surrounds it.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The experiments were run on five video sequences from UCSD. Two of which do not contain any anomalies, while the remaining three do. A quantitative analysis of results is conducted over the first deep learning module, which is responsible for the microscale analysis. In tables 1-2 precision and recall (see equations 2-3) for YOLOv5 and DeepSORT are reported. The second deep learning module is still currently being developed. Only qualitative results can be shown 7 to give the big picture of the consistency of clusters of people. As it can be noticed in table 1, YOLOv5 reaches high precision rates on all tests up to 0.98 while recall is penalised by some false negatives. Occlusion and overlapping cause a drop of performances on pedestrian detection. DeepSORT also achieves good precision rates even though sometimes the tracking shows some mismatch. Recall values drop by 10% on average if compared to precision. Nevertheless, the combination of the two supervised learning modules gains decent performances. As described in section 3.2, the supervised deep learning module allows the extraction of high-level features such as spatial coordinates, velocity and trajectories. On top of that, some parameters are to be fine-tuned, respectively, Δt and the distance from the normal distribution. The latter has a sample evaluated as anomaly, trigger a sort of alert to the crowd behaviour analysis system. Some fine-tuning has been necessary in order to find the right trade-off performances and computational load. δt has been set to 5 seconds, while 5 pixel/second has been selected as the distance threshold from the normal distribution of velocities.

The experiments on the automatic optimisation of the given advertisement layouts and images have been carried out on a 13-inch Mac-book Pro with 16 GB of RAM, 2.4 GHz Quad-Core Intel Core i5, Intel Iris Plus Graphics 655 1536 MB.

Table 1. YOLOv5 Precision and Recall in 5 tests over UCSD

No. of Test	Precision	Recall
Test 1	0.98	0.75
Test 2	0.93	0.72
Test 3	0.95	0.71
Test 4	0.94	0.78
Test 5	0.92	0.70

Table 2. DeepSORT Precision and Recall in 5 tests over UCSD

No. of Test	Precision	Recall
Test 1	0.85	0.74
Test 2	0.89	0.72
Test 3	0.83	0.69
Test 4	0.86	0.68
Test 5	0.87	0.72

5 Conclusions

This paper showcases the effectiveness of an integrated solution consisting of three main modules: pre-processing, supervised learning, unsupervised learning. The main goal is to perform crowd behaviour analysis by considering several variables such as velocity, spatial coordinates and trajectories. The first two have been used to detect anomalies in the test set at the microscale level. Successively, the unsupervised learning module ingests velocities and trajectories to initialise clusters of people according to cohesive movements. The microscale analysis task has been entirely carried out with supervised deep learning models such as YOLOv5 and DeepSORT. Cohesive movement-based clustering has been tackled by the Coherent Neighbour Invariance technique. Further experiments are underway to improve precision and recall rates, especially on the pedestrian tracking task. Furthermore, some other alternatives are in consideration to detect anomalies by combining physical properties like velocity and trajectories and semantic features such as objects whose only presence might represent a danger within a given environment. Furthermore, some work is to be done to adapt the method to different datasets and environments.

Acknowledgements This work is a part of the S4AllCities project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 883522. Content reflects only the authors' view and the Research Executive Agency (REA)/European Commission is not responsible for any use that may be made of the information it contains.

References

1. Smart spaces safety and security: www.s4allcities.eu: Greece, https://www.s4allcities.eu/
2. Arbab-Zavar, B., Sabeur, Z.A.: Multi-scale crowd feature detection using vision sensing and statistical mechanics principles. Machine Vision and Applications **31**(4), 1–16 (2020)
3. Ardizzone, E., Bruno, A., Mazzola, G.: Scale detection via keypoint density maps in regular or near-regular textures. Pattern Recognition Letters **34**(16), 2071–2078 (2013)
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
5. Bruno, A., Ardizzone, E., Vitabile, S., Midiri, M.: A novel solution based on scale invariant feature transform descriptors and deep learning for the detection of suspicious regions in mammogram images. Journal of Medical Signals and Sensors **10**(3), 158 (2020)
6. Bruno, A., Greco, L., La Cascia, M.: Video object recognition and modeling by sift matching optimization. In: Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods. pp. 662–670 (2014)
7. Cheng, Z., Qin, L., Huang, Q., Yan, S., Tian, Q.: Recognizing human group action by layered model with multiple cues. Neurocomputing **136**, 124–135 (2014)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dogbe, C.: On the modelling of crowd dynamics by generalized kinetic models. Journal of Mathematical Analysis and Applications **387**(2), 512–532 (2012)

10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
12. Hou, X., Wang, Y., Chau, L.P.: Vehicle tracking using deep sort with low confidence track filtering. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2019)
13. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Fang, J., imyhxy, Michael, K., Lorna, V, A., Montes, D., Nadar, J., Laughing, tkianai, yxNONG, Skalski, P., Wang, Z., Hogan, A., Fati, C., Mammana, L., AlexWang1900, Patel, D., Yiwei, D., You, F., Hajek, J., Diaconu, L., Minh, M.T.: ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference (Feb 2022). https://doi.org/10.5281/zenodo.6222936, https://doi.org/10.5281/zenodo.6222936
14. Khan, A., Ali Shah, J., Kadir, K., Albattah, W., Khan, F.: Crowd monitoring and localization using deep convolutional neural network: A review. Applied Sciences **10**(14), 4781 (2020)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
17. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5099–5108 (2019)
18. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
19. Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y.: Density-aware person detection and tracking in crowds. In: 2011 International Conference on Computer Vision. pp. 2423–2430. IEEE (2011)
20. Sabeur, Z., Arbab-Zavar, B.: Crowd behaviour understanding using computer vision and statistical mechanics principles. In: Crowd Dynamics, Volume 3, pp. 49–71. Springer (2021)
21. Saqib, M., Khan, S.D., Blumenstein, M.: Texture-based feature mining for crowd density estimation: A study. In: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE (2016)
22. Singh, U., Determe, J.F., Horlin, F., De Doncker, P.: Crowd monitoring: State-of-the-art and future directions. IETE Technical Review **38**(6), 578–594 (2021)
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence (2017)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
25. Tripathi, G., Singh, K., Vishwakarma, D.K.: Convolutional neural networks for crowd behaviour analysis: a survey. The Visual Computer **35**(5), 753–776 (2019)

26. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3645–3649. IEEE (2017). https://doi.org/10.1109/ICIP.2017.8296962
27. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 2054–2060. IEEE (2010)
28. Xu, D., Song, R., Wu, X., Li, N., Feng, W., Qian, H.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing **143**, 144–152 (2014)
29. Zhang, C., Vinyals, O., Munos, R., Bengio, S.: A study on overfitting in deep reinforcement learning. arXiv preprint arXiv:1804.06893 (2018)
30. Zhou, B., Tang, X., Wang, X.: Coherent filtering: Detecting coherent motions from crowd clutters. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 857–871. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)