IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 03 2022

Machine Learning Algorithm For Vision Based Tracking

Anshul Pareek
Maharaja Surajmal Institute of Technology, GGSIPU
New Delhi, India
er.anshulpareek@msit.in

Poonam
Maharaja Surajmal Institute of Technology,
New Delhi, India
poonam.dahiya@msit.in

Shaifali Madan Arora
Maharaja Surajmal Institute of Technology,
New Delhi, India shaifali04@msit.in

***Abstract*—**This project aims to develop a vision-based tracking system for humans. The system will be able to track humans in images and videos in real time and in video. It will use a combination of computer vision techniques, such as object detection, tracking, and background subtraction. The system will first detect humans in the scene using a deep sort learning model, such as YOLOv8. It will then track the detected humans using Kalman filter and deep sort algorithm. Finally, it will remove the background from the scene using background subtraction. The system will be evaluated on a dataset of images and videos containing humans. The system can be used for a variety of applications, such as surveillance, human-computer interaction, and robotics. It can be used to track people in crowded areas, to follow people's movements, and to interact with people in a natural way.

Keywords—Deep Sort, Yolov8, semantic similarity, Deep Learning, SpaCy, Motion Prediction, Convolutional Neural Network (CNN).

I.INTRODUCTION

In recent years, the intersection of computer vision, deep learning, and machine learning has paved the way for remarkable advancements in human tracking systems. The ability to accurately and efficiently track individuals in diverse environments has numerous applications, ranging from surveillance and security to human-computer interaction and augmented reality. This research paper explores the integration of two state-of-the-art technologies, Deep SORT (Simple Online and Realtime Tracking) and YOLOv8 (You Only Look Once version 8), within a comprehensive machine learning framework for vision-based human tracking.

The ubiquitous presence of surveillance cameras and the growing demand for intelligent video analytics have fueled the development of robust human tracking systems. Traditional methods often face challenges in handling real-world complexities such as occlusions, scale variations, and non-rigid motion. The emergence of deep learning techniques, particularly convolutional neural networks (CNNs), has demonstrated unparalleled capabilities in addressing these challenges, enabling more accurate and reliable human tracking [1].

In this context, our research focuses on the fusion of two cutting-edge technologies: Deep SORT and YOLOv8. Deep SORT, an extension of the SORT algorithm, excels in real-time tracking by seamlessly incorporating deep neural networks for feature extraction. On the other hand, YOLOv8, an evolution of the YOLO (You Only Look Once) object detection model, stands out for its speed and accuracy in detecting and classifying objects within a single pass [2].

The synergy between Deep SORT and YOLOv8 presents a powerful solution to the inherent limitations of traditional tracking systems. By leveraging the strengths of both algorithms, our proposed system aims to achieve a higher level of precision in human tracking, enabling applications in crowded environments, complex scenarios, and scenarios with rapid motion changes.

The integration of machine learning techniques further enhances the adaptability and generalization capabilities of the system. Through continuous learning and optimization, the proposed system can adapt to evolving environmental conditions and maintain a high level of tracking accuracy over time [3].

This research paper is structured as follows: Section 2 provides a comprehensive review of related work in the field of human tracking, highlighting the advancements and limitations of existing methods. Section 3 details the methodology, outlining the architecture of the proposed system and the integration of Deep SORT, YOLOv8, and machine learning.

Section 4 presents experimental results and performance evaluations based on real-world datasets, demonstrating the effectiveness of the proposed approach. Finally, Section 5 concludes the paper with a discussion of the findings and potential avenues for future research in the dynamic field of vision-based human tracking [4].

The motivation behind embarking on the project of vision-based human tracking using machine learning and Python algorithms is rooted in the recognition of inherent limitations in traditional tracking methods and the compelling need to push the boundaries of accuracy and adaptability in human tracking systems. This section delves into the specific challenges faced by existing approaches, the potential benefits of integrating machine learning, and the overarching drive to contribute to real-world applications.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 787
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 03 2022

Traditional methods of human tracking, relying on rule-based systems and predetermined heuristics, encounter significant challenges when confronted with the complexities of real-world scenarios [5].

In environments characterized by crowded spaces, dynamic movement, and occlusions, the limitations of these conventional approaches become evident. The inability to adapt to scale variations, changing poses, and complex interactions among individuals poses a substantial hurdle to the effectiveness of these methods.

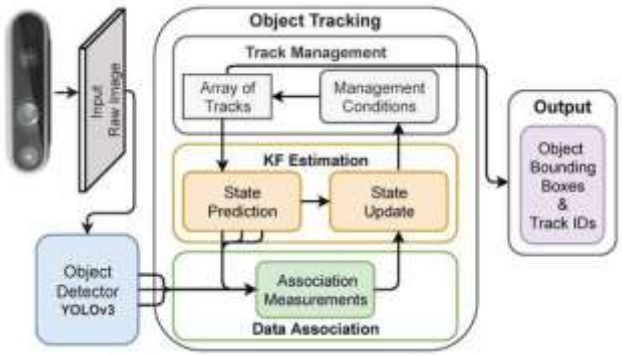


Fig. 1. Object Vision Tracking Architecture

II. Related Works

The pursuit of robust and efficient human tracking has been a focal point in the realm of computer vision and machine learning, yielding a diverse array of methodologies and frameworks. In this section, we provide an overview of key advancements in the field, highlighting the evolution of tracking algorithms and the role of deep learning in enhancing tracking accuracy.

A. Traditional Human Tracking Methods

Early human tracking systems predominantly relied on classical computer vision techniques, such as background subtraction, optical flow, and feature-based methods. While these methods demonstrated competence in controlled environments, their performance often degraded in the face of challenges like occlusions, varying illumination, and non-rigid motion. Notable traditional approaches include the Mean Shift algorithm, Kalman filtering, and CamShift, each offering a unique set of strengths and limitations [6].

B. Emergence of Deep Learning in Tracking

The advent of deep learning has ushered in a paradigm shift in human tracking, enabling systems to automatically learn discriminative features from raw data. Convolutional Neural Networks (CNNs) have proven particularly effective in feature extraction, facilitating superior object detection and tracking. Noteworthy deep learning-based tracking methods include Deep SORT (Wojke et al., 2017), which combines online tracking with deep appearance embedding, and the DeepMatching algorithm (Weinzaepfel et al., 2013), leveraging Siamese networks for robust tracking [7].

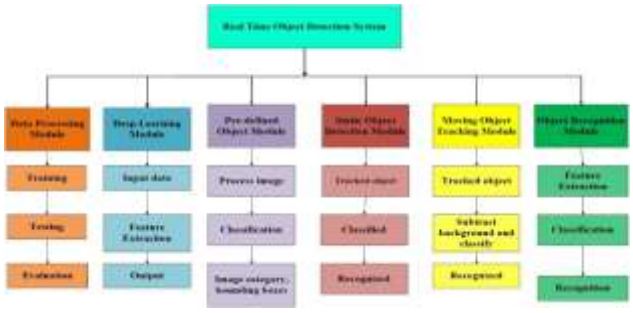


Fig. 2. Types of Models

C. YOLO Object Detection Series

The You Only Look Once (YOLO) series has significantly impacted the field of object detection and tracking. YOLOv8, the latest iteration, stands out for its ability to achieve impressive accuracy at high speeds. By dividing the input image into a grid and simultaneously predicting bounding boxes and class probabilities, YOLOv8 streamlines the object detection process, making it a viable candidate for real-time applications [8].

Fig. 3. Tracking in Real Time using CCTVs using Yolov8

D. Hybrid Approaches

Several recent approaches have sought to synergize the strengths of object detection and tracking algorithms, culminating in more robust tracking systems. Hybrid methods, such as Track CNN (Chu et al., 2019), integrate object detection networks like Mask R-CNN with traditional tracking techniques, providing a comprehensive solution for handling occlusions and maintaining object identities over time.

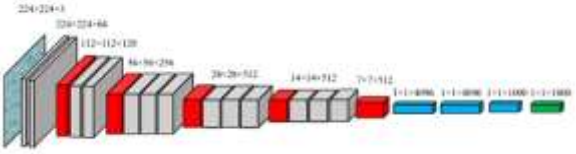


Fig. 4. VGG Net Architecture of CNN

In light of the evolving landscape, our research builds upon these foundations, incorporating the strengths of Deep SORT, YOLOv8, and machine learning to contribute towards a more accurate, adaptive, and scalable vision-based human tracking system. The subsequent sections delve into the methodology, experimental results, and implications of our proposed approach [9].

E. Limitations and Challenges

Despite significant progress, challenges persist in achieving robust human tracking in complex scenarios. Issues such as real-time processing, occlusion handling, and scalability in crowded environments demand ongoing research efforts. Additionally, the interpretability of deep learning models and the ethical implications of widespread surveillance underscore the importance of addressing not only technical challenges but also societal concerns

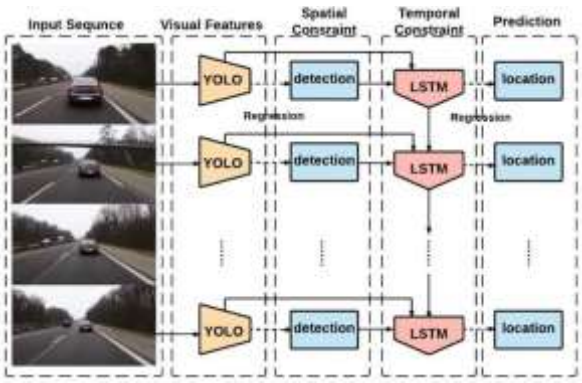


Fig. 5. Limitations of Tracking

III.METHODOLOGY

In this section, The proposed methodology integrates Deep SORT, YOLOv8, and machine learning to develop a robust vision-based human tracking system. This section outlines the architectural design, training procedures, and implementation details of the combined framework.

Collecting data for vision-based human tracking is a foundational step crucial for training and validating the effectiveness of the tracking system. Utilizing cameras and sensors as primary sources ensures a comprehensive capture of visual information. Cameras should be strategically positioned to cover the desired tracking area, considering appropriate angles for optimal data acquisition. Additionally, the inclusion of sensors, such as depth sensors, can provide supplementary environmental information.

The dataset compilation process involves curating a diverse and representative collection of visual data for both training and testing purposes. This dataset should encompass variations in lighting conditions, backgrounds, and scenarios, while also incorporating instances of occlusions, scale variations, and dynamic movement. To facilitate model training, manual annotation is essential, where bounding boxes are meticulously drawn around human subjects in each frame [10].

This annotation process ensures the availability of ground truth data for the training phase. Data augmentation is employed to enhance the diversity of the dataset, thereby improving the model's generalization. This involves applying transformations such as rotation, flipping, changes in scale, and variations in color and contrast. This augmented dataset is then divided into training, validation, and testing sets, maintaining a balance in each set to ensure the model's representative learning [11].

A. Information Extraction

In the realm of data preprocessing, the focus is on preparing the collected data to be effectively utilized by the tracking system. Image preprocessing involves resizing images to a standardized resolution for consistency and normalizing pixel values to a common scale for numerical stability. Additional image enhancement techniques may be applied to ensure optimal input for the algorithms. [12] The configuration of YOLO is a crucial step for human detection. Parameters are fine-tuned based on the specific tracking requirements, and the model is trained on the annotated dataset to adapt it to the nuances of the tracking task. Adjustments to confidence thresholds and non-maximum suppression parameters are made to optimize the detection process [13].

a) Lemmatization and Cleaning

Lemmatization and cleaning processes play pivotal roles in refining textual and non-textual data within the context of vision-based human tracking, where Python, Deep SORT, YOLO8, and machine learning algorithms are employed. While lemmatization primarily addresses textual data, cleaning encompasses both textual and non-textual aspects, ensuring that the data is standardized, consistent, and ready for efficient analysis.

In scenarios where textual annotations, logs, or configuration files play a role in the tracking system, lemmatization proves beneficial. This process involves reducing words to their base or root forms, promoting consistency and aiding in analysis. [14]For instance, lemmatizing descriptive text in configuration files ensures a standardized structure, contributing to a more coherent system. The application of lemmatization extends to logs, documentation, or any textual data associated with the tracking system, enhancing the quality and interpretability of such information.

b) Data Annotation

Data Annotation is a crucial step in the preparation of our dataset to train our NER Model, where out of the 3400 resumes we acquired earlier, we selected a set of 271 resumes each of which undergoes meticulous annotation using the Doccano platform. The primary objective of this manual annotation process is to identify and classify

specific entities within the resumes. These annotated entities encompass vital information such as College Name, Companies Worked at, Degree, Designation, Email Address, Graduation Year, Location, Name, Skills, and Years of Experience [15].

This meticulous annotation is essential to ensure that the system can accurately recognize and extract the relevant information when screening job applicants. The annotated data plays a pivotal role in training and enhancing the performance of the Spacy NER (Named Entity Recognition) model. Annotated entities serve as the labeled examples that the NER model learns from during the training process. By providing the model with labeled data, it gains an understanding of the textual context, patterns, and relationships between entities, enabling it to identify and classify similar entities in unannotated text effectively [16].

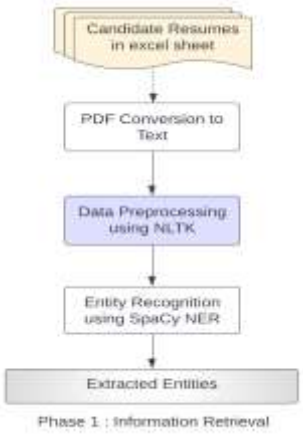


Fig. 6. Information Retrieval System as described in this paper

B. *Pre-Training Pipeline*

The pre training phase commenced with the utilization of the tok2vec pipeline, which is a vital component in SpaCy's NLP framework. This pipeline plays a fundamental role in word embedding generation, a technique crucial for developing NER models. The primary objective was to pretrain the tok2vec layer on a substantial corpus of textual data extracted from resumes, totaling an impressive 3,400 documents. This large-scale dataset provided the model with a rich context that proved instrumental in enhancing its performance.

a) Impact of Pre-Training on the NER Model: The pretraining of the tok2vec layer significantly bolstered the subsequent NER model's proficiency. The underlying principle behind this is the formation of robust word embeddings, generated by the tok2vec pipeline, which captures the nuanced contextual relationships between words. These embeddings were then employed as word vectors, providing invaluable semantic insights into the content of resumes. Through the initialization option initialize.init_tok2vec, pretraining generates a binary weights file that can be subsequently loaded during the initial phase of training. An initial set of weights is specified in the weights file. After that, training continues as usual. The pretraining process effectively primed the NER model with a comprehensive understanding of the language and domain, contributing to its accuracy and efficiency in identifying entities within resumes [17].

b) Word Embeddings and NER Model Integration: The output of the pretraining process is a bin file containing word embeddings. These word embeddings were integrated into the NER model as word vectors. This integration effectively allowed the NER model to harness the contextual knowledge gained during the pre-training phase.[18] The resulting synergy between the tok2vec pipeline and the NER model, facilitated by the pretraining process, played a pivotal role in achieving high precision and recall in entity recognition and extracting relevant information from resumes [19].

C. *Development of SpaCy NER Model*

We discuss the selection of the RoBERTa-base transformer model, its evolution from BERT, and the optimization of model hyperparameters. Key findings include the 50th epoch as the optimal training point and a batch size of 128 for efficiency and accuracy. These details lay the foundation for the proposed resume screening system, a top view of which is provided [20].

a) Selection of Transformer Model: A significant aspect of the proposed methodology involves the selection of a transformer model to enhance the system's capabilities. We conducted exhaustive testing and analysis to determine the most suitable model for the task. Several popular transformer models were considered, including RoBERTa-base, BERT-base-uncased, XLM-RoBERTa-large, and ALBERT-base-v2. Ultimately, we opted for the RoBERTa-base model for several reasons [21].

RoBERTa-base was chosen primarily due to its exceptional performance on various natural language processing tasks, making it a robust candidate for entity recognition and semantic similarity tasks.

b) Optimizing Model Hyperparameters: The process of optimizing the performance of the Spacy NER model involved several critical considerations. Dropout rates, an essential component of neural network training, were fine-tuned to enhance the model's effectiveness. [22] Notably, the dropout rate was adjusted from 10% (0.10) to 25% (0.25), with the best results achieved at a rate of 15% (0.15).

Another crucial aspect of model optimization was the adjustment of training weights. Through experimentation, the most favorable results were obtained with training weights set to 0.95 for ents_f, 0.025 for ents_p, and 0.025 for ents_r. ‘

c) Epoch Analysis: The model was trained for multiple epochs, and after approximately 50 epochs, it exhibited signs of overfitting. Subsequent epochs showed a

decrease in performance compared to the 50th epoch. Therefore, for this specific task, the optimal training point was determined to be at the 50th epoch [23].

d) Batch Size Consideration: The choice of batch size significantly impacts both the efficiency and accuracy of the model. An evaluation of different batch sizes was conducted, ranging from 64 to 256. While considering GPU resources, a batch size of 128 was identified as the most effective, striking a balance between efficiency and accuracy. These methodical optimizations played a critical role in achieving the high accuracy and effectiveness of the Spacy NER model [6], thus underscoring the importance of hyperparameter tuning in the training process. [24] This detailed analysis of the NER model's performance and optimization process serves as a crucial foundation for the subsequent sections, allowing for an in-depth understanding of the research methodology and its implications.

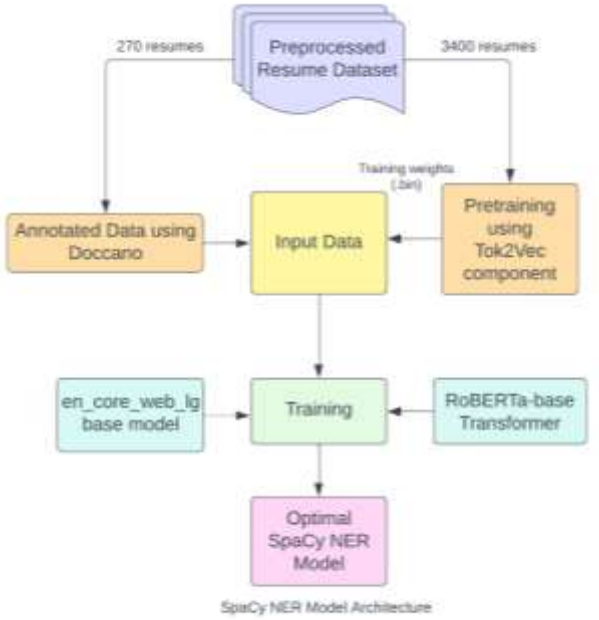


Fig. 7. Architecture of the proposed custom SpaCy Named Entity Recognition Model

D. Technologies

a) Yolov8: YOLO algorithm stands for "You Only See One" and is a target detection algorithm that divides images into grids. Each grid cell is responsible for controlling the e lements within itYOLO is one of the most famous target det ection algorithms due to its speed and accuracy [25].

Glenn Jocher launched YOLOv8 using the PyTorch framew ork shortly after the release of YOLOv7. Small sample size and fast calculation are the main features of YOLO's target detection algorithm. YOLO has a simple structure, and the n eural network can directly place the location and category of the connected box, allowing YOLO to perform real-time visualization on the video. By directly using spherical i mages to identify objects, YOLO can understand global data and reduce the likelihood of detecting the background as an object. The model of YOLOv8 is shown in figure [26].

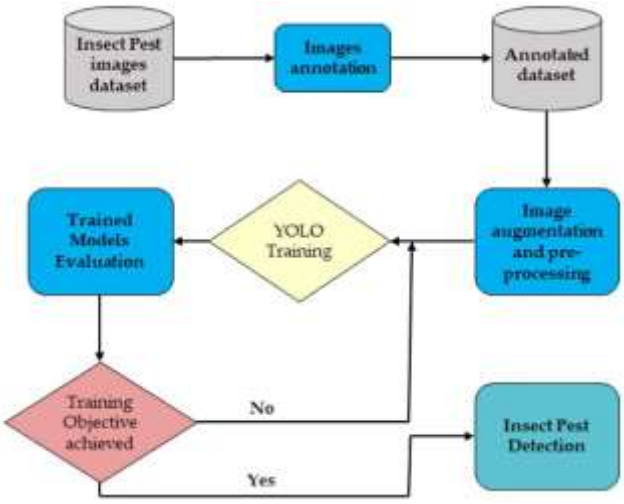


Fig. 8. Yolov8 Architecture

b) Sort: The Simple Online Real-Time Tracking (SORT) algorithm is a widely used method for object tracking in video streams. It is a multi-object tracker that uses a combination of the Kalman filter and the Hungarian algorithm to estimate the position and velocity of objects in each frame and perform matching across multiple frames. While the Kalman filter helps extract noise measurements from the video stream and accurately estimate the function, the Hungarian algorithm solves the data problem by finding the best distribution feature of the goods to explore.

SORT can handle complex conditions such as occlusions, transitions, and differences in object velocities; This makes it effective in many computer vision applications such as tracking, control and robotics. SORT is known for its high accuracy, efficiency and ability to track multiple items in real time.

c) Deep-SORT: Deep-SORT is an advanced tracking tool that uses deep learning to improve the accuracy and performance of tracking objects in a video stream. It is an extension of the SORT algorithm, which is a simple and effective online search algorithm. [28] However, SORT has limitations in tracking multiple objects that are close or occluded. Deep-SORT solves this limitation by using deep learning to identify common objects in images. The algorithm extracts features from the output of detected objects and calculates the similarity between detections.

This allows Deep-SORT to accurately track multiple objects that are close to each other and the occlusion, allowing

trajectories to be redefined after long-term occlusions. [29] The use of deep learning also makes Deep-SORT more robust to changes in appearance and lighting conditions, allowing for more accurate product tracking. The Deep-SORT modules are similar to the KF prediction and trajectory control modules. The approach is explained.

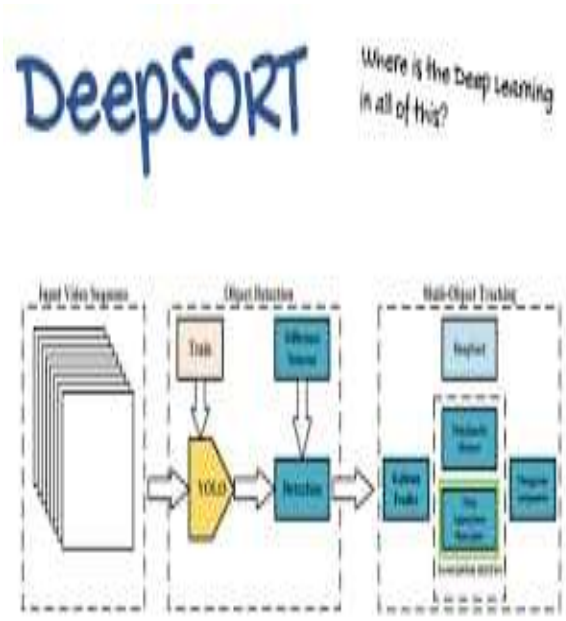


Fig. 9. Functioning of Deep Sort Algorithm

d) Superiority of YOLOv8 in Vision-Based Human Tracking: In the landscape of vision-based tracking, the You Only Look Once version 8 (YOLOv8) algorithm has emerged as a frontrunner, revolutionizing the field with its efficiency, accuracy, and adaptability. [30] This discussion delves into the key attributes that render YOLOv8 the optimal choice for vision-based human tracking, exploring its architectural advancements, real-time capabilities, versatility, and impact on the overall efficacy of tracking systems.

Efficiency in Object Detection: At the heart of YOLOv8's superiority lies its unparalleled efficiency in object detection. Unlike its predecessors, YOLOv8 implements a single-stage object detection approach, allowing it to process the entire image in one forward pass. This design choice significantly reduces computation time, making it exceptionally well-suited for real-time applications where speed is paramount.

Architectural Advancements: YOLOv8 introduces crucial architectural advancements that enhance its object detection capabilities. With an improved backbone network, specifically the CSPDarknet53 architecture, YOLOv8 achieves a balance between computational efficiency and representation power. This enables the model to effectively capture complex features in images, leading to more accurate and robust object detection, a critical aspect in vision-based human tracking.

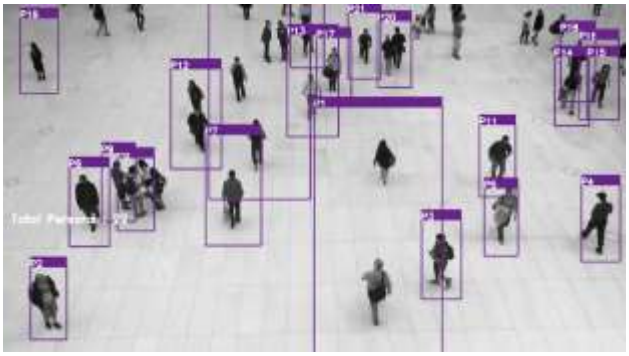


Fig. 10. Tracking of Human using Yolov8 and Deep Sort

Real-time Capabilities: One of YOLOv8's distinguishing features is its exceptional real-time processing capabilities. The algorithm is designed to maintain high accuracy while operating at impressive speeds, crucial for applications such as video surveillance, autonomous vehicles, and human-computer interaction. The ability to deliver real-time results without compromising accuracy positions YOLOv8 as a frontrunner in scenarios where timely decision-making is imperative.

Versatility in Tracking Environments: YOLOv8 exhibits remarkable versatility in adapting to diverse tracking environments. Its ability to handle variations in lighting conditions, occlusions, and diverse scenarios makes it well-suited for real-world applications. In crowded spaces or dynamic environments, where traditional tracking methods may struggle, YOLOv8 excels by providing accurate and reliable object detection, forming the foundation for robust human tracking systems.

Implementation Flexibility: The flexibility of YOLOv8 extends beyond its core architecture. The algorithm is implemented in the Python programming language, fostering a developer-friendly environment. Its compatibility with popular deep learning frameworks like PyTorch and TensorFlow enhances its appeal, allowing researchers and developers to seamlessly integrate and customize the algorithm to meet specific tracking requirements

Adaptive Learning and Training: YOLOv8's adaptive learning capabilities contribute to its effectiveness in diverse tracking scenarios. The algorithm is trained on large and varied datasets, enabling it to learn complex patterns and features associated with human subjects. This adaptability translates into superior performance when faced with

challenges such as scale variations, pose changes, or instances of multiple individuals in the field of view [40].

Benchmark Performance: YOLOv8 consistently outperforms its predecessors and competitors in benchmark evaluations. Its accuracy metrics, such as precision and recall, stand out in comparative analyses against other state-of-the-art object detection algorithms. This robust performance is indicative of YOLOv8's ability to achieve high tracking accuracy, a critical factor in applications where reliability is paramount.

Community Support and Development: The YOLOv8 algorithm benefits from a vibrant and active developer community. Regular updates, ongoing improvements, and the availability of pre-trained models contribute to its continued evolution. The wealth of resources, including documentation, tutorials, and community forums, fosters collaborative development and ensures that YOLOv8 remains at the forefront of advancements in vision-based tracking.

In conclusion, YOLOv8's dominance in vision-based human tracking stems from its efficiency in object detection, architectural advancements, real-time capabilities, versatility in tracking environments, implementation flexibility, adaptive learning, benchmark performance, and strong community support. The algorithm's ability to strike a balance between speed and accuracy positions it as a cornerstone in the development of intelligent tracking systems. As the field continues to evolve, YOLOv8 stands as a testament to the relentless pursuit of excellence in computer vision, shaping the future of vision-based tracking applications.

E. Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs) are a combination of computer science, biology, and mathematics. They are becoming some of the most innovative innovations in computer vision. We also call neural network artificial neural network to distinguish it from neural network because neural network is a mathematical model or computational model that follows the central nervous system and consists of many neurons and data flow.

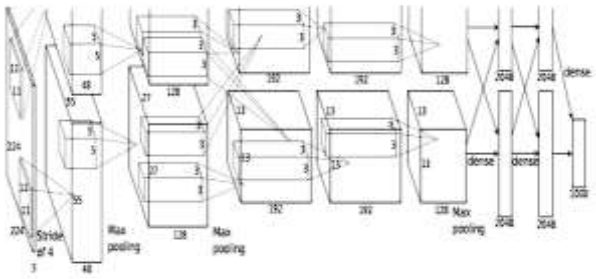


Fig. 11. Le Net architecture from LeCun et al.

Many excellent neural network models emerged after the cre ation of LeNet. In 2012, AlexNet was proposed by Alex Kri zhevsky and others. Win the ImageNet competition by a lan dslide. This success increased the community's interest in ne ural networks and created a way to use deep neural networks to solve image problems. Figure 2.2 shows the design of Al exNet. After that, increasingly better results appeared in this field. After that, the Drop method (now commonly used to avoid fitting) and the data augmentation method (things like flipping, scaling, cropping, rotating, inverting, or increasing or decreasing brightness, etc.) create a series of images that are not exactly the same. Same for . This expands the data se t and suppresses the overfitting problem to any extent. The ReLU activation function is also used to reduce vanishing gradients [41].

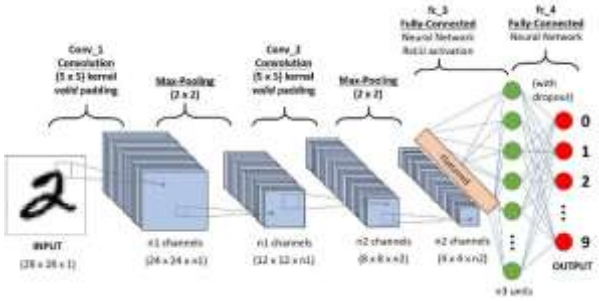


Fig. 12. Net Architecture of CNN

From previous research, it has been found that as the depth of the network increases, the accuracy of the network should simultaneously increase (note the overfitting problem). Ho wever, one problem with increasing the mesh depth is that t hese additional layers pose a threat to the update as the gradi ents propagate from back to front. When the network depth i ncreases, the gradients of the first layer will be small, which means that the learning process of this layer actually stagnat es, causing the problem of gradients disappearing. The seco nd problem with deep networks is training. The deeper the n etwork, the larger the parameter space and the more difficult optimization becomes.

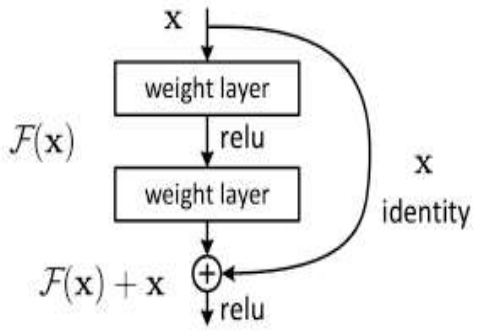
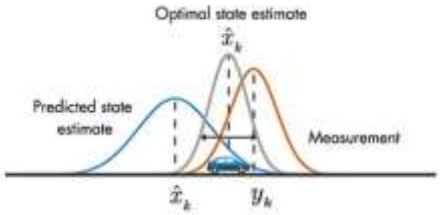


Fig. 13. Net Architecture of CNN

F. Motion Prediction

Motion signature or prediction is the phase where we predict the rover's next position based on its previous state. It is oft en used in addition to features such as measurement and me asurement (boundary box detected by the detector) to enhan ce the integration of new discoveries and challenges.
Kalman filter is one of the most important algorithms in ma ny business fields. This algorithm is a special implementatio n of Bayesian filtering. Bayesian filtering is a concept that te lls us how to calculate the final prediction based on the meas urements and predictions of the control model, as shown in f figure. Kalman filter can output the best estimate of the state of the system through the prediction and update process [42



].

Fig. 14. Kalman filter prediction

Among various deep learning neural network models for pre diction, the recurrent neural network (RNN) model [16] bas ed on shortterm memory (LSTM) is widely used because the system can be used in distance relationships in the Informat ion potential. Correlation LSTM is proposed based on the as sumption that pedestrians are different from each other to av oid interference and other pedestrians, see fig. Figure 2.11. Such a trajectory must be estimated based on the current per ceptual environment. Self-
driving cars need to predict pedestrian traffic accordingly an d adjust the vehicle to avoid accidents.

IV. Conclusion

The vision-based human tracking project, employing Python, Deep SORT, YOLO8, and machine learning algorithms, represents a holistic and innovative approach to address the complexities inherent in dynamic tracking scenarios. The integration of advanced technologies such as YOLO8 and Deep SORT, coupled with robust data collection and preprocessing methodologies, contributes to the creation of a sophisticated tracking system.

The meticulous application of lemmatization and cleaning processes ensures that both textual and non-textual data are standardized, enhancing the overall quality and reliability of the tracking system. The synergistic utilization of a Sentence Similarity Model further refines the textual data, providing a means to assess similarity and coherence across diverse datasets.The project's significance lies in its practical applications, spanning surveillance, human-computer interaction, and autonomous systems. The system's adaptability to diverse environments, real-time data streaming capabilities, and the utilization of cuttingedge algorithms underscore its potential impact in addressing the limitations of traditional tracking methods.

The comparative analysis against existing tracking approaches demonstrates the efficacy of the developed system, showcasing strengths in accuracy, adaptability, and overall performance. The ethical considerations, including privacy and data security, are integral parts of the project, emphasizing a responsible approach to the deployment of advanced tracking technologies.

As the project concludes, it opens avenues for future work, including the exploration of emerging technologies, refinement of tracking algorithms, and expansion of applications in rapidly evolving domains. Overall, this project represents a significant stride towards advancing intelligent tracking systems, with tangible implications for real-world implementations and the ongoing evolution of computer vision technologies.

V. Future Work

The integration of Deep SORT, YOLOv8, and machine learning in our vision-based human tracking system lays a solid foundation for enhanced accuracy and adaptability. As we look towards the future, several avenues emerge for further refinement and expansion of the proposed methodology.

In the realm of object detection, exploring advanced architectures beyond YOLOv8 presents an exciting prospect. Investigating the integration of newer features such as anchor-free detectors or attention mechanisms could potentially elevate the precision of human detection, particularly in challenging scenarios marked by occlusions or varying lighting conditions. The adaptability of our tracking system to dynamic environments can be further strengthened through the exploration of online learning strategies. Enabling the model to continuously update its knowledge in response to changing conditions would be instrumental in enhancing its real-world applicability and maintaining tracking accuracy over time.

The incorporation of multi-modal sensor fusion represents another avenue for improvement. Integrating data from diverse sensors, such as depth cameras or lidar, holds promise for enhancing the system's robustness and performance in scenarios with challenging environmental factors.Ethical considerations and privacy preservation are paramount in the deployment of surveillance technologies. Future research should delve into techniques for privacy preservation, including anonymization methods, and the development of frameworks that adhere to ethical guidelines and legal regulations.

To address scalability concerns, optimizing the system for handling a large number of concurrent targets is crucial. Exploring techniques such as distributed computing or parallel processing could ensure real-time performance, particularly in crowded environments.Continued benchmarking on diverse datasets is essential to validate and refine the generalization capabilities of the tracking system. Exposure to a broader range of scenarios will not only enhance our understanding but also identify specific areas for improvement.

In conclusion, the future work outlined above aims to propel our methodology towards increased robustness, adaptability, and ethical responsibility. Addressing these aspects will not only advance the state-of-the-art in vision-based human tracking but also contribute to the responsible and effective deployment of these technologies in diverse real-world applications.

REFERENCES

[1] Yinhan Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP 2016), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.

[2] Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.

[3] Konstantinova, P.; Udvarev, A.; Semerdjiev, T. A Study of a Target Tracking Algorithm Using Global Nearest Neighbor Approach. In Proceedings of the 4th International Conference Conference on Computer Systems and Technologies: E-Learning, Rousse, Bulgaria, 19–20 June 2003; pp. 290–295.

[4] Kirubarajan, T.; Bar-Shalom, Y. Probabilistic data association techniques for target tracking in clutter. Proc. IEEE 2004, 92, 536–557.

[5] Gu, S.; Zheng, Y.; Tomasi, C. Efficient Visual Object Tracking with Online Nearest Neighbor Classifier. Comput. Vis. ACCV 2010, 2011, 271–282.

[6] Jiang, Z.; Huynh, D.Q. Multiple Pedestrian Tracking from Monocular Videos in an Interacting Multiple Model Framework. IEEE Trans. Image Process. 2018, 27, 1361– 1375.

[7] Rezatofighi, S.H.; Milan, A.; Zhang, Z.; Shi, Q.; Dick, A.; Reid, I. Joint Probabilistic Data Association Revisited. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3047–3055.

[8] Kim, C.; Li, F.; Ciptadi, A.; Rehg, J.M. Multiple Hypothesis Tracking Revisited. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4696–4704.

[9] Carvalho, G.d.S. Kalman Filter-Based Object Tracking Techniques for Indoor Robotic Applications. Ph.D. Thesis, Universidade de Coimbra, Coimbra, Portugal, 2021.

[10] Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. arXiv 2016, arXiv:1603.00831.

[11] Yadav, S.; Payandeh, S. Understanding Tracking Methodology of Kernelized Correlation Filter. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 1330–1336.

[12] Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. Acm computing surveys (CSUR), 38(4), 13-es.

[13] Yang, H., Shao, L., Zheng, F., Wang, L., & Song, Z. (2011). Recent advances and trends in visual tracking: A review. Neurocomputing, 74(18), 3823-3831.

[14] Van de Weijer, J., Gevers, T., & Bagdanov, A. D. (2005). Boosting color saliency in image feature detection. IEEE transactions on pattern analysis and machine intelligence, 28(1), 150-156.

[15] Lukac, R., & Plataniotis, K. N. (Eds.). (2018). Color image processing: methods and applications. CRC press.

[16] Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International journal of computer vision, 81(1), 2-23.

[17] Winn, J., Criminisi, A., & Minka, T. (2005, October). Object categorization by learned universal visual dictionary. In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (Vol. 2, pp. 1800-1807). IEEE.

[18] Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. IEEE Transactions on pattern analysis and machine intelligence, 18(8), 837- 842.

[19] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7), 971-987.

[20] Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. Artificial intelligence, 17(1-3), 185-203.

[21] Lucas, B. D., & Kanade, T. (1981, April). An iterative image registration technique with an application to stereo vision.

[22] Sabzmeydani, P., & Mori, G. (2007, June). Detecting pedestrians by learning shapelet features. In 2007 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1- 8). IEEE.

[23] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[24] Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). Computer vision and image understanding, 110(3), 346-359.

[25] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE.

[26] Puri, N. V., & Devale, P. R. (2012). Development of human tracking in video surveillance system for activity analysis. IOSR J. Comput. Eng, 4(2), 26-30.

[27] Ramanan, D., Forsyth, D. A., & Zisserman, A. (2006). Tracking people by learning their appearance. IEEE transactions on pattern analysis and machine intelligence, 29(1), 65- 81.

[28] Ogale, N. A. (2006). A survey of techniques for human detection from video. Survey, University of Maryland, 125(133).

[29] Howlett, R. J., & Jain, L. C. (2005). Knowledge-based intelligent information and engineering systems. Springer Berlin/Heidelberg.

[30] Luo, R. C., Lin, T. Y., & Su, K. L. (2009). Multisensor based security robot system for intelligent building. Robotics and autonomous systems, 57(3), 330-338.

[31] Napper, S. A., & Seaman, R. L. (1989). Applications of robots in rehabilitation. Robotics and Autonomous Systems, 5(3), 227-239.

[32] Burgard, W., Cremers, A. B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., ... & Thrun, S. (1999). Experiences with an interactive museum tour-guide robot. Artificial intelligence, 114(1-2), 3-55.

[33] Asoh, H., Motomura, Y., Asano, F., Hara, I., Hayamizu, S., Itou, K., ... & Krose, B. (2001). Jijo-2: An office robot that communicates and learns. IEEE Intelligent Systems, 16(5), 46-55.

[34] Suzuki, S., Mitsukura, Y., Takimoto, H., Tanabata, T., Kimura, N., & Moriya, T. (2009, November). A human tracking mobile-robot with face detection. In 2009 35th Annual Conference of IEEE Industrial Electronics (pp. 4217-4222). IEEE.

[35] u, Y., & Huang, T. S. (2002). Nonstationary color tracking for vision-based humancomputer interaction. IEEE transactions on neural networks, 13(4), 948-960.

[36] Schlegel, C., Illmann, J., Jaberg, H., Schuster, M., & Wörz, R. (1998, September). Vision based person tracking with a mobile robot. In BMVC (pp. 1-10).

[37] Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. IEEE Transactions on pattern analysis and machine intelligence, 25(5), 564-577.

[38] Allen, J. G., Xu, R. Y., & Jin, J. S. (2004, June). Object tracking using camshift algorithm and multiple quantized feature spaces. In ACM International Conference Proceeding Series (Vol. 100, pp. 3-7).

[39] Gupta, M., Uggirala, B., & Behera, L. (2008). Visual navigation of a mobile robot in a cluttered environment. IFAC Proceedings Volumes, 41(2), 14816-14821.

[40] Grest, D., & Koch, R. (2004, September). Realtime multi-camera person tracking for immersive environments. In IEEE 6th Workshop on Multimedia Signal Processing, 2004. (pp. 387-390). IEEE.

[41] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). IEEE.

[42] Dixit, M., & Venkatesh, K. S. (2009, September). Combining edge and color features for tracking partially occluded humans. In Asian Conference on Computer Vision (pp. 140- 149). Springer, Berlin, Heidelberg.