Information Research Communications, 2024; 1(2):65-73.
https://inforescom.org

Review Article

Detection and Tracking of People in a Dense Crowd through Deep Learning Approach-A Systematic Literature Review

Muhammad Firdaus Mohamed Badauraudine[1,*], Megat Norulazmi Megat Mohamed Noor[1], Mohd Shahizan Othman[2], Haidawati Binti Mohamad Nasir[1]

[1]Universiti Kuala Lumpur Malaysian Institute of Information Technology, Universiti Teknologi MALAYSIA.
[2]Department of Applied Computing and Artificial Intelligent, Faculty of Computing, Universiti Teknologi Malaysia.

ABSTRACT
Crowd-related incidents, such as the Hillsborough Disaster and the Kanjuruhan Stadium stampede, often result from poor crowd management, leading to tragedies like suffocation and crushing. To mitigate human error in crowd control, this research explores the use of deep learning for the detection and tracking of individuals in dense crowds. The study focuses on implementing artificial intelligence for automated crowd monitoring through a localization map, with an emphasis on re-identification accuracy and auto-annotation of targets in datasets. A Systematic Literature Review (SLR) was conducted following the PRISMA guidelines, analyzing 4384 articles published between 2019 and 2024 across five databases. 13 primary studies met the inclusion criteria and were analyzed to address questions related to the accuracy of crowd tracking and detection. This SLR aims to provide insights and reference points for further research in artificial intelligence, particularly in the areas of auto annotation and re-identification for crowd monitoring.

Keywords: Annotation, Deep Learning, Dense Crowd, Localisation, Re-Identification, Track.

Correspondence:
Mr. Muhammad Firdaus Mohamed Badauraudine
Universiti Kuala Lumpur Malaysian Institute of Information Technology, Universiti Teknologi MALAYSIA.
Email: mfirdaus.badauraudine@s.unikl.edu.my

Received: 25-11-2024;
Revised: 27-12-2024;
Accepted: 02-01-2025.

INTRODUCTION

Crowd detection by Multiple Head Tracking (MHT) is currently one of the most essential and challenging research topics in the computer vision community. Due to the common availability of high-quality low-cost video cameras and considering the inefficiency of manual surveillance and monitoring systems, automated video surveillance is now essential for today's crowded and complex environments. Accurate information about numbers plays a vital role in operational and security efficiencies for monitoring, controlling, and protecting crowds. The counting and tracking of many persons pose as a challenge due to several reasons that are occlusions, the constant displacement of people, different perspectives and behaviours, varying illumination levels, and because, the bigger the crowd, the lower the allocation of pixels per person.

However, the counting and tracking of people in crowds is important to be implemented. This is due to public domains such as stadiums, airports, and, and even religious gathering areas tend to be difficult to track the number of people in crowds due to its density. This difficulty has proven to be deadly such as the Hillsborough disaster on 15th April 1989, the human "crush and stampede" event that occurred for the Hajj pilgrims on 24th September 2015, and the human crush at Kanjuruhan Stadium on 1st October 2022. Due to these incidents, a question that usually pops up is how this can be avoided or lesson. One of the solutions that was being researched upon was crowd detection using MHT.

Crowd detection by MHT is currently one of the most essential and challenging research topics. However, methods developed by different researchers only produce satisfactory results in sparse crowd setting for crowd detection and tracking. Hence, this area of research needs further elaboration to implement this area of research in dense crowd benchmarks.

It is important to analyse the methods that can be used for tracking and detecting people in a crowd. It is also important to take a step back and understand the flow of how the tracking and detection was originally to how much it has evolved so that the current research can understand more on how the gaps were filled with each new evolution. Similar works to this research need to be researched to understand the methods used by different research to detect and track people in crowds and their gaps for the current research to fill them to be novel research.

There have been many review studies focusing on localisation to track or detect people in a crowd. However, some of those studies sometimes only detect and does not track. Another limitation is the accuracy of tracking and detection. The output of the studies

OPEN ACCESS
ScienScript

DOI: 10.5530/irc.1.2.10

Publishing Partner : ScienScript Digital. [www.scienscript.com.sg]

can determine the accuracy of the detection and tracking of people in a dense crowd. For example, a topological output will have a higher accuracy in the study compared to a density map output. Since this area of study deals with the physical security of humans, it is important to maintain a high accuracy of tracking and detection. Current architectures implemented in several papers such as YOLO was developed to detect objects in normal situations and not in a highly-dense environment. Another contribution to the accuracy of the output is the re-identification of people in a dense crowd. This is because if the same person exited the tracking parameter and rejoined, they can be counted as another person and increase the crowd count. Hence, it is important to further filter the papers to review research that are more similar to the current research.

In this paper, our aim is to help other researchers by making an SLR of related studies to using deep learning in tracking and detection of people in a dense crowd. The SLR's purpose is to research and compare how different studies approach tracking and detection involving dense crowds using deep learning approaches to find out which approach is best suited to answer the study's research question. The SLR encompass studies in the last five years that are from 2019 to 2024. The main contributions of this systematic review can be summarised as follows:

1. Searching articles from various databases with specific keywords that are broadly related to current research. Mainly, the research must include tracking in a crowd using deep learning. The reason why only tracking is used as the keyword and not detection is because to track, the experiment must detect the people first. Hence, both tracking and detection goes together. There are 4384 papers collected using the specified keywords from seven databases that are ACM Digital Library, Wiley Online Library, SpringerLink, ScienceDirect, IEEExplore, and PubMed. The research carried out must have a published date within the past five years' timeframe.

2. A large-scale literature review is completed using the PRISMA process methodology. The current research proposes three research questions. Firstly, the research questions were formulated. Then, the related studies were collected, the inclusion and exclusion criteria were defined, and from there, the primary studies focused on the related topics were identified with an acceptable quality score.

3. A thorough analysis has been implemented in the primary studies. Each reviewed paper is based on research questions, involving the current methods of detection and tracking of people in a dense crowd using deep learning approach and whether it is applicable in public places. The aim of this is to provide researchers with methods to track and detect people in dense crowds and how their output will be to assist them more in their research.

The remainder of this paper is structured as follows:

1. Section 2: The research methodology involving research questions and research protocols.

2. Section 3: The results of the literature review to provide answers and discussions to the initial research questions.

3. Section 4: The limitations of the study.

4. Section 5: The conclusion of the research.

MATERIALS AND METHODS

The methodology used for the current SLR follows the original guidelines proposed by S. Keele in 2007 in the paper "Guidelines for Performing Systematic Literature Reviews in Software Engineering". This section presents the method used to undertake the current SLR for the detection and tracking of people in a dense crowd. Figure 1 shows the methodology's flowchart for searching for the related papers, filtering them, and writing the report based on the primary studies.

Based on the flowchart, there are 6 phases of the methodology that highlight the process of preparing the SLR.

Planning

The planning phase is where the actions required to achieve the research objectives are decided. In this study, the objectives are related to the topic of the methods and algorithms used in the tracking and detection of people in a dense crowd as well as the accuracy of the output and what datasets are used for the experiment. The keywords to be used to search for papers in the five databases were also planned during this stage. Hence, this step is the foundation of the SLR's methodology.

Formulation of Research Questions

The research questions were discussed and laid out during this phase. The main questions proposed were related to the research study's output and what datasets were used for the research study's experiment. The method and architecture used for the detection and tracking of people in a dense crowd also serves as questions needed to be answered as different research provides different methods and algorithms that might not be as accurate as other studies and not suitable for dense crowds. The annotation method is also considered to further research on auto annotation of the individuals in the video frames of the dataset used. Some research focuses on detection only and not tracking. Due to the reasons stated, the research questions formulated are:

RQ1: What method will be used for detecting and tracking people in a dense crowd.

RQ2: What method is used for the annotation of individuals in a dense crowd.

RQ3: How accurate is the detection and tracking of people in a dense crowd.

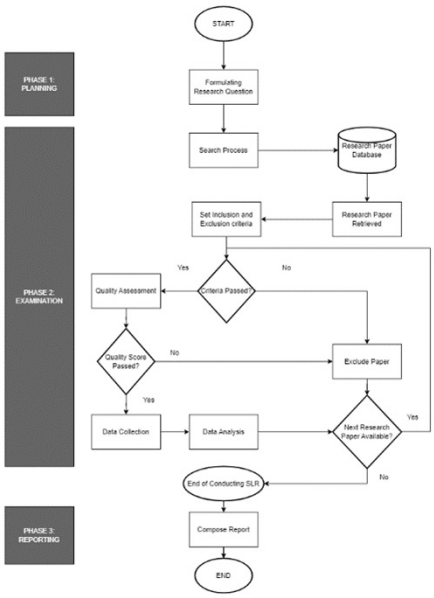


Figure 1: SLR's Methodology.

Table 1: Search query used for each database and number of papers gotten from them.

Database	Search Query	Number of Papers
IEEExplore	("CROWD") AND ("DEEP LEARNING") AND ("TRACK" OR "TRACKING")	99
ACM Digital Library	("CROWD") AND ("DEEP LEARNING") AND ("TRACK" OR "TRACKING")	1170
Scopus	("CROWD") AND ("DEEP LEARNING") AND ("TRACK" OR "TRACKING")	195
Wiley Online Library	("CROWD") AND ("DEEP LEARNING") AND ("TRACK" OR "TRACKING")	420
PubMed	("CROWD") AND ("DEEP LEARNING") AND ("TRACK" OR "TRACKING")	5

Search Process

In this subsection, each article to be used for comparing similar works is identified. The purpose of this phase is to extract relevant experimental studies on the tracking and detection of people in a dense crowd and if they can be implemented in public places. The databases used to get these articles are IEEExplore, ACM Digital Library, Scopus, Wiley Online Library, and PubMed. The original keywords planned during the planning phase is used to search for related articles in these databases but with a few adjustments. For example, instead of using "DENSE CROWD", the keyword used will be "CROWD". This is due to the number of papers being retrieved if the query is too strict. Some research papers might not use "DENSE CROWD" but they use different names to signify the same meaning. The papers will be more accurate to the current study in the future steps, especially the quality assessment step. Publish or Perish is used to obtain the papers and compile them as a CSV file. These papers will be filtered in the further stages. Table 1 shows the keywords used for the search query and articles gotten from each repository.

Inclusion and Exclusion Criteria

The studies included in the SLR were based on specific criteria. This is to filter out articles into papers that complement the current research. If the papers don't meet the criteria, they will be excluded and filtered out of the study. Among the most important criteria is that it must be within the appropriate timeframe, which is the past five years of the current studies, to make sure that they are still relevant to the current day and age. Another important factor is that the paper must be written in English. Papers that do not meet these criteria are filtered out. The filtering occurs step by step to achieve higher accuracy in filtering the papers. Table 2 refers to the inclusion and exclusion criteria and the number of papers excluded.

Quality Assessment

Table 3 shows the quality assessment questions.

These questions are to find research papers that can help answer the research questions proposed at the start of the SLR. The quality assessment questions are formulated similarly to the current study such as finding papers that implement dense crowds and localisation method.

The studies filtered from the previous step were further filtered through the quality assessment criteria. The main reason for this is although the papers gathered are somewhat related to the current topic, they are not able to help answer the research questions proposed during the research questions' formulation phase. The quality assessment phase is executed to filter out the best of the best papers that can help answer the research questions proposed in this SLR and extract data from the remaining research studies. The quality assessment checklist is as shown in Table 3.

The quality criteria are as follows in Table 4.

The remaining research papers were evaluated against the quality assessment questions shown in Table 3. If the score is at least 3, then the paper is accepted in this SLR. As the quality assessment goes through each of the paper, the scores related to each QA of the papers were done on an Excel sheet and can be downloaded and then viewed through the link, https://github.com/firdauskotp/SLR-DETECTION-AND-TRACKING-OF-PEOPLE-IN-A-CROWD-THROUGH-DEEP-LEARNING-APPROACH/blob/main/Publish%20or%20Perish%20Excel%20Sheets/Quality%20Assessment.xlsx.

Table 2: Inclusion and Exclusion criteria and papers excluded.

Inclusion Criteria	Exclusion Criteria	Number of Papers Excluded
Articles must be unique	There are duplicated articles.	140
Paper must be published within 5 years of when the SLR is written (2019 to 2024).	Paper is not published within 2019 to 2024.	52
Have an abstract related to the current research	No abstract or abstract not related to current research.	71
Paper must be in English	Paper is in a language other than English.	2
Articles' titles must relate to the current research	Articles not related to current research. Focus is on crowds. If crowd is not present, the article is excluded.	1497
Have a DOI or URL	Doesn't have both URL and DOI.	11
The same article could still be obtained	The same article cannot be found through the search engine used.	2
The article can be fully accessed	The article is not open-accessed even when using tools such as Zotero and Sci-Hub.	26

Data Collection

After the quality assessment phase, data extraction is implemented on the remaining papers which are now used as the primary studies. The reason is to find the strong point of each research paper that can answer a specific research question. Table 5 shows the research paper along with their most respected research question and how it helped to answer it.

Figure 2 shows the PRISMA process implemented during the SLR from the papers gotten until the papers used as primary studies.

The final part of the process is cleaning the data. This is because Publish or Perish mixes up the data between different databases.

The screening process from getting the papers in the database to getting the primary studies where the important details are fixed can be found in the Excel sheet, https://github.com/firdauskotp/SLR-DETECTION-AND-TRACKING-OF-PEOPLE-IN-A-CROWD-THROUGH-DEEP-LEARNING-APPROACH/blob/main/Publish%20or%20Perish%20Excel%20Sheets/PoP%20Primary%20Studies%20Screening%20Process.xlsx.

Table 3: Quality Assessment Questions.

Quality Assessment	Respective RQ
QA 1: The research uses localisation as its deep learning approach.	Q1
QA 2: The crowd involved in the study is a dense crowd.	Q1
QA 3: The research uses auto annotation for tracking the individuals in a dense crowd.	Q2
QA 4: One trained dataset is used on different test datasets.	Q2
QA 5: Re-identification of the same person is implemented in the tracking.	Q3

Table 4: Quality Assessment Scoring Criteria.

Quality Assessment Scoring	Score
The author(s) demonstrated a clear, detailed and explicit explanation on the answers for the specific RQ	High = H = 1
The author(s) provided a general, non-detailed, and non-explicit explanation on the answers for the specific RQ	Medium = M = 0.5
The author(s) either provided no or very few technical details on answering the specific RQ	Low = L = 0

Data Analysis

This is the final step of the SLR methodology where the primary studies and the data extracted from them are analyzed to answer the research questions according to the strong points of each paper.

RESULTS AND ANALYSIS

In this SLR, there are 13 primary studies obtained after the PRISMA process from 1889 studies collected from five different databases. These primary studies will be used to answer the research questions proposed at the start of the SLR. A summary of the primary studies can be found in Table 6.

VOSviewer was used to conduct a brief data analysis on the keywords found in the primary studies' title and abstract to determine the relationships and key aspects between them. As Publish or Perish gives a CSV file and it couldn't be converted to a RIS file which is accepted by VOSviewer, Zotero is used to get the DOIs and convert them into a RIS file. The RIS file is then used by VOSviewer to obtain the key aspects and relationships between the primary studies. Figure 3 shows the output of VOSviewer.

Based on Figure 3 and Figure 4, technical terms such as deep learning and crowd are repeated among the primary studies. This is to show that the primary studies involved are correlated with the current research. However, terms such as dense are not shown as not all of the primary studies involve dense crowds,

Table 5: Primary Studies and its most respected RQ.

Primary Study	Respected RQ
Learning how to analyse crowd behaviour using synthetic data.	RQ2
Tracking Hundreds of People in Densely Crowded Scenes with Particle Filtering Supervising Deep Convolutional Neural Networks.	RQ2
Crowd Abnormal Behavior Detection Combining Movement and Emotion Descriptors.	RQ2
Fast intensive crowd counting model of Internet of Things based on multi-scale attention mechanism.	RQ3
UUCT-HyMP: Towards Tracking Dispersed Crowd Groups from UAVs.	RQ2
Multi-Scale Occluded Pedestrian Detection Based on Deep Learning.	RQ1
An Aerial Crowd-Flow Analyzing System for Drone Under YOLOv5 and StrongSort.	RQ2
Fusion of CCTV Video and Spatial Information for Automated Crowd Congestion Monitoring in Public Urban Spaces.	RQ1
Enhancing Real-Time Human Tracking using YOLONAS-DeepSort Fusion Models.	RQ3
Less Is More: Learning from Synthetic Data with Fine-Grained Attributes for Person Re-Identification.	RQ2
A Survey on Multi-Target Multi-Camera Tracking Methods.	RQ3
Handling Heavy Occlusion in Dense Crowd Tracking by Focusing on the Heads.	RQ1
Topology and channel affinity reinforced global attention for person re-identification.	RQ1

and those that do uses different names across the studies such as high-density crowds. Re-identification is not shown here as well as they are written in the primary studies in different spellings such as Re-indentification, Re-identification, and also shorten to Re-ID. An important term is the synthetic data, which plays a role in auto annotation of some of the primary studies.

RQ1: What method will be used for detecting and tracking people in a dense crowd

There were not many papers that used the localisation method. Hence, if they were incorporating any kind of localisation method, they were researched further and counted towards the quality assessment. Most of the primary studies involve dense crowd but not many use any kind of localisation methods. Hence, three primary studies were mainly used to answer RQ1. PS13's

Figure 2: PRISMA Diagram for the current SLR.

topology localisation method was one of the closest to the current research as it uses the method with modules for re-identification of people with the help of Spatial Topology Information (STI) and Channel Affinity Information (CAI). Multi-Informant Fusion Reinforced Global Attention (MIFGA) is a module proposed by the researchers of PS13 in their research for more comprehensive information about STI and CAI. Although the study doesn't deal with a dense crowd, it helps further the research on involving detection and tracking through localisation with re-identification. PS12 also shows how to deal with occlusions when tracking people in a dense crowd. The research doesn't use localisation methods. However, it shows how to re-identify a person from the dense crowd. The research also focuses on head tracking of the people in the crowd. By avoiding occlusions, which are overlapping of the boundary boxes or something blocking the annotated target for detection, we can further improve the accuracy for RQ3. PS8 solves the average for all three RQs. It implements a localisation method by using a topology constraint as information for the spatial connectivity of egresses. It also involves re-identification and uses a method that the current research's experiment is doing that is verifying models not associated with the trained dataset to verify the accuracy. The reason for this is to check if the datasets need to be manually labeled or if a different dataset can automatically annotate the unlabeled datasets based on our desired target. If we were to combine the research of all three primary studies, we could achieve the current research objectives through the experiment conducted.

RQ2: What method is used for the annotation of individuals in a dense crowd

Most of the primary studies use manual annotation. However, there are primary studies like PS8 that train different models on different datasets to test the accuracy of their annotation.

Table 6: Summary of Primary Studies.

ID	Title	References	Publisher	Type	Respected RQ	Year
PS1	Learning how to analyse crowd behaviour using synthetic data	(A.R. Khadka *et al.*, 2019)	ACM Digital Library	Conference Paper	RQ2	2019
PS2	Tracking Hundreds of People in Densely Crowded Scenes with Particle Filtering Supervising Deep Convolutional Neural Networks	(Gianni Franchi *et al.*, 2020)	IEEE	Conference Paper	RQ2	2020
PS3	Crowd Abnormal Behavior Detection Combining Movement and Emotion Descriptors	(Xiao Li *et al.*, 2020)	ACM Digital Library	Conference Paper	RQ2	2020
PS4	Fast intensive crowd counting model of Internet of Things based on multi-scale attention mechanism	(Dong Liu *et. al.*, 2022)	The Institute of Engineering Technology	Article	RQ3	2022
PS5	UUCT - HyMP: Towards Tracking Dispersed Crowd Groups from UAVs	(Tonmoay Deb *et. al.*, 2021)	IEEE	Conference Paper	RQ2	2021
PS6	Multi-Scale Occluded Pedestrian Detection Based on Deep Learning	(Fang Li *et. al.*, 2022)	IEEE	Journal Article	RQ1	2022
PS7	An Aerial Crowd-Flow Analyzing System for Drone Under YOLOv5 and StrongSort	(Kuan-Hao Yeh *et. al.*, 2022)	IEEE	Conference Paper	RQ2	2022
PS8	Fusion of CCTV Video and Spatial Information for Automated Crowd Congestion Monitoring in Public Urban Spaces	(Vivian W. H. Wong *et. al.*, 2023)	MDPI	Article	RQ1	2023
PS9	Enhancing Real-Time Human Tracking using YOLONAS-DeepSort Fusion Models	(Athilakshmi R. *et. al.*, 2024)	ResearchGate	Conference Paper	RQ3	2024
PS10	Less Is More: Learning from Synthetic Data with Fine-Grained Attributes for Person Re-Identification	(Suncheng Xiang *et. al.*, 2021)	arXiv	Article	RQ2	2021
PS11	A Survey on Multi-Target Multi-Camera Tracking Methods	(Temitope Ibrahim Amosa *et. al.*, 2023)	ScienceDirect	Survey Paper	RQ3	2023
PS12	Handling Heavy Occlusion in Dense Crowd Tracking by Focusing on the Heads	(Yu Zhang *et. al.*, 2023)	arXiv	Article	RQ1	2023
PS13	Topology and channel affinity reinforced global attention for person re-identification	(Xile Wang *et. al.*, 2021)	Wiley Online Library	Research Article	RQ1	2021

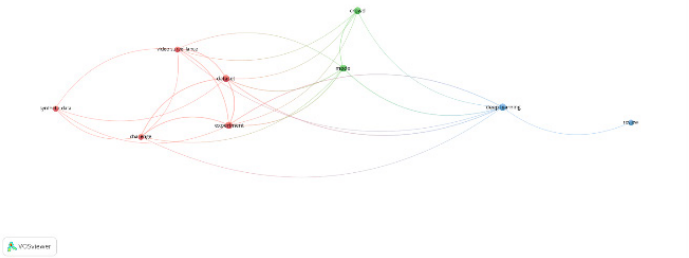
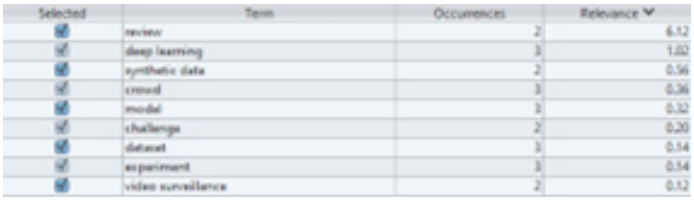


Figure 3: Key items of the primary studies visualised by VOSviewer.

However, most of the auto annotation from these primary studies come from synthetic data. Let's take PS1 for example where there was a discussion on exploiting abundantly available unlabeled crowd imagery in a learning-to-rank framework. PS2 overcomes the issue of manual annotation for videos. The method they used to achieve this is by having three Diffusion-Convolutional Neural Network (DCNN) monitored by a Particle Filter (PF). The first DCNN is the detector that detects the pedestrians via RetinaNet algorithm and the ResNet 50 architecture. The second DCNN was pretrained on the synthetic MPI-Sintel dataset. It is referred to as DCNN O. Flow, which is Deepflow. It takes two images as inputs and outputs the optical flow. The final DCNN is known as the DCNN Corrector where its training is supervised by the PF. Figure 5 summarises how the DCNNs were used together for tracking.

The researchers conducted two experiments where for the first, they used sequence 6 of an extensive annotated dataset of dense crowd sequence where the tracking accuracy threshold was set to 15 of 133 frames. The last 200 frames were used to train the DCNN Detection which produced a very high accuracy of 98%. The second experiment includes a dataset similar to the current research's objective that is the dense crowd during Makkah's pilgrimage. However, they use two different datasets composed of different times during the Makkah pilgrimage. For the DCNN Detection, a training set with no tracking annotation is used. The only annotations that were present are the bounding boxes that is around each person's heads. The second tracking dataset is based on seeds sampled mostly in high-density crowds during Makkah's Hajj pilgrimage. Their annotations span only a field of view in which human annotations can still be done without ambiguity. They compared the results between both algorithms where the first algorithm uses optical flow to track the pilgrims while the second algorithm is based on Neural Marching Cube (NMC). The DCNN Corrector which was trained with supervised learning earlier improves the accuracy when implemented. Figure 6 summarizes the results of their studies where PFMRF is the coupling of the Particle Filter and Markov Random Field model,



Selected	Term	Occurrences	Relevance
☑	review	2	6.12
☑	deep learning	3	1.02
☑	synthetic data	2	0.56
☑	crowd	3	0.36
☑	model	3	0.32
☑	challenge	2	0.20
☑	dataset	3	0.14
☑	experiment	3	0.14
☑	video surveillance	2	0.12

Figure 4: Terms of key items from primary studies that are visualised by VOSViewer.

DCNN Cor is short for the DCNN Connector, MOTP is Multiple Object Tracking Precision, MOTA is Multiple Object Accuracy, and TA is Tracking Accuracy. The amount of training data used in the model (T data) is tested with different epochs.

From the result, it can be seen that this research proves that different datasets can be used with an dissociated model and still achieve decent results. Besides that, the summary of this paper's experiments is that by training on diverse datasets, the DCNN learns to handle various conditions and crowd densities, while PF helps in refining detection. Through the combined approach, the researchers managed to overcome the issue of occlusions and maintaining constant tracking of the pedestrians. This combination improves the accuracy and reliability of auto-annotation, as the model can generalize better across different scenarios. The trained model can then be used to annotate new video frame frames by having a high accuracy in detecting and tracking people, as well as providing ground truths that are reliable for further training and evaluation. This methodology demonstrates that leveraging multiple datasets and integrating sophisticated tracking mechanisms can significantly enhance the performance of auto-annotation systems.

PS5 also has an interesting approach to the annotation of their models. The researchers created their own generator to generate the annotations. Their generator takes four steps to work that are data synthesis, diversity, environment simulation, and annotations. The first step that is data synthesis is where the generator creates synthetic crowd scenes by simulating both various crowd behaviours and movement patterns. Then diversity is introduced to the data in terms of variability which includes crowd density, group dispersion, and individual trajectories to mimic reality. Then, it is time to simulate the environment as the generator includes different kinds of environmental settings. These settings include urban and rural landscapes. The purpose of these settings is to ensure robustness of the tracking algorithms. Finally, each synthetic scenario is annotated with ground truth data. This will provide a benchmark for training and evaluation. By combining this primary study with the others, it is highly possible to create an accurate annotation generator for future researchers in this field so that they can reduce the labour on annotating and can focus more on other areas to improve the quality of tracking people in dense crowds.

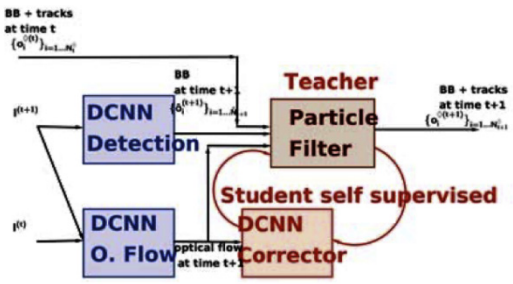


Figure 5: How the three DCNNs work together with the PF for tracking people in a dense crowd.

RQ3: How accurate is the detection and tracking of people in a dense crowd

All the primary studies that involve a dense crowd show a high accuracy in detection and tracking of people in it such as PS1 that shows a high accuracy as they had very low MAE through using Cross-Modal Transfer Learning (CMTL) and Convolutional Neural Network for Counting and Density Map Estimation (CSRNet) with real world and syntactic dataset. They had an accuracy of 88.06, 17.0, and 300.2 MAE for Shanghai Tech Part A, Part B, and UCF-50 datasets respectively through CMTL. Most of the high accuracy tests come from testing different architectures such as YOLO and GCN. However, there are primary studies that managed to further improve accuracy with the help of re-identification to reduce the rate of false positives. PS1, 2, 3, 6, and 7 are studies that involve dense crowds but don't have re-identification. These studies don't mention their false positives as well. Their accuracy measurement comes from different sources such as Mean Absolute Error (MAE) for PS1 and training accuracy via the amount of training datasets used with different epochs for PS2. For the research that both utilises and explains their use of re-identification, they are included as primary studies that respect RQ3. These primary studies include PS4, 9, and 11. In the primary studies that respect RQ3, it is shown that most of the research uses a tracking algorithm called Deep Association Metric (DeepSORT). DeepSORT uses a simple person Re-ID model architecture, making it possible to easily re-identify a person from a dense crowd. DeepSORT can be used to reduce false positives according to a paper called "False positive elimination in object detection method for videos" as it helps to re-identify the person in the scene. Hence, that person will not be able to increase the crowd count when re-identified. PS9 explains further about DeepSORT such as how it employs Kalman filtering and Hungarian matching to ensure consistent object tracking across frames. It also explained how DeepSORT produces tracked objects with corresponding IDs. This helps to reduce false positives as since every object, in the case of the research, humans, have their own IDs, they cannot be identified again as a new person. Instead, they will be re-identified by their original ID. The other PS such as 5, 6, 12, and 13 that doesn't delve much into re-identification still uses DeepSORT as its primary algorithm for re-identification. Hence, not only is it open-sourced but also widely used and trusted. From the public studies and research, what seemed to be a difficult task in re-identifying a person in the high-density crowd becomes an easier approach with the help of DeepSORT.

	MOTP	MOTA	TA
flow	22%	16%	39%
NMC	54%	61%	78%
PF	59%	66%	73%
PFMRF	60%	70%	78%
PFMRF DCNN Cor 2 epochs $T_{data} = 5$	60%	70%	78%
PFMRF DCNN Cor 5 epochs $T_{data} = 5$	**64%**	**75%**	**84%**
PFMRF DCNN Cor 10 epochs $T_{data} = 5$	62%	72%	82%
PFMRF DCNN Cor 5 epochs $T_{data} = 10$	62%	73%	82%
PFMRF DCNN Cor 5 epochs $T_{data} = 50$	63%	73%	83 %

Figure 6: Results of PS2's experiments.

LIMITATIONS

The SLR about the tracking and detection of people through a deep learning approach is conducted based on 13 primary studies between 2019 to 2024. However, there is a possibility that the result of the SLR is affected by a few factors. These factors are the coverage of the search strategy and quality assessment inaccuracy. These factors have been discussed and explained further in this section.

The coverage of the search strategy influences the number of papers gotten across different online databases. The keywords used need were related to the current research such as topology and dense crowd. The Boolean expression used can also influence the number of papers retrieved from the databases. While using the AND Boolean expression, that keyword must be available in the topic.

The filtering of the papers also affects the number of papers gotten. This is due to how EndNote files and papers gotten from Scopus don't give abstracts, influencing the step where the papers without abstracts. Due to them not getting the abstract from the beginning, these papers were not removed.

Publish or Perish and Zotero were used to retrieve the data. However, Publish or Perish mixes up the data in different column headers as the SLR retrieves data from different databases. Publish or Perish also causes some of the paper's dates to be wrong and having the wrong DOI. Hence, it was needed to manually check some of the papers that were screened to make sure that they can be counted as a primary study.

The quality assessment might also affect the SLR. This is due to how the final exclusion of the papers before the current study gets the primary studies relies on the quality for the papers. If the papers do not meet the quality score set during the section, the paper will be excluded. There are only 5 research questions that are closely related to the current study that the quality score is based on. The strictness of the quality set and the number of research questions will influence the final number of primary studies. Since there are not many papers that use topological localisation, the research that incorporated any kind of topological method was counted.

CONCLUSION

Detection and tracking of people in a dense crowd is essential research as there have been many accidents that have happened due to bad crowd management. To further the current study that is to track and detect people in a dense crowd through a deep learning approach, a SLR must be carried out to understand the methods and architectures used for tracking and detection of people in a crowd by similar studies. Even when we have the data for tracking the people in a dense crowd, we will need to manually annotate the people to be detected. Hence, the research also needed to cover on how to auto annotate the targets in our dataset. Research questions were formulated for the current study to be solved by the primary studies of the SLR. 1889 articles were originally obtained when five online repositories were searched using specific keywords. The number of articles was filtered down to 13 primary studies published from 2019 to 2024. The filtering includes exclusion of papers that are not related to the topic, not English, and doesn't achieve the target of the quality score set. The primary studies are then used to answer the three research questions where each primary study is researched more to see which research question is most suitable to fill in the gap. Finally, the limitations of the studies were researched and the future directions for researchers who are keen on this area of research were discussed. The aim of this paper is to serve as a guide for research in this area of study through the help of a systematic review of the methods and architectures of detection and tracking of people in a dense crowd and improve the productivity of research in terms of annotating the targets for labeling.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

Amosa, T. I., Sebastian, P., Izhar, L. I., Ibrahim, O., Ayinla, L. S., Bahashwan, A. A., Bala, A., & Samaila, Y. A. (2023). Multi-camera multi-object tracking: A review of current trends and future advances. Neurocomputing, 552, 126558. https://doi.org/10.1016/j.neucom.2023.126558

Athilakshmi, R., Chandan Sainagakrishna, P. S., Chaitanya Chowdary Kota, S. S., Kiran Teja, M. C., Venkatesh, T., & Prasad, V. J. (2023). Enhancing real-time human tracking using YOLONAS-DeepSort fusion models 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2023 (pp. 1118–1125). https://doi.org/10.1109/ICECA58529.2023.10394864

Deb, T., Rahmun, M., Bijoy, S. A., Raha, M. H., Khan, M. A., & UUCT. (2021). UUCT – HyMP: Towards tracking dispersed crowd groups from UAVs International Joint Conference on Neural Networks (IJCNN), 2021 (pp. 1–8). https://doi.org/10.1109/IJCNN52387.2021.9533600

Dubey, S. K., Satyanarayana, J. V., & Krishna Mohan, C. K. (2024, April 3). False positive elimination in object detection methods for videos. Proceedings of the SPIE 13072, Sixteenth International Conference on Machine Vision (ICMV 2023), 130720J. https://doi.org/10.1117/12.3023362

Franchi, G., Aldea, E., Dubuisson, S., & Bloch, I. (2020). Tracking hundreds of people in densely crowded scenes with particle filtering supervising deep convolutional neural networks IEEE International Conference on Image Processing (ICIP), 2020 (pp. 2071–2075). https://doi.org/10.1109/ICIP40778.2020.9190953

Khadka, A. R., Oghaz, M. M., Matta, W., Cosentino, M., Remagnino, P., & Argyriou, V. (2019). Learning how to analyse crowd behaviour using synthetic data. https://doi.org/10.1145/3328756.3328773

Li, J. (2023). A systematic literature review on feature selection for machine learning-based attack classification for IoT security. International Journal of Sensor Networks. X, Y4, 000–000.

Li, X., Yang, Y., Xu, Y., Wang, C., & Li, L. (2020). Crowd abnormal behavior detection combining movement and emotion descriptors 2nd International Conference on Industrial Control. Network and System Engineering Research. https://doi.org/10.1145/3411016.3411166

Liu, D., Wang, Z., & Meng, X. (2022). Fast intensive crowd counting model of Internet of Things based on multi-scale attention mechanism. IET Image Processing, ipr2.12686. https://doi.org/10.1049/ipr2.12686

Wang, X., Gao, C., Xin, M., Zhang, S., & Zhang, M. (2021). Topology and channel affinity reinforced global attention for person re-identification. International Journal of Intelligent Systems, 36(9), 5136–5160. https://doi.org/10.1002/int.22506

Wong, V., & Law, K. (2023). Fusion of CCTV video and spatial information for automated crowd congestion monitoring in public urban spaces. Algorithms, 16(3), 154. https://doi.org/10.3390/a16030154

Xiang, S., You, G., Guan, M., Chen, H., Yan, B., Liu, T. *et al*. (2021). Less is more: Learning from synthetic data with fine-grained attributes for person re-identification. arXiv. https://doi.org/10.48550/ARXIV.2109.10498

Yeh, K.-H., Hsu, I.-C., Chou, Y.-Z., Chen, G.-Y., & Tsai, Y.-S. (2022). An aerial crowd-flow analyzing system for drone under YOLOv5 and StrongSort International Automatic Control. Conference (CACS), 2022 (pp. 1–6). https://doi.org/10.1109/CACS55319.2022.9969785

Zhang, Y., Chen, H., Bao, W., Lai, Z., Zhang, Z., & Yuan, D. (2023). Handling heavy occlusion in dense crowd tracking by focusing on the heads. In arXiv. https://doi.org/10.48550/ARXIV.2304.07705.

Cite this article: Badauraudine MFM, Noor MNMM, Othman MS, Nasir HBM. Detection and Tracking of People in a Dense Crowd through Deep Learning Approach-A Systematic Literature Review. Info Res Com. 2024;1(2):65-73.