**CHAPTER 1**

**INTRODUCTION**

**1.1 INTRODUCTION TO THE PROJECT**

In the modern world, public safety and surveillance have become critical aspects of urban infrastructure. With the exponential growth of CCTV installations, there is a growing demand for intelligent systems that can automate the process of monitoring and analyzing video footage. Manual monitoring is not only time-consuming but also prone to human error. This project aims to develop an automated CCTV Footage Person Attribute Extraction System using deep learning and computer vision techniques. The system will allow users to manually upload surveillance videos, which will then be analyzed to extract vital person attributes.

The key objective of this project is to identify and classify multiple soft biometric attributes from video frames, such as age, weight, height, clothing type, and walking style (gait). These attributes are extremely useful in forensic investigations, crowd analysis, and missing person identification. By extracting such information directly from video, authorities can gain insights into suspects or individuals of interest without needing clear facial recognition or ID verification, especially when camera quality or conditions are suboptimal.

To accomplish this, the system utilizes a pipeline of deep learning models that work in stages: first detecting humans in the video, then segmenting their bodies and faces, and finally estimating different attributes. The architecture integrates state-of-the-art models for detection and classification, trained on publicly available datasets. Gait and posture are analyzed using motion-based patterns extracted over sequential frames, whereas visual features like clothing and body proportions are handled using frame-wise object recognition models.

This project not only contributes to the growing field of smart surveillance but also addresses real-world challenges such as occlusion, low resolution, varying lighting conditions, and crowded scenes. It is designed to be robust, extensible, and usable in multiple environments, including public spaces, transportation hubs, and secure facilities. By automating the process of attribute extraction from surveillance videos, the system enhances both the efficiency and accuracy of person tracking and identification.

**1.2 INTRODUCTION TO TECHNOLOGIES USED**

Some of the core technologies used in this project is YOLO (You Only Look Once), Pose Estimation, Convolutional Neural Networks (CNNs) and temporal analysis models such as LSTM (Long Short-Term Memory).

**1.2.1 YOLO (You Only Look Once )** :

YOLO (You Only Look Once) is utilized as the primary object detection model to accurately identify and locate human figures in the frames of manually uploaded CCTV footage. YOLO divides each video frame into a grid and simultaneously predicts bounding boxes and class probabilities for each region, enabling fast and precise detection of persons. This allows the system to efficiently extract the relevant regions of interest, significantly reducing the computational load for subsequent attribute analysis stages. By leveraging YOLO’s robust detection capability, the system ensures high accuracy in identifying individuals even in complex environments, such as crowded scenes or low-quality footage, which are common in surveillance videos.

**1.2.2 Pose estimation** :

Pose estimation plays a vital role in understanding the body structure and movement patterns of individuals detected in CCTV footage. Once a person is localized in the video frame, pose estimation techniques are used to identify and map key body joints such as shoulders, elbows, knees, and ankles. These key points help define the person's posture and orientation, which is crucial for further analysis like gait recognition, height approximation, and walking pattern extraction. By analyzing the spatial relationships between joints over consecutive frames, the system can extract meaningful motion features without relying solely on facial data. This makes pose estimation especially valuable in cases where the face is occluded or too blurry for recognition. Additionally, pose data enhances the system’s ability to differentiate between clothing types based on body segments, improving the accuracy of overall person attribute extraction in surveillance videos.

**1.2.3 Convolutional Neural Networks (CNNs) :**

Convolutional Neural Networks (CNNs) are used as the core deep learning architecture for extracting and analyzing visual features from detected human regions in the uploaded CCTV video. After identifying and cropping the person from each frame using YOLO, CNNs process these regions to classify a variety of attributes such as age, clothing type, body shape, and facial features. CNNs are particularly well-suited for this task due to their ability to automatically learn hierarchical patterns in image data, such as textures, edges, and shapes, which are crucial for distinguishing subtle visual differences in human appearance. By leveraging CNNs, the system can perform robust and scalable attribute recognition even under variations in pose, lighting, and background clutter, making them essential for the success of person attribute extraction from surveillance footage.

**1.2.4 Long Short-Term Memory (LSTM) :**

Long Short-Term Memory (LSTM) networks are employed to recognize a person's gait or walking style from sequential video frames. Since gait is a temporal pattern that unfolds over time, LSTMs are well-suited for capturing such dynamic motion information. After detecting and tracking a person across multiple frames, relevant pose or body movement features are extracted and passed to an LSTM model. The LSTM processes these features in sequence, learning the temporal dependencies between body movements to identify unique gait characteristics. This enables the system to distinguish between different walking styles, which can serve as a soft biometric for person identification, even in cases where facial details are not clearly visible. LSTM’s ability to retain important motion cues over time makes it a critical component for gait recognition in surveillance-based attribute extraction.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 INTRODUCTION**

In recent years, the integration of deep learning and computer vision in surveillance systems has transformed the way human identification and behaviour analysis are performed. Traditional CCTV systems rely heavily on manual monitoring, which is time-consuming and error-prone. With the advancements in artificial intelligence, especially convolutional neural networks (CNNs) and pose estimation techniques, it is now possible to automatically extract detailed person attributes from video footage, including age, gender, height, weight, clothing type, and gait.

Our system leverages deep learning models to analyze video input which are uploaded manually. The system processes each frame to detect individuals and estimate their soft biometrics and physical attributes using image and motions .This capability is especially valuable for smart surveillance, public safety, and forensic investigations. To build a robust and accurate model, it is essential to review existing literature on person attribute recognition, gait analysis, pose estimation, and multi-modal biometric systems. The literature survey provides insights into previously proposed models, datasets, methodologies, and real-world challenges, forming a foundation for designing and improving the system.

**2.2 LITERATURE SURVEY**

Rohit Kumar Gupta et al. (2022), [1] proposed a CNN-based model for detecting age and gender from facial images, including real-time CCTV footage. The model uses convolutional, pooling, and fully connected layers with a softmax classifier to categorize gender (male or female) and age into specific groups. Trained on a Kaggle dataset with varied lighting and poses, the model treats age prediction as a classification task, enhancing accuracy in uncontrolled environments. This approach serves as a strong reference for our project, where we adapt their architecture and strategy for broader surveillance tasks like person identification and attribute-based filtering.

Sivachandiran et al. (2022), [2] developed an automated deep learning model named ADCNN-AGC for classifying age and gender from facial images in surveillance systems. The model uses MTCNN for detecting faces, EfficientNet for feature extraction, and 1D-CNN for classification. Tested on the UTKFace dataset, the model achieved 95.29% accuracy for gender and a mean absolute error of 2.89 in age prediction. Compared to other recent models like GRA-Net and RAN, ADCNN-AGC demonstrated superior results in both efficiency and accuracy. This research provides a robust and scalable solution for real-time demographic analysis and is highly relevant for CCTV-based systems where facial attributes need to be extracted quickly and accurately under uncontrolled conditions.

Nikouei et al. (2018), [3] proposed a real-time human detection system for edge computing environments using a Lightweight Convolutional Neural Network (L-CNN). Designed with resource constraints in mind, the model employs depthwise separable convolutions and is based on the SSD (Single Shot Multibox Detector) architecture, enabling efficient detection of human figures in surveillance footage. Tested on a Raspberry Pi 3, the L-CNN demonstrated competitive performance, achieving an average speed of 1.79 frames per second (FPS) and a false positive rate of 6.6%, while using significantly less memory than other standard models such as SSD-GoogleNet. This model is particularly relevant to edge-based smart surveillance systems, offering a viable approach for efficient person detection under limited hardware. Its application as a frontend human detector makes it a practical reference for projects like ours that require low-latency, high-accuracy person attribute extraction from CCTV footage.

Guruh Fajar Shidik et al. (2019), [4] conducted a systematic literature review analyzing 220 journal publications on intelligent video surveillance systems from 2010 to 2019. The study categorizes research trends into three main areas: visual surveillance, intelligent surveillance integration, and system infrastructure design. It provides a detailed overview of machine learning techniques—especially deep learning, SVM, and fuzzy logic—used for surveillance tasks such as object detection, behavior analysis, and activity recognition. The review highlights key public datasets and evaluates five widely cited surveillance frameworks (e.g., SSF, RISE, and EDCAR), offering a valuable foundation for modular and scalable surveillance solutions. Although the paper lacks experimental validation, it serves as a rich knowledge base for developing advanced systems. For our CCTV-based person attribute extraction project, this review offers strategic insights into system design, suitable datasets, and integration frameworks, supporting the development of a robust, real-time surveillance solution focused on identifying multiple soft biometric traits.

Hitesh Panchal (2016), [5] The paper "CCTV Video Abstraction and Object Detection for Video Surveillance System" by Hitesh Panchal introduces an innovative algorithm for key frame extraction from CCTV footage, addressing the challenges of analyzing extensive video data. By employing video segmentation and automatic shot boundary detection, the algorithm efficiently summarizes video content, allowing for quick retrieval of relevant frames. This work highlights the importance of intelligent video management in surveillance systems, paving the way for further research in person attribute extraction. The findings emphasize the potential for enhancing video analysis techniques, which can be beneficial for developing advanced surveillance applications.

Joseph Redmon et al. (2018), [6] The paper "YOLOv3: An Incremental Improvement" by Joseph Redmon and Ali Farhadi presents significant enhancements to the YOLO object detection framework. The authors introduce a new classifier network that improves accuracy while maintaining high processing speed, achieving 28.2 mAP at 320x320 resolution. YOLOv3 employs multiscale predictions and a multilabel classification approach, allowing for effective detection of overlapping labels. This work highlights the advancements in real-time object detection, making it a crucial reference for projects focused on person attribute extraction from CCTV footage, where speed and accuracy are paramount for effective surveillance analysis.

Xiao Ke et al. (2020), [7] introduced a deep learning approach for extracting human attributes from surveillance images. It integrates SSD-based pose estimation and multi-feature fusion to effectively identify clothing attributes, addressing issues like pixel resolution and background interference. This approach is particularly relevant for CCTV attribute extraction systems, where accurate human region isolation is crucial for robust performance.

Prof. Nandhini N et al. (2019), [8] discussed a deep learning approach for identifying anomalies in surveillance footage. It leverages CNNs for feature extraction and anomaly detection, providing a robust framework for real-time analysis of high-dimensional data. This method can be adapted for person attribute extraction by focusing on specific human characteristics and movement patterns, making it a valuable reference for surveillance systems that require precise behavior monitoring.

Hiren Galiyawala et al. (2022), [9] presented a deep learning approach for person identification based on soft biometrics like age and clothing type. It uses Mask R-CNN for accurate person detection and attribute recognition, achieving high retrieval accuracy with fewer attributes. This method is highly relevant for CCTV-based person attribute extraction systems, providing a streamlined approach to identifying individuals in complex surveillance environments.

Fabbri et al. (2017), [10] proposed a tri-network approach (ResNet classifier + occlusion-resistant DCGAN + super-resolution DCGAN) for low-resolution surveillance video-based attribute classification, achieving a state-of-the-art benchmark on RAP under 80% occlusion/low resolution. It outperformed DeepMAR/ACN by >6% mAP by recovering classifiable features. The module-based architecture is stronger but limits real-time use.

Shoitan et al. (2023 ), [11] proposed a spatio-temporal person retrieval method in video surveillance using a combination of ByteTrack for robust tracking and two attribute recognition models—APR and ALM—to provide higher accuracy. Unlike conventional methods, their method relates the bounding boxes from frames to reduce detection errors and enhance attribute recognition. Evaluated on the SoftBioSearch dataset, the system achieved a 93.21% true positive, 14% better than state-of-the-art. While it performs well in occlusion and low visibility, its reliance on advanced tracking can be an issue for real-time applications. However, it addresses a significant loophole in attribute-based person retrieval.

Yaghoubi et al. (2020), [12] provided an exhaustive survey of Human Attribute Recognition (HAR), condensing state-of-the-art contributions in terms of most significant challenges such as data imbalance, occlusion, and attribute correlation. Different from previous surveys, they formulated a challenge-oriented taxonomy and critically examined deep learning methods, datasets, and measures, including sub-areas such as pedestrian and clothing attribute recognition. The survey recognizes gaps in the literature including the absence of integrated data, occlusion, and model explainability. It recognizes the use of CNNs, GCNs, and RNNs in filling the gaps, providing insightful information in the design of more trustworthy and interpretable HAR systems.

Amirgaliyev et al. (2025), [13] gave a comprehensive overview of ML and DL methods for person detection, tracking, identification, and face recognition, focusing on the shift from traditional features to deep CNNs like YOLO and FaceNet. Using the PRISMA method, they evaluated over 140 articles and encountered issues of occlusion, night vision, and ethical concerns. The research points to efficient, privacy-sensitive models and rich data sets as crucial, with directions for future research in smart surveillance systems suggested.

Haritha et al. (2025), [14] developed an AI-powered surveillance framework using YOLOv8 for object detection and LSTM for anomaly detection to promote public safety through automated crowd monitoring and prevention of crime. The system can operate in real-time using CCTV footage to monitor crowd density and spotting suspicious activity so that alerts can be given when an anomaly or over-crowding occurs. The authors report high detection results (95.4%) and anomaly detection recognition (92.7%) along with a 30% computation overhead; allowing the framework to be built on existing CCTV hardwares. The model uses contextual filtering and low latency processing to assist secure alterations, and scalability or operational effectiveness in workplace environments, public areas and industrial locations. While the framework showed a successful proof of concept, the system's ability to be optimized to other environments, along with the ambiguity of privacy implications will require future evaluation.

Amirgaliyev et al. (2025), [15] presented a systematic review of over 140 studies focused on machine learning and deep learning techniques for person detection, tracking, identification, and face recognition. They analyze classical approaches like HOG and Kalman filters alongside modern deep models such as YOLO, ArcFace, and DeepSORT. The paper highlights real-world applications in surveillance, transportation, and smart cities while addressing challenges such as occlusion, real-time constraints, and ethical concerns. This review serves as a valuable reference for developing intelligent video surveillance systems, especially for projects involving CCTV-based person attribute extraction using deep learning.

Iyshwarya Ratthi et al. (2024), [16] introduced an AI-based human height estimation model for surveillance, leveraging monocular cameras and YOLOv7 with a hybrid attention mechanism (HAM). Designed to aid in missing child retrieval, the system uses camera calibration and a new dataset (“Sense-Height”) featuring adults and children. Unlike traditional models, this approach handles occlusion, diverse lighting, and motion conditions with high accuracy (error as low as 0.02 cm). The paper provides strong empirical validation and proposes a field-of-view (FOV) zoning strategy. This work is significant for integrating height as a soft biometric in intelligent video surveillance.

Taha et al. (2024), [17] proposed a gait recognition model using IMU data instead of conventional video. Their system collects gait features from shoe-embedded sensors and processes them using stacked sparse autoencoders. The high-level features are then clustered to identify physical characteristics like age, gender, and body size. The model shows greater robustness to occlusion and environmental variation than visual gait recognition systems. While it is not directly usable for CCTV-based projects, its deep learning approach and gait-based soft biometric extraction provide a strong conceptual base for designing attribute recognition models using motion cues in surveillance footage.

Gururaj et al. (2024), [18] presented a detailed review of face recognition (FR) systems, covering traditional techniques like PCA and LDA, and advanced deep learning methods including CNN-based hybrid models. The paper explores FR challenges such as pose variation, occlusion, and aging, while classifying existing approaches into appearance-based, landmark-based, and hybrid methods. It also discusses video-based FR systems, dataset availability, and future directions. Although it does not introduce new models, this survey offers valuable insights into the selection of algorithms and datasets that can aid in developing accurate and real-time person attribute recognition from CCTV surveillance footage.

**2.3 SUMMARY OF LITERATURE SURVEY**

Table 2.1 shows the summary of literature survey done.

**Table 2.1: Observations of Literature Survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author Name** | **Title of Paper** | **Methodology used** | **Advantages** | **Future Work** |
| Rohit Kumar Gupta, Shivaprasad M B, Dr. S. Srividhya | Age & GenderDetection using Convolutional Neural Network. | Multi-Branch Deep Neural Network architecture (ReLU, pooling, normalization). | Simple CNN architecture. Utilizes Keras, TensorFlow, OpenCV. Classifies into defined age/gender classes. | Improve accuracy under challenging conditions. Expand attributes beyond age/gender. Real-time multi-person detection. |
| S. Sivachandiran, Dr. K. Jagan Mohan, Dr. G. Mohammed Nazer | Automated Deep Learning based Age and Gender Classification Model using Facial Features for Video Surveillance. | ADCNN-AGC (Automated Deep Convolutional Neural Network for Age and Gender Classification) model. | Employs MTCNN for robust face detection. High accuracy (95.29%). | Improve lightweight deployment for edge devices. Real-time optimization for surveillance. |
| S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan | Real-Time Human Detection as an Edge Service Enabled by a Lightweight CNN. | Lightweight Convolutional Neural Network (L-CNN) optimized for real-time human detection. | Low computational complexity, real-time performance (1.79–2.06 FPS), low memory usage. | Expand to full attribute recognition- Integrate tracking and behavior analysis. Support multiple human instances. |
| G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto, and E. J. Kusuma | A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets. | Deep learning techniques such as (YOLOv5, MobilNetv2, Local Binary Pattern Histogram). | Covers 220 studies from 2010–2019- Identifies trends, datasets, ML methods. Reviews use cases like crime, traffic, healthcare. | Guide future framework development- Build unified benchmarks.  Improve integration of multi-sensor data. |
| Hitesh Panchal | CCTV Video Abstraction and Object Detection for Video Surveillance System. | Key-frame extraction using histogram matching and shot boundary detection. | Reduces data volume, improves efficiency in browsing and object detection. | Extend to larger datasets and real-time applications. |
| J. Redmon and A. Farhadi | YOLOv3: An Incremental Improvement. | Improved YOLO architecture with Darknet-53 for object detection. | Faster and more accurate than previous versions and competitors. | Explore applications in real-time systems and ethical implications. |
| Xiao Ke, T. Liu, and Z. Li | Human Attribute Recognition Method Based on Pose Estimation and Multiple-Feature Fusion. | Pose estimation with SSD, multi-feature fusion, and MAP allocation. | Improved accuracy in attribute recognition under poor conditions. | Integrate more attributes and improve robustness. |
| Nandhini N et al. | Anomaly Detection System in CCTV Derived Videos. | CNN-based deep learning for anomaly detection in surveillance videos. | High accuracy in detecting anomalies, adaptable to different scenarios. | Enhance real-time performance and reduce false positives. |
| Hiren Galiyawala et al. | Person Retrieval in Surveillance Videos Using Attribute Recognition. | Mask R-CNN for detection, attribute weighting, and ranking. | State-of-the-art performance with fewer attributes. | Address gender bias and improve attribute recognition models. |
| Matteo Fabbri et al. | Generative Adversarial Models for People Attribute Recognition in Surveillance. | DCGAN for image enhancement, part-based attribute classification. | Handles occlusion and low resolution effectively. | Develop an end-to-end model for automatic enhancement selection. |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**2.4 COMPARISON WITH EXISTING SYSTEMS**

Traditional CCTV analysis systems predominantly rely on manual monitoring, basic motion detection, or facial recognition. These methods, while foundational, suffer from critical limitations. Manual observation is slow, exhausting, and vulnerable to human error—especially during extended surveillance tasks. Motion detection can only identify movement, offering no information about the person involved. Facial recognition, though more advanced, is heavily dependent on high-resolution images and favorable lighting conditions, often failing when faces are angled, occluded, or poorly lit. Additionally, many existing research models focus on isolated attributes such as age or gender, lacking the capability to integrate and analyze multiple traits simultaneously. Commercial surveillance tools, on the other hand, tend to be expensive and limited to surface-level analytics, making them impractical for dynamic, real-world security environments.

Our project—CCTV Footage Person Attribute Extraction System—offers a significant leap forward. Unlike traditional approaches, our system performs real-time, multi-attribute extraction using advanced computer vision and deep learning techniques. We utilize YOLO for accurate and fast person detection and ByteTrack for reliable tracking. Once individuals are detected, we apply deep learning models like EfficientNet and processing tools such as OpenCV to extract key attributes such as height, weight, age, clothing color, body tone, and gait from CCTV footage.

This integration allows us to generate detailed, actionable profiles of individuals that can aid in faster and more informed decision-making in surveillance, forensic investigations, and public safety scenarios. Our system also features real-time dashboards using Streamlit or Grafana, enabling security personnel to visualize insights instantly. Alerts can be sent through platforms like Twilio or Slack, allowing immediate responses to suspicious activities. By combining speed, accuracy, and usability, our system fills the gaps left by conventional solutions and sets a new benchmark for intelligent video surveillance.

**2.5 PROPOSED SYSTEM**

The proposed CCTV Footage Person Attribute Extraction System leverages cutting-edge computer vision and deep learning techniques to automate the process of identifying and extracting key human attributes from surveillance videos. At the core of this system lies a robust video analysis pipeline, which takes in manually uploaded CCTV footage and transforms it into actionable insights. The process begins with the detection of individuals using high-performance object detection algorithms such as YOLO, which accurately identifies human figures even in crowded or low-quality frames. This detection is followed by tracking the individual’s movement across frames using ByteTrack, ensuring continuity and reducing duplication errors.

Once the individual is tracked, the system proceeds to the attribute extraction stage. Here, deep learning models like EfficientNet and frameworks like OpenCV are employed to identify and classify multiple human traits such as height, weight, age group, clothing color, and gait. Unlike systems that rely solely on facial recognition or single-feature analysis, our solution offers a comprehensive profile generation based on soft biometrics and body posture, making it highly effective even when facial visibility is limited due to occlusions, low resolution, or poor lighting.

To make this technology practically useful for law enforcement and security teams, the system includes a real-time visualization dashboard built using platforms like Streamlit or Grafana. These dashboards display extracted attributes alongside corresponding video segments, enabling fast interpretation and decision-making.

Scalability and performance have been core design priorities. The backend is optimized to handle large volumes of video input and concurrent processing tasks using parallel computing strategies and hardware acceleration, ensuring that the system remains responsive during intensive operations. Cloud integration is supported to facilitate remote deployment, centralized data management, and elastic resource allocation based on real-time demand.

Moreover, the system emphasizes user-friendliness and operational adaptability. With minimal manual intervention, it can be operated by personnel without deep technical expertise. The modular architecture allows for easy updates, model replacements, or integration with existing surveillance infrastructure. Security and data privacy are also addressed through local processing options and encrypted storage, safeguarding sensitive footage from unauthorized access.

Overall, this system bridges the gap between traditional CCTV monitoring and intelligent surveillance. It transforms raw footage into rich, structured information, empowering security personnel with faster, smarter, and more informed decision-making capabilities in real-time environments such as public places, transportation hubs, and critical infrastructure zones.

**2.6 OBJECTIVES**

* To develop a system that automatically extracts multiple human attributes such as height, weight, age, clothing color, and gait from CCTV footage.
* To implement advanced object detection and tracking techniques like YOLO and ByteTrack to accurately locate and follow individuals in video streams.
* To apply deep learning models such as EfficientNet along with image processing libraries like OpenCV for reliable multi-attribute recognition.
* To reduce dependency on manual CCTV monitoring by automating the identification process and improving accuracy in real-time surveillance scenarios.
* To ensure the system can function effectively under challenging conditions like low resolution, poor lighting, occlusion, and crowd density.
* To create detailed individual profiles using extracted attributes that can assist in forensic investigations, missing person cases, and security monitoring.
* To build an intuitive and real-time dashboard using tools like Streamlit or Grafana for visualizing attribute data and person profiles.
* To design a scalable backend system that can efficiently handle large volumes of video data and support cloud-based deployment.
* To create a modular, flexible, and user-friendly architecture that can be easily integrated into existing surveillance systems with minimal setup.

**REFERENCES**

[1] R. K. Gupta, M. B. Shivaprasad and S. Srividhya, "Age & Gender Detection using Convolutional Neural Network," *International Journal of Engineering Research & Technology (IJERT)*, vol. 11, no. 6, pp. 441–443, Jun. 2022.

[2] S. Sivachandiran, K. J. Mohan, and G. M. Nazer, "Automated Deep Learning based Age and Gender Classification Model using Facial Features for Video Surveillance," *Journal of Algebraic Statistics*, vol. 13, no. 2, pp. 621–633, 2022.

[3] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Real-Time Human Detection as an Edge Service Enabled by a Lightweight CNN," *arXiv preprint*, arXiv:1805.00330, Apr. 2018.

[4] G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto, and E. J. Kusuma, "A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets," *IEEE Access*, vol. 7, pp. 170457–170480, 2019, doi: 10.1109/ACCESS.2019.2955387.

[5] H. Panchal, "CCTV Video Abstraction and Object Detection for Video Surveillance System," *International Journal of Current Research*, vol. 8, no. 1, pp. 25277–25280, Jan. 2016.

[6] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint*, arXiv:1804.02767, Apr. 2018.

[7] X. Ke, T. Liu, and Z. Li, "Human Attribute Recognition Method Based on Pose Estimation and Multiple-Feature Fusion," *Signal, Image and Video Processing*, vol. 14, pp. 1441–1449, Apr. 2020, doi: 10.1007/s11760-020-01690-8.

[8] N. Nandhini, M. R. Barath Kumar, L. Sharma, and A. Gupta, "Anomaly Detection System in CCTV Derived Videos," *International Research Journal of Engineering and Technology (IRJET)*, vol. 6, no. 5, pp. 1202–1204, May 2019.

[9] H. Galiyawala, M. S. Raval, and M. Patel, "Person Retrieval in Surveillance Videos Using Attribute Recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 1–17, May 2022, doi: 10.1007/s12652-022-03891-0.

[10] M. Fabbri, S. Calderara, and R. Cucchiara, "Generative Adversarial Models for People Attribute Recognition in Surveillance," *in Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Lecce, Italy, Aug. 2017, pp. 1–6, doi: 10.1109/AVSS.2017.8078511.

[11] R. Shoitan, M. M. Moussa, and H. A. El Nemr, "Attribute-based Spatio-Temporal Person Retrieval in Video Surveillance*," Alexandria Engineering Journal*, vol. 63, pp. 441–454, 2023, doi: 10.1016/j.aej.2022.07.053.

[12] E. Yaghoubi, F. Khezeli, D. Borza, S. V. A. Kumar, J. Neves, and H. Proença, "Human Attribute Recognition—A Comprehensive Survey*," Applied Sciences*, vol. 10, no. 16, p. 5608, Aug. 2020, doi: 10.3390/app10165608.

[13] J. C. Antony, C. L. S. Chowdary, N. P. B., E. Murali, and A. Mayan, "Advancing Crowd Management through Innovative Surveillance using YOLOv8 and ByteTrack*," in Proc. 2024 Int. Conf. on Advances in Computing, Communication, and Control (ICAC3),* 2024, pp. 1–6.

[14] K. Haritha, N. S. Geethika, K. Venkateswarlu, R. H. Kumar, and Y. Ramakrishna, "Enhancing Public Safety with AI & ML-Based CCTV Surveillance," *International Journal for Modern Trends in Science and Technology (IJMTST)*, vol. 11, no. 3, pp. 322–331, Apr. 2025, doi: 10.5281/zenodo.15171437.

[15] B. Amirgaliyev, M. Mussabek, T. Rakhimzhanova, and A. Zhumadillayeva, "A Review of Machine Learning and Deep Learning Methods for Person Detection, Tracking and Identification, and Face Recognition with Applications," *Sensors*, vol. 25, no. 5, p. 1410, Feb. 2025, doi: 10.3390/s25051410.

[16] K. I. Ratthi, B. Yogameena, and S. S. Perumaal, "Human Height Estimation Using AI-Assisted Computer Vision for Intelligent Video Surveillance System*," Measurement*, vol. 236, p. 115133, Jun. 2024, doi: 10.1016/j.measurement.2024.115133.

[17] K. Taha, P. D. Yoo, Y. Al-Hammadi, S. Muhaidat, and C. Y. Yeun, "Learning a Deep-Feature Clustering Model for Gait-Based Individual Identification*," Computers & Security*, vol. 136, p. 103559, Oct. 2023, doi: 10.1016/j.cose.2023.103559.

[18] H. L. Gururaj, B. C. Soundarya, S. Priya, J. Shreyas, and F. Flammini, "A Comprehensive Review of Face Recognition Techniques, Trends, and Challenges," *IEEE Access*, vol. 12, pp. 107903–107921, Aug. 2024, doi: 10.1109/ACCESS.2024.3424933.