This is a repository copy of *Smart IoT cameras for crowd analysis based on augmentation for automatic pedestrian detection, simulation and annotation*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/148173/

Version: Accepted Version

Smart IoT Cameras for Crowd Analysis based on augmentation for automatic pedestrian detection, simulation and annotation

1st Antoine Rimboux
Science, Engineering and Computing
Kingston University
London, UK
A.Rimboux@kingston.ac.uk

2nd Rob Dupre
Science, Engineering and Computing
Kingston University
London, UK
R.Dupre@kingston.ac.uk

3rd Thomas Lagkas
University of Sheffield International Faculty
CITY College
Thessaloniki, Greece
tlagkas@citycollege.sheffield.eu

4th Panagiotis Sarigiannidis
University of Western Macedonia
Informatics and Telecommunications Engineering
Kozani, Greece
psarigiannidis@uowm.gr

5th Paolo Remagnino
Science, Engineering and Computing
Kingston University
London, UK
P.Remagnino@kingston.ac.uk

6th Vasileios Argyriou
Science, Engineering and Computing
Kingston University
London, UK
Vasileios.Argyriou@kingston.ac.uk

Abstract—Smart video sensors for applications related to surveillance and security are IOT-based as they use Internet for various purposes. Such applications include crowd behaviour monitoring and advanced decision support systems operating and transmitting information over internet. The analysis of crowd and pedestrian behaviour is an important task for smart IoT cameras and in particular video processing. In order to provide related behavioural models, simulation and tracking approaches have been considered in the literature. In both cases ground truth is essential to train deep models and provide a meaningful quantitative evaluation. We propose a framework for crowd simulation and automatic data generation and annotation that supports multiple cameras and multiple targets. The proposed approach is based on synthetically generated human agents, augmented frames and compositing techniques combined with path finding and planning methods. A number of popular crowd and pedestrian data sets were used to validate the model, and scenarios related to annotation and simulation were considered.

Index Terms—crowd analysis, data augmentation, crowd behavior

I. INTRODUCTION

IoT technologies incorporate a vast amount of specialized protocols and schemes [1], which have allowed them to enter a variety of domains. On this basis, they are eventually expected to dominate in a plethora of use cases, including very promising ones such as aerial communications for UAV-based monitoring applications [2], where pervasive collection, annotation and processing of data is required, realizing the vision of ubiquitous computing. In the same time, multimedia traffic exhibits a continuous growth which is attributed to the evolution of computing devices (especially mobile devices), the increasing demands from the multimedia users' side, the improvements of video quality and content variety, as well as the significant developments in the telecommunications and networking infrastructure. Contemporary network multimedia

services are now enabled through modern cyber-physical systems and can be provided through IoT autonomous distributed architectures utilizing agent-based middleware solutions [3]. Furthermore, video streams (e.g. for surveillance) generated by IoT devices can now be routed over ad hoc wireless links and/or the fronthaul domain of the telecommunication system utilizing the flexibility provided by Software Defined Networking (SDN) and advanced techniques for optimized federated management over heterogeneous networks [4], [5].

In the present video surveillance and situation security assume a certain level of understanding regarding the environment and the human or crowd behaviour [6]–[8]. IoT applications such as surveillance, security, and market analysis require accurate crowd and pedestrian detection and behaviour recognition. Video processing and computer vision approaches based on deep *convolutional neural network* (CNN) architectures provide accurate estimates [9], [10], but their accuracy depends on the quality and the amount of annotations in the training data sets. In order to improve the overall detection and recognition accuracy, a significant amount of labelled data is required, while the annotation process is time consuming, tedious, cost inefficient, prone to error, and often leads to performance degradation. Furthermore, the problem of how to perform comparative studies to simulation algorithms is considered in this work. The lack of a single and unified form of comparison between different simulation, detection and modelling approaches is still an issue for the crowd and pedestrian simulation and analysis methods. This often means that a given methodology is developed and evaluated for a specific purpose, with its wider abilities and properties left unconfirmed. Generally, the employed evaluation techniques and related measures can be broadly split into qualitative [11] and quantitative [12], [13]. The former including assessment

arXiv:1906.03994v1 [cs.CV] 6 Jun 2019

made by experts in the field or context of the intended application [14], as well as category based rating systems [15] designed to define the capabilities of an algorithm.

A number of quantitative measures have been suggested to provide a numeric measure of accuracy for detection and simulation, to mention a few: bounding boxes, tracks, speed, pedestrian density, number of steps taken to destination and duration. These evaluation techniques tend to be data driven, and as such require ground truth data for testing purposes. The concept of an evaluation framework has been suggested before [16]–[21]; with most deducing various metrics based on a simulation in an effort to rate simulation algorithms or tune parameters. Many of these evaluation frameworks are affected by problems tightly related with the data collection and annotation process, such as cost, time, privacy and suitability and availability of large outdoor environments.

With the advent of so much research in the area of pedestrian detection and crowd analysis, the ability to generate realistic data of different modalities, including ground truth has been researched over the last few years. It is worth mentioning the work of [11], [12], [15], [22]–[28] on data generation and simulation algorithms for crowd and pedestrian analysis.

In [16] the creation of a tool is proposed that characterises and generates outlying behaviour in simulated videos. In [18] an entropy score is computed to generate simulated data closer to real world data. The *stochastic variational dual hierarchical Dirichlet process* (SV-DHDP) model is introduced in [29], where groups of similar trajectories (trending paths) and subpixel motion flows [30]–[32] can be combined to generate an overall path pattern for an environment offering higher realism. In [33], [34] the concept of *look and feel* of a crowd is proposed by comparing an agent's actions at a given moment in time with a database of observed actions, providing more realistic simulated videos. In [21] the issue of tracking generalised paths in crowds is tackled using four dimensional histograms to describe and generate movement within a crowd. Additionally the work in [19], [20], [35], [36] propose interesting approaches and metrics to generate realistic crowd behaviours.

The use of synthetic data generation to train and evaluate machine learning models based on deep architectures was introduced recently. Methods in the literature use the process to generate labelled data sets for different applications, including pedestrian and crowd analysis. These methods aim to generate pedestrians and groups of people at different locations in a given scene, supporting a variety of appearances. However, current methods are either restricted to a single camera in terms of their usage or support only single frames without motion to be considered. Additionally, all of them are based on 3D virtual environments with quality not comparable with real video sequences affecting the obtained models and underperforming in real scenes. In this paper the proposed framework supports multiple cameras and moving agents, generating videos instead of single images based on compositing, a technique used to generate realistic videos by superimposing virtual object in real scenes. As such, the following novel
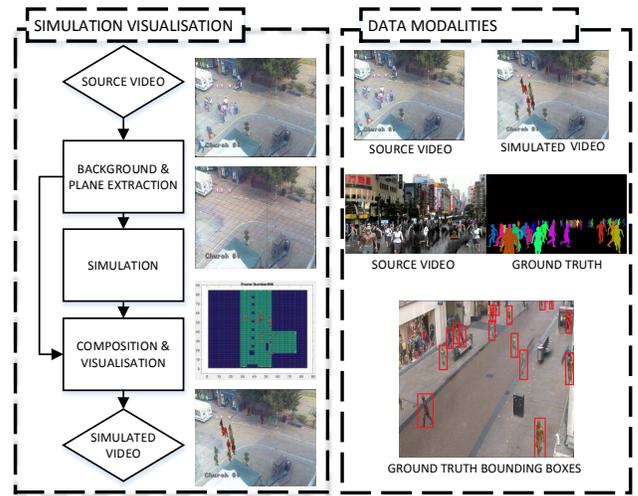

Fig. 1. Overview of the CCF framework.

framework is suggested which reduces the complexity of crowd and pedestrian realistic simulation, allows the automatic generation of data sets and annotation, supporting different data modalities, aiming to improve the detection accuracy of existing deep architecture focusing on pedestrian detection, pose estimation and tracking.

II. SIMULATION EVALUATION USING AUGMENTATION

The proposed modular *crowd composition framework* (CCF) provides a method of pedestrian and crowd data augmentation which considers realistic simulated human walking behaviours, multi-view and multi-modal data. Data generation can be implemented on a frame by frame basis or for a sequence as a whole, providing flexibility on how simulated data and ground truth are extracted. Additionally, the proposed methodology requires no track or path information, allowing the user to control parameters related to the number of pedestrians, their behaviour and the modality of the data.

The proposed framework takes as input source video footage and generates an augmented output video using composition techniques. The process utilises background subtraction techniques as well as methods to extract 3D data from 2D images. This allows the construction of a 3D space in which virtual agents can navigate around. Through the use of composition, a final visualisation combining this background and 3D space is generated to form the simulated video sequence in which artificial agents are superimposed onto the background of the source video data. Fundamentally the framework is made up of two components: simulation visualisation and annotated video data generation. The modular nature of the framework supports inputs of any simulation algorithm or video analysis techniques, depending on application. Furthermore, it retains the ability to produce realistic synthetic annotated crowd data. Figure 1 provides a more specific overview of the CCF framework as it is utilised in this work.

As the proposed framework uses videos to generate synthetic data, firstly a simulated video must be constructed.

Initially, using the source video sequence, the background is obtained. Next, a two dimensional plane is extracted representing a top down view of the given scene. Simulations are then run to produce paths for the virtual agents to follow based on the extracted plane. The visualisation component is then used to create a composite of the extracted 2D background image and 3D rendered agents as they follow the simulated paths. Frames are output from the visualisation into a final simulated video sequence. Once both a simulated and source video are available, the ground truth and other data modalities such as depth can be exported. Finally, *tracklets*, pose, skeleton, flow, and density measures can be evaluated and used for training and evaluation in deep architectures or other learning techniques.

To allow the composition of the simulated video to be created, the background of the source video sequence is required [37]. Once the background image has been subtracted the process of defining the perspective grid is applied. The perspective grid allows scale mapping of an environment from the viewpoint of the source video camera pose. The resultant grid represents a top down environment map of the viewable area and is used during agent simulation. Using the concept of perspective scale along a line we can, through the definition of two parallel lines that run to the vanishing point of an image, estimate distance in arbitrary units of measure within this perspective space (Figure 2b). This unit can be based upon an object in the scene with known dimensions or using pedestrians [38].

Initially the user defines the points $\mathbf{i}$ and $\mathbf{j}$, in the 2D image space, forming a line along an edge that leads to the vanishing point of the image. A second line is defined by the points $\mathbf{k}$ and $\mathbf{l}$, such that it runs *parallel*, relative to the vanishing point in the 3D space of the captured image, to the line defined by points $\mathbf{i}$ and $\mathbf{j}$ (Figure 2a). At a location along the line $\mathbf{ij}$ the user defines another point $\mathbf{u}_1$, such that the line $\mathbf{iu}_1$ represents the unit of distance m from which all further perspective points are defined. An additional point $\mathbf{u}_2$ is defined on top of the line $\mathbf{ik}$ which represents the same relative distance in 3D space as m. For the next step of the proposed algorithm the reference points $\mathbf{T}_{vanish}$, $\mathbf{R}$, $\mathbf{R}_0$ and $\mathbf{T}_{n-1}$ are initialised automatically (Figure 2a). In more detail, the vanishing point $\mathbf{T}_{vanish}$ is defined as the point at which the lines $\mathbf{ij}$ and $\mathbf{kl}$ intersect, this may well be at a position outside the image plane. As such $\mathbf{T}_{vanish}$ is defined as

$$\mathbf{T}_{vanish} = f(\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l}) \tag{1}$$

An arbitrary point $\mathbf{R}$ is selected at a random location outside the triangle $\mathbf{iT}_{vanish}\mathbf{k}$. The point $\mathbf{T}_{n-1}$ is defined as the point of intersection of the lines $\mathbf{iR}$ and $\mathbf{kT}_{vanish}$

$$\mathbf{T}_{n-1} = f(\mathbf{i}, \mathbf{R}, \mathbf{k}, \mathbf{T}_{vanish}) \tag{2}$$

Finally the point $\mathbf{R}_0$ is defined.

$$\mathbf{R}_0 = f(\mathbf{u}_1, \mathbf{T}_{n-1}, \mathbf{R}, \mathbf{T}_{vanish}) \tag{3}$$

With these points initialised, a recursive algorithm is applied to calculate equidistant points along the line $\mathbf{iT}_{vanish}$ in 3D
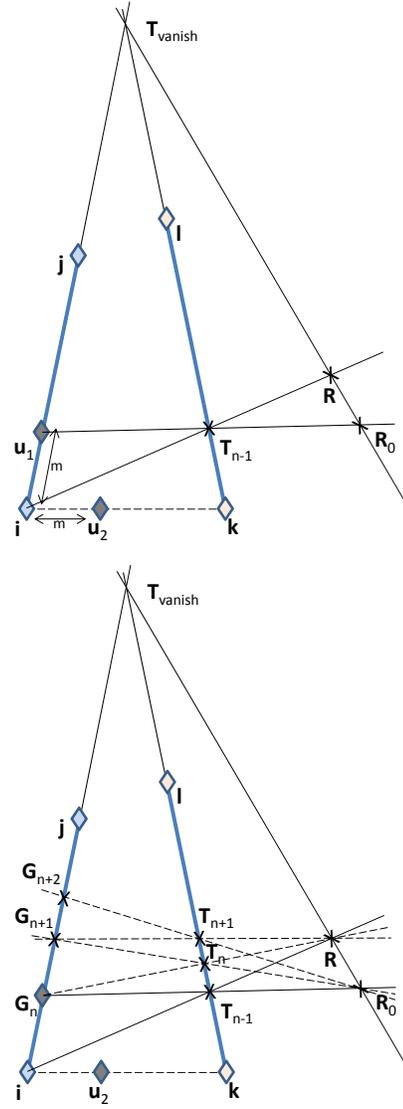

Fig. 2. (a) User defined points and initialisation. (b) The first two iterations of the recursive algorithm.

space. As the user has already defined the first of these points $\mathbf{u}_1$, for the purposes of the recursive step, these will be relabeled as $\mathbf{G}_n$. This is a two-step iterative process, with the point $\mathbf{T}_n$ being defined as the intersection of the lines $\mathbf{G}_n\mathbf{R}$ and $\mathbf{kT}_{vanish}$.

$$\mathbf{T}_n = f(\mathbf{G}_n, \mathbf{R}, \mathbf{k}, \mathbf{T}_{vanish}) \tag{4}$$

and during the second step the next equidistant point $\mathbf{G}_{n+1}$ on the line $\mathbf{iT}_{vanish}$ is expressed as a function of

$$\mathbf{G}_{n+1} = f(\mathbf{R}_0, \mathbf{T}_n, \mathbf{i}, \mathbf{T}_{vanish}) \tag{5}$$

This process is repeated until $\mathbf{G}_{n+1}$ is no longer within the borders of the original background image. The grid is initially defined using all the equidistant points on the line $\mathbf{ik}$, using the distance $\mathbf{iu}_2$ as a unit. Lines are defined between each of these points and the vanishing point $\mathbf{T}_{vanish}$ of the image. The
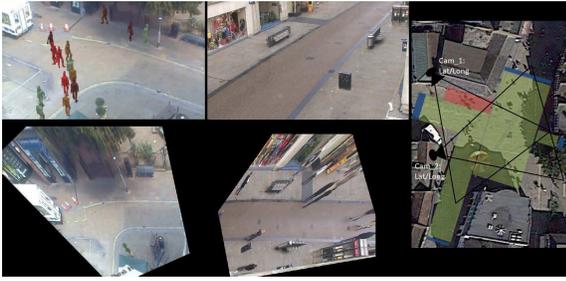
Fig. 3. An example of multiple camera simulation and the corresponding projections.

scale points **G** are plotted along each of these newly defined lines forming the grid. Additionally, if required, the recursive process can be inverted to create points moving away from the vanishing point. This ensures that the entire image plane is encapsulated by the defined grid, regardless of where the user has defined their points. The resultant grid represents the perspective plane of the source image. On that grid the areas (cells) with obstacles (i.e. cells where pedestrians cannot walk) are annotated as is information about entrance/exit locations. By adding a few additional calibration variables, during the plane extraction phase, the composition and visualisation process can be extended to multi camera setup steps, both in overlapping and non-overlapping scenes (Figure3). In these cases, rather than using a single virtual camera within a composition scene, additional cameras are added to represent the other camera views. During the calibration of these cameras, the same plane extraction technique is used, however additional relational measurements (latitude/longitude and orientation or distance and bearing from the source camera) are required to allow the positioning of these additional cameras within the 3D environment. The same process of background image alignment is also used on the additional virtual cameras to allow the generation of composite videos from these new views. The principle advantage here is that now a single crowd simulation can be viewed from several different camera angles whilst still having access to the ground truth data and crowd statistics.

III. EVALUATION OF THE CROWD COMPOSITION FRAMEWORK

This part of the paper is focused on the evaluation of the proposed CCF framework using a set of different deep neural networks trained for applications related to pedestrian detection. The first model used was [39] and is a *region proposal network* (RPN), used for the detection of pedestrians in abnormal situations. This model takes as input images of size 960×720 and, returns bounding boxes for all pedestrians present in the images (see figure 4 left). The second model proposed in [40] is based on the ResNet-101 deep network and aims to estimate the pose of different persons on a image. This network was trained using the COCO data set [41] where annotations are not made of boxes but keypoints which are,

Fig. 4. Left: Bounding boxes estimated by the RPN network. Right: Skeleton and the corresponding bounding boxes detected by the VGG-19 network.

for pedestrians, their joints (such as knee, ankle and neck) as we can see in figure 4 right.

In order to evaluate the contribution of the proposed CCF framework these networks are retrained using additional simulated data. Therefore, the data set that was used for the training and the comparison with the synthetic data is the *Town Centre* data set [42] (see figure 5) which represents the footage of a crowded street captured by a CCTV camera. This video is supplied with an annotation file which contains, the coordinates of the boxes bounding for the pedestrians present for each frame. Note that this data set will also be used as a reference to build the synthetic data. As such the background is extracted based on the proposed framework and is used int he composition of the synthetic data (see figure 5 right). Also, the synthetic data are designed to be similar to real life data so there will be a need to analyze the path of the pedestrians in the video to make the simulated pedestrians follow similar paths (e.g. select same entrance and exit points in the scene).

Fig. 5. Left: A frame from the 'Town Centre' video dataset. Right: Extracted Background.

A. Evaluation Metrics

In order to evaluate the accuracy of the models, metrics must be defined to determine success of failure in the detection of a pedestrian. Since the provided data set provides annotation files, the ground truth is provided as bounding boxes. Additionally, as it was previously stated, when an image is input, the model returns several boxes and the corresponding scores (which can be viewed as the probability for the region of the image located inside the box to actually represent a pedestrian). If the output is not formatted like so, for instance as the pose estimation model [40], estimation of the corresponding bounding boxes is performed. In this particular case, the top left corner of the box is defined by the lowest x and y-coordinates of the human joints. The bottom right corner is selected based on the highest x and y-coordinates.

Thus, a means to compute the accuracy of the model would be to compare, for a given image, the ground-truth boxes from the annotation file with the highest scored boxes that have been returned by the model for the same image. For this work

intersection over union (IoU) is used to determine the accuracy of the selected deep networks trained with and without the simulated data. So for two boxes, the ground truth and the prediction, computing this overlap consists of dividing the area of intersection between the boxes by the area of union. With A and B the areas of the two boxes we have

$$IoU = \frac{A \cap B}{A \cup B} \cdot 100 \qquad (6)$$

To compute the global accuracy, each predicted box with a score higher than a given threshold (most likely to represent a pedestrian) is compared with their closest ground-truth box (using the *Manhattan* distance). The overlap percentage for this ground-truth box is then added to the annotation file. Therefore, this file is made of the ground-truth data and the overlap for each box, which represents the accuracy of the box predicted by the model. Moreover, if a ground-truth box does not have an added overlap, that means that no predicted boxes where close enough to have a non-null union. In that case, the model did not manage to predict the location of the pedestrian represented by this ground-truth box. Also, if several predicted boxes are the closest to the same ground-truth box, the highest overlap is kept.

Finally, to compute the accuracy of the model in an image, the last part is to sum the overlap for each ground-truth box and divide it by the number of pedestrians present in the image to obtain the average accuracy. If an overlap for a ground-truth box passes a given threshold (e.g. 80%) the pedestrian is considered as found.

B. Crowd simulation process

The next step of the evaluation process is to build a crowd simulation made of synthetic agents using the proposed CCF framework, capturing the video frames and automatically generated ground truth.

Firstly, using CCF we extract the background of a sequence of images or the whole video, compute a grid that represents the walkable area for the synthetic agents and, then provide the characteristics of the synthetic agents (e.g. height, speed, entrance and exit points). We estimate the path of each agent but also its dimensions and orientations according to the obtained perspective. In this particular case, using the *Town Centre* data set, the first step is to extract the background of the video as it was analysed above. Once the background is extracted, the next step is to build the grid (or map) that will represent the area on which the agents will walk, following perspective of the scene. According to the CCF framework, we provide two parallel lines and a unit (i.e. a square of size 1×1 meter) and the framework returns the map (see figure 6a). All is left to do is to decide which part of the map is walkable, which parts are obstacles, and where the entrances and exits are situated. Then, with the details about the characteristic of the agents, the framework computes the path of each agent (see figure 6b) using a crowd simulation algorithm and returns their positions, orientation and scale information that is used for the scene simulation and the annotated data extraction.
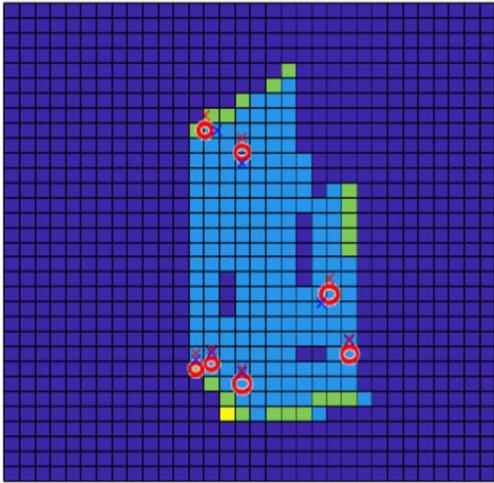
Fig. 6. a) Computed grid, partially filled. b) Top down view of the computed grid and visualization of the agents paths.

Fig. 7. a) Superposed map on the extracted background. b) Crowd simulation composited frame.

Note that, in order to provide an adequate amount of images for the training, approximately 100 agents are selected. The entrance and exit points are assigned for each agent randomly but with weighted probabilities based on the observed behaviours in the real video. Examples of the superposed map on the extracted background showing the walkable areas and the obtained composited crowd simulation are shown in figure 7. Next the simulation is run showing the agents walking on the map emulating a real crowd. Finally, for each rendered frame, the RGB image and the corresponding ground truth data are saved. At each frame, the agents in the scene are rendered according to their different characteristics (coordinates, rotation) in the map coordinate space. So at each frame, the CCF framework provides the ground truth as an image and as a list of bounding boxes. The generated synthetic data set includes 2845 images and it is used to retrain the available networks.

C. Fine-tuning the selected networks

The fine-tuning and retraining process for the selected RPN, and ResNet-101 networks, employed mainly to re-scale the

Model-Accuracy	Original	Proposed CCF
ResNet-101	39.09%	**50.04**%
RPN+	20.77%	**21.32**%

input images and format the ground truth to the expected dimensions and order. For each network the original parameters were selected and the new simulated data added to obtain the new models.

D. Results

Evaluation is carried out using the metrics previously presented. The table I shows the average accuracy. For the validation step, the read frames from the *Town Centre* data set were used. Results demonstrate that the synthetic data set generated using the proposed CCF framework can be utilised to train deep neural network and significantly improve their accuracy to detect pedestrians in real video sequences. Qualitative results for both training and testing stages are shown in figure 8.
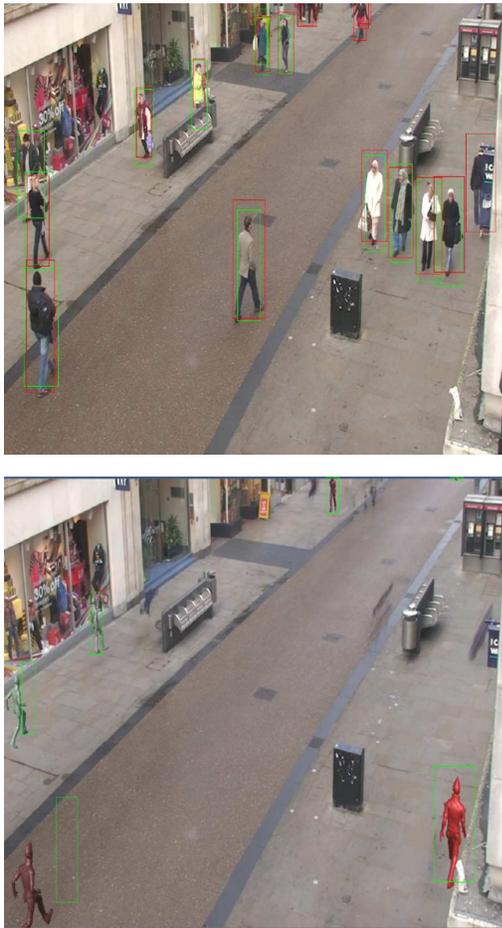
Fig. 8. a) Results using the ResNet-101. b) Examples of False positives using the simulate data.

IV. CONCLUSIONS

A novel *crowd composition framework* was presented which provides simulated annotated data using the composition process for pedestrian detection. Through the use of a modular system, any crowd or pedestrian simulation model or data annotation system supporting multiple cameras can be evaluated and compared by generating agent motion for use in the final visual simulation. Additionally, any video analysis feature can be utilised to evaluate similarity. Our experiments showed that the proposed framework improved the performance of deep networks in terms of pedestrian detection and crowd analysis.

ACKNOWLEDGMENT

REFERENCES

[1] A. Triantafyllou, P. Sarigiannidis, and T. D. Lagkas, "Network Protocols, Schemes, and Mechanisms for Internet of Things (IoT): Features, Open Challenges, and Trends," 2018. [Online]. Available: https://www.hindawi.com/journals/wcmc/2018/5349894/abs/

[2] T. Lagkas, V. Argyriou, S. Bibi, and P. Sarigiannidis, "UAV IoT Framework Views and Challenges: Towards Protecting Drones as Things," *Sensors*, vol. 18, no. 11, p. 4015, Nov. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/11/4015

[3] G. Eleftherakis, D. Pappas, T. Lagkas, K. Rousis, and O. Paunovski, "Architecting the IoT Paradigm: A Middleware for Autonomous Distributed Sensor Networks," *International Journal of Distributed Sensor Networks*, vol. 11, no. 12, p. 139735, Dec. 2015. [Online]. Available: https://doi.org/10.1155/2015/139735

[4] P. Bellavista, C. Giannelli, T. Lagkas, and P. Sarigiannidis, "Multi-domain sdn controller federation in hybrid fiwi-manet networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 103, May 2018. [Online]. Available: https://doi.org/10.1186/s13638-018-1119-0

[5] P. Bellavista, C. Giannelli, T. Lagkas, and P. Sarigiannidis, "Quality management of surveillance multimedia streams via federated sdn controllers in fiwi-iot integrated deployment environments," *IEEE Access*, vol. 6, pp. 21 324–21 341, 2018.

[6] V. Bloom, V. Argyriou, and D. Makris, "G3di: A gaming interaction dataset with a real time detection and evaluation framework," in *Computer Vision - ECCV 2014 Workshops*. Cham: Springer International Publishing, 2015, pp. 698–712.

[7] V. Bloom, V. Argyriou, and M. Dimitrios, "Hierarchical transfer learning for online recognition of compound actions," *Comput. Vis. Image Underst.*, vol. 144, no. C, pp. 62–72, Mar. 2016. [Online]. Available: https://doi.org/10.1016/j.cviu.2015.12.001

[8] V. Bloom, D. Makris, and V. Argyriou, "Clustered spatio-temporal manifolds for online action recognition," in *2014 22nd International Conference on Pattern Recognition*, Aug 2014, pp. 3963–3968.

[9] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" *CoRR*, vol. abs/1607.07032, 2016.

[10] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou, "Pedestrian-Synthesis-GAN: Generating Pedestrian Data in Real Scene and Beyond," *ArXiv e-prints*, Apr. 2018.

[11] A. Portz and A. Seyfried, "Analyzing stop-and-go waves by experiment and modeling," in *Pedestrian and Evacuation Dynamics*, 2010, pp. 577–586.

[12] S. Kim, S. J. Guy, D. Manocha, and M. C. Lin, "Interactive simulation of dynamic crowd behaviors using general adaptation syndrome theory," *SIGGRAPH*, vol. 1, no. 212, p. 55, 2012.

[13] M. Asano, T. Iryo, and M. Kuwahara, "A pedestrian model considering anticipatory behaviour for capacity evaluation," *Transportation*, vol. 18, p. 28, 2009.

[14] F. Klugl, G. Klubertanz, and G. Rindsfuser, "Agent-based pedestrian simulation of train evacuation integrating environmental data," in *Lecture Notes in Computer Science*, vol. 5803, 2009, pp. 631–638.

[15] D. C. Duives, W. Daamen, and S. P. Hoogendoorn, "State-of-the-art crowd motion simulation models," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 193–209, 2013.

[16] P. Charalambous, I. Karamouzas, S. Guy, and Y. Chrysanthou, "A data-driven framework for visual crowd analysis," *CGF*, vol. 33, no. 7, pp. 41–50, 2014.

[17] D. Wolinski, S. Guy, A. Olivier, M. Lin, D. Manocha, and J. Pettr, "Parameter estimation and comparative evaluation of crowd simulations," *Computer Graphics Forum*, vol. 2, no. 33, pp. 303–312, 2014.

[18] S. J. Guy, J. van den Berg, W. Liu, R. Lau, M. C. Lin, and D. Manocha, "A statistical similarity measure for aggregate crowd dynamics," *ACM Transactions on Graphics*, vol. 31, no. 6, p. 1, 2012.

[19] M. Kapadia, M. Wang, S. Singh, G. Reinman, and P. Faloutsos, "Scenario space: Characterizing coverage, quality, and failure of steering algorithms," *SIGGRAPH*, vol. 1, p. 53, 2011.

[20] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *ICCV*, 2011, pp. 1235–1242.

[21] S. R. Musse, V. J. Cassol, and C. R. Jung, "Towards a quantitative approach for comparing crowds," *CAVW*, vol. 23, no. 1, pp. 49–57, 2012.

[22] H. Hattori, V. N. Boddeti, K. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in *CVPR*, June 2015, pp. 3819–3827.

[23] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *CVPR*, June 2016, pp. 3234–3243.

[24] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding realworld indoor scenes with synthetic data," *CVPR*, pp. 4077–4085, 2016.

[25] E. Papadimitriou, G. Yannis, and J. Golias, "A critical assessment of pedestrian behaviour models," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 3, pp. 242–255, 2009.

[26] S. Singh, M. Kapadia, P. Faloutsos, and G. Reinman, "Steerbench: A benchmark suite for evaluating steering behaviors," *Computer Animation and Virtual Worlds*, vol. 20, no. 5-6, pp. 533–548, 2009.

[27] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Q. Xu, "Crowd analysis: A survey," *MVA*, vol. 19, no. 5-6, pp. 345–357, 2008.

[28] J. Pettré, J. Ondrej, A.-h. Olivier, A. Cretual, and S. Donikian, "Experiment-based modeling, simulation and validation of interactions between virtual walkers," *SIGGRAPH*, vol. 2009, p. 189, 2009.

[29] H. Wang, J. Ondej, and C. O'Sullivan, "Path patterns: Analyzing and comparing real and simulated crowds," in *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games - I3D '16*, 2016, pp. 49–57.

[30] V. Argyriou, "Sub-hexagonal phase correlation for motion estimation," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 110–120, Jan 2011.

[31] V. Argyriou and T. Vlachos, "Sub-pixel motion estimation using gradient cross-correlation," in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, vol. 2, July 2003, pp. 215–218 vol.2.

[32] V. Argyriou, M. Petrou, and S. Barsky, "Photometric stereo with an arbitrary number of illuminants," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 887 – 900, 2010.

[33] A. Lerner, Y. Chrysanthou, A. Shamir, and D. Cohen-Or, "Data driven evaluation of crowds," *Lecture Notes in AI*, vol. 5884 LNCS, pp. 75–83, 2009.

[34] A. Lerner, Y. Chrysanthou, and A. Shamir, "Context-dependent crowd evaluation," *Computer Graphics Forum*, vol. 29, no. 7, pp. 2197–2206, 2010.

[35] B. Banerjee and L. Kraemer, "Evaluation and comparison of multi-agent based crowd simulation systems," in *Agents for games and simulations II*. Springer, 2011, pp. 53–66.

[36] F. Zanlungo, D. Brščić, and T. Kanda, "Pedestrian group behaviour analysis under different density conditions," *TRP*, vol. 2, pp. 149–158, 2014.

[37] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, 2004, pp. 28–31.

[38] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," *CVPR*, 2008.

[39] S. Huang and D. Ramanan, "Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters," in *CVPR*, July 2017, pp. 4664–4673.

[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[42] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *CVPR*, June 2011, pp. 3457–3464.