Dense crowds analysis: Challenges and state of the art

November 8, 2025

Abstract

Dense crowd analysis is crucial for ensuring safety, security and good resources usage during large-scale gatherings, particularly the Hajj and Umrah pilgrimages, where high crowd density, limited available space and continuous motion, challenge conventional computer vision and deep learning models. Although lot of progress has been made in crowd counting, tracking, and identification, existing research remains fragmented, and to our knowledge, no dedicated state-of-the-art review focused on pilgrimage crowds currently exists. This work addresses that gap by synthesizing recent AI-based approaches and evaluating their applicability under extreme cultural, visual, and logistical conditions of the Hajj.

Dense crowds, Hajj, Umrah, CNN, Dataset, Related work

1 Introduction

Monitoring and analyzing the dynamics of dense crowds is a critical research area with profound implications for public safety, infrastructure planning, and large-scale event management. Among the most challenging and significant environments for crowd analysis are the annual Islamic pilgrimages of Hajj and Umrah in Mecca, where millions of individuals converge in confined spaces within limited timeframes. These gatherings present extreme cases of high-density crowds, characterized by continuous motion, limited maneuverability, frequent occlusions, and cultural or religious behaviors that constrain individual autonomy. Effective crowd management in such contexts necessitates robust systems capable of real-time counting, individual identification, trajectory tracking, and motion analysis.

Traditional computer vision methods often fail under such dense and dynamic conditions, motivating the adoption of artificial intelligence (AI) techniques, particularly those based on deep learning. Neural networks, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures, have demonstrated substantial improvements in handling occlusion, scale variation, and scene complexity. These models can extract and integrate spatiotemporal features, enabling accurate people counting, identity tracking, and motion prediction even in highly congested scenarios like the Mataf area around the Kaaba or during the Tawaf and Sa'i rituals.

This article provides a focused review of recent AI-driven approaches for crowd monitoring, with an emphasis on the four core challenges: (1) crowd counting in dense environments, (2) individual identification and re-identification in large, homogeneous groups, (3) multi-object tracking under continuous motion and occlusion, and (4) speed and trajectory estimation at individual and collective levels. We pay particular attention to research applied to or inspired by the specific conditions of the Hajj and Umrah pilgrimages, where data scarcity, privacy constraints, and religious sensitivities pose additional challenges to algorithm design and deployment.

By synthesizing state-of-the-art techniques and analyzing their applicability to pilgrimage crowds, this paper aims to inform both academic research and practical implementations in real-world crowd monitoring systems. We highlight the key neural network models employed, evaluate their strengths and limitations in high-density contexts, and outline open challenges and future directions for AI-based tracking and movement analysis in sacred and sensitive environments.

2 Background - Key Concepts

This section presents the key concepts and issues most commonly addressed in crowd monitoring and management.

2.1 Crowd Tracking

Crowd tracking is the process of detecting and following the movements of individuals or groups across consecutive video frames to understand behaviour and flow patterns in crowded environments. This task usually begins with object detection (e.g., using YOLOv5/YOLOv8 or CenterTrack), followed by multi-object tracking (MOT) algorithms like ByteTrack and FairMOT. These methods combine appearance features, spatial positioning, and motion consistency to maintain accurate tracking of individuals even in dense, occluded scenes ([36]. Recent advancements such as OCSORT and StrongSORT have further improved tracking robustness under complex motion and occlusions ([9], [11]).

2.2 Crowd Counting

Crowd counting aims to estimate the number of individuals in a particular scene, especially in conditions of high density or occlusion where traditional object detection techniques fail. Instead of detecting each individual explicitly, modern methods rely on density map estimation, in which a convolutional neural network predicts a density value for each pixel. The sum of these values gives the estimated crowd count (Li et al. 2018).

One of the foundational models in this area is CSRNet, which uses dilated convolutions to capture a larger context without reducing resolution, allowing high accuracy density estimation in complex scenes ([22]). One of the foundational models in this area is CSRNet, which uses dilated convolutions to capture a larger context without reducing resolution, allowing for high-accuracy density estimation in complex scenes ([22]). Recent models improve on this by incorporating attention and semantic refinement. For instance, SSPNet enhances performance by integrating semantic priors into a spatial pyramid module (Zhao Wang, 2022), while MAN leverages hierarchical attention to better focus on relevant areas in variable crowd densities (Lin et al., 2022).

2.3 Crowd Control and Management

Crowd control involves proactive measures to ensure the safe movement and behaviour of individuals within a space, often involving physical interventions like barriers or flow redirection. In contrast, crowd management encompasses both technological and logistical approaches to planning and monitoring large gatherings.

AI-driven crowd management systems combine real-time data from sensors, cameras, and social media to predict congestion, detect anomalies, and suggest interventions (Madhu et al., 2025; Springer, 2024). These systems are increasingly using simulation-based decision support models, digital twins, and predictive analytics for risk assessment and strategic planning (MDPI, 2024).

2.4 Comparison of Tasks

While crowd tracking and counting are often used together, they serve different purposes. Tracking focuses on movement analysis and identity preservation, making it ideal for behaviour prediction and anomaly detection. Counting provides aggregate statistics used in capacity planning and density estimation. Crowd control and management, on the other hand, focus on macro-level interventions and planning, relying heavily on the outputs of tracking and counting systems to drive decisions (Madhu et al., 2025; Springer, 2024).

2.5 Deep Learning in Crowd Analysis

Deep learning has become central to crowd analysis due to its strength in extracting meaningful features and recognizing complex spatial-temporal patterns. Convolutional Neural Networks (CNNs) are commonly used to extract spatial features, such as density maps or individual object representations, from raw video or image data (Li et al., 2018). These features are then used for downstream tasks like crowd counting, tracking, anomaly detection, and crowd flow analysis.

Beyond CNNs, advanced models now integrate Recurrent Neural Networks (RNNs) or Graph Neural Networks (GNNs) to capture temporal and relational dynamics, while transformer-based architectures are emerging for attention-based scene understanding (Mohamed et al., 2021; Huang et al., 2023).

3 Methods and Current Approaches

Crowd analysis has been approached through a variety of computational techniques, broadly classified into two categories: traditional image processing methods and modern deep learning-based frameworks. Each class brings distinct advantages and limitations, particularly in the context of extreme crowd densities such as those observed during the Hajj pilgrimage.

3.1 Traditional Techniques

Traditional methods for crowd analysis rely on low-level visual cues and handcrafted algorithms. These approaches are relatively lightweight and require less computational power, making them suitable for early implementations and real-time surveillance on limited hardware.

3.1.1 Optical Flow

Optical flow techniques compute pixel-wise motion vectors by analyzing intensity changes between consecutive frames [16]. These vectors can be aggregated to infer global motion patterns of crowds, detect abnormal behaviors, or estimate local velocities. Optical flow is especially effective in moderate-density crowds where occlusions are minimal.

However, in high-density settings, such as during Hajj, overlapping individuals and homogeneous appearances significantly degrade its accuracy. Noise from small camera movements, lighting changes, or shadows can also distort flow fields. Improvements such as dense optical flow and multi-resolution approaches attempt to enhance robustness, but they still struggle under occlusion-heavy scenarios.

3.1.2 Background Subtraction

Background subtraction separates moving foreground objects from static backgrounds by constructing a model of the scene over time. Popular techniques include Gaussian Mixture Models (GMM) [30], ViBe, and codebook models. These methods have seen widespread use in surveillance systems due to their computational efficiency and adaptability.

In crowded scenarios, however, traditional background subtraction faces numerous challenges. Constant motion in dense crowds prevents stable background modeling. Moreover, environmental factors like dynamic lighting, shadows, and repetitive movements lead to high false-positive rates. Advanced methods incorporate shadow detection and adaptive thresholds to mitigate these effects [8, 29], but they remain insufficient for complex crowd patterns.

3.1.3 Bloc Detection and Clustering

Some traditional systems segment moving regions into blobs (connected components), which are then tracked over time. Blob features such as size, position, and shape are used to approximate the number of people. Although computationally simple, blob-based methods lack precision in dense environments where individuals are not visually separable. Overlapping blobs or merged contours result in undercounting or identity confusion.

3.2 Limitations of Traditional Methods

Despite their simplicity and low computational requirements, traditional computer vision methods exhibit several limitations that hinder their applicability in high-density, dynamic environments such as the Hajj pilgrimage.

- **Occlusion Sensitivity:** One of the primary limitations of traditional approaches like optical flow and blob tracking is their poor performance in the presence of partial or full occlusions. In dense crowds, individuals frequently overlap, making it difficult to isolate motion vectors or distinguish separate entities. This results in inaccurate tracking, fragmented trajectories, or merged detections, especially when camera views are not elevated or panoramic [**?**].
- **Scalability and Generalization Issues:** These methods are often rule-based and tuned for specific environments. They struggle to generalize to unseen conditions, such as different camera angles, lighting variations, or crowd densities. In events like the Hajj, where density and environmental conditions vary drastically over time and space, traditional models require continuous manual adjustment and are thus not scalable for long-term deployment [**?**].

- **Inadequate Feature Representation:** Traditional methods rely on handcrafted features such as edges, gradients, or background color models. These low-level features are insufficient to encode complex semantics like behavioral patterns, visual clothing cues (e.g., white ihram garments), or group dynamics. Consequently, these systems cannot distinguish between normal and abnormal crowd behaviors or identify individuals in visually homogeneous scenes [**?**].
- **Lack of Temporal Consistency:** Background subtraction and blob analysis often work frame-by-frame and lack memory of temporal context. This can cause abrupt object disappearance, flickering false positives, or identity switches when tracking through congested areas.
- **Vulnerability to Environmental Noise:** Minor environmental changes such as shadows, reflections, camera shake, or lighting fluctuations can significantly degrade the performance of traditional techniques. Even small scene disturbances can trigger misclassifications, especially in algorithms relying on fixed thresholds or non-adaptive models [8].
- **Limited Use in Predictive Analytics:** These methods generally serve as detection tools but cannot perform high-level inference tasks like motion prediction, density forecasting, or individual re-identification. This limits their utility in proactive crowd management systems that require anticipatory capabilities.

In summary, while traditional techniques provide a computationally efficient foundation for early-stage crowd monitoring, they fall short in the face of real-world complexities such as high-density human flows, dynamic behavior, and environmental variability. Their inability to semantically interpret scenes or adapt to rapidly changing conditions underscores the need for deep learning-based approaches that can learn robust and transferable representations from large-scale data.

3.3 Deep Learning-Based Methods

With the rise of large annotated datasets and GPU-accelerated computing, deep learning has significantly advanced crowd analysis. Unlike traditional approaches, deep learning models learn hierarchical feature representations directly from data, enabling better generalization and robustness in complex environments.

3.3.1 CNN-Based Models

Convolutional Neural Networks (CNNs) extract multiscale spatial features from input images. Early CNN-based models such as MCNN (Multi-column Convolutional Neural Network) by Zhang et al. [37] use parallel convolutional branches with varying kernel sizes to capture information at different densities. This architecture proved effective in mixed-density scenes.

Subsequent models like Switch-CNN introduced region-wise adaptation to density levels, while SANet incorporated self-attention mechanisms for enhanced focus on dense regions. These models outperform classical techniques on public datasets such as ShanghaiTech and UCF-QNRF.

3.3.2 Density Map Estimation

Instead of detecting individuals, many modern methods estimate a continuous density map, where each pixel indicates crowd concentration. The total count is obtained by summing the density map. CSRNet [23], a leading architecture in this domain, uses dilated convolutions to maintain receptive field size without downsampling, preserving fine-grained spatial details in dense scenes.

These methods are particularly effective for crowd counting in extreme densities where detection-based models fail. They also serve as a precursor for anomaly detection and flow estimation.

3.3.3 Transformer and Attention-Based Architectures

Inspired by advancements in natural language processing, transformer-based models are being introduced for crowd analysis. Vision Transformers (ViTs) and attention-based CNN hybrids offer superior long-range context modeling and robustness to scale variation. For instance, models like CrowdFormer leverage self-attention to correlate features across distant regions, improving detection accuracy in highly congested settings [32].

3.3.4 Multi-task Learning Approaches

Recent research has explored joint learning of multiple tasks such as counting, segmentation, and tracking within a single network. This allows shared representations and reduces redundant computation. Multi-task networks can simultaneously produce density maps, semantic segmentations, and motion trajectories, which is beneficial for comprehensive crowd monitoring.

3.4 Limitations of Deep Learning Methods

Despite their superior performance, deep learning models also face limitations:

- **Data Dependence:** They require large, annotated datasets for effective training, which may be unavailable for specific scenarios like the Hajj.
- **Computational Overhead:** High inference latency can hinder real-time applications unless optimized for edge devices.
- **Domain Shift Sensitivity:** Pretrained models often fail when deployed in novel environments due to differences in camera angle, crowd behavior, or lighting.

Nevertheless, ongoing research into lightweight architectures, transfer learning, and synthetic data augmentation is addressing these challenges. The integration of deep models into embedded systems and smart cameras makes them increasingly viable for field deployment.

4 Hajj-Specific Applications Specificity

The Hajj pilgrimage represents one of the largest annual human gatherings, attracting over two million pilgrims to Mecca. Crowd scenarios observed during the Hajj and Umrah pilgrimages represent some of the most densely populated and dynamic human gatherings in the world, with millions of individuals moving simultaneously within confined sacred spaces such as the Grand Mosque in Mecca. Unlike conventional urban or event-based crowds, these gatherings often exceed critical density thresholds, where personal space is minimal and occlusion rates are extremely high [3**?**].

This level of density presents significant challenges for existing deep neural network (DNN)-based monitoring systems, which are typically trained and validated on datasets representing more heterogeneous, lower-density, or urban pedestrian environments [17, 38]. Furthermore, the unique dress code followed by pilgrims—such as the white ihram garments worn by all male participants—reduces the visual variability that DNN models rely on for tasks such as individual identification, re-identification, and tracking [6]. The homogeneity in clothing color, texture, and silhouette complicates the extraction of discriminative features and increases the likelihood of identity switches and tracking failures. In addition, minimal visual accessories and head coverings reduce the effectiveness of appearance-based models [28].

As a result, direct application of standard DNN-based approaches in these contexts often yields suboptimal performance, emphasizing the need for specialized models and datasets that account for the visual and behavioral characteristics specific to pilgrimage crowds.

Artificial Intelligence (AI) technologies have been increasingly employed to address these challenges, tailored to the unique cultural and environmental context of the Hajj.

Context-Aware Model Customization

Conventional AI models often underperform in the Hajj setting due to specific factors:

- **Homogeneous Attire:** Pilgrims typically wear similar clothing (e.g., white ihram for men [**?**]), complicating individual identification by computer vision systems.
- **Ritual-Specific Movements:** Activities such as Tawaf (circumambulation of the Kaaba as shown in 2) and Sa'i (walking between Safa and Marwah) involve unique movement patterns not commonly found in other crowd scenarios [3, 6].
- **Group Dynamics:** Pilgrims often move in organized groups based on nationality or language, requiring models to account for group behaviors and leader-follower dynamics [13].
- **Fixed Camera Constraints:** Surveillance systems are often limited to fixed positions due to infrastructural constraints, affecting the adaptability of AI models to different viewpoints [26].

Figure 1: Pilgrims wearing Ihram

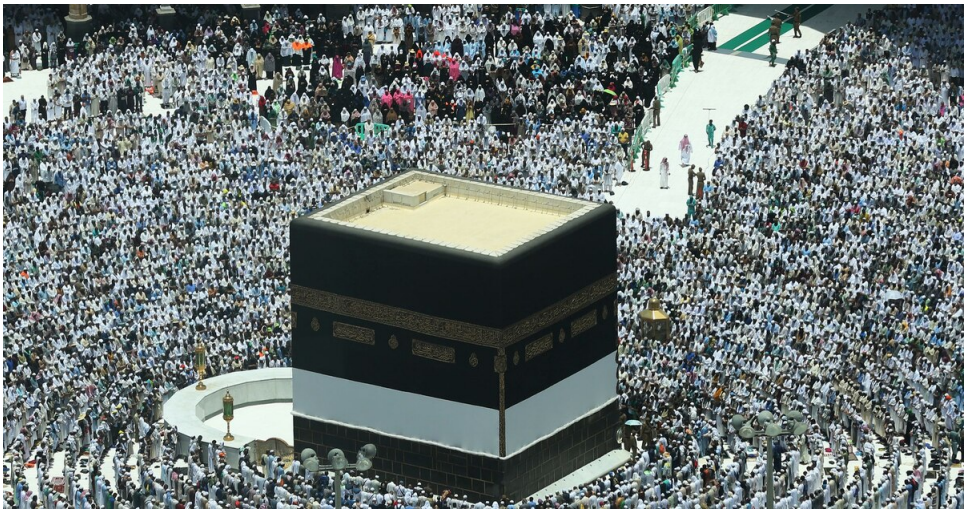

Figure 2: Pilgrims during Tawaf around Kaaba

To address these issues, AI systems are:

- **Fine-Tuned on Hajj-Specific Datasets:** Utilizing datasets that reflect the unique attire, movement patterns, and crowd densities of the Hajj [6, 15].
- **Adjusted for Environmental Conditions:** Accounting for factors such as low lighting in prayer halls, outdoor heat distortion, and varying crowd flows during different times of the day [2, 26].

Real-World Implementations

Several AI-driven initiatives have been deployed to enhance crowd management during the Hajj:

- **Surveillance and Crowd Heatmaps:** Vision-based systems generate real-time density heatmaps around critical areas like the Grand Mosque and Jamarat Bridge, aiding in monitoring and decision-making [26].
- **Congestion Management:** AI systems alert authorities when crowd density exceeds safe thresholds, enabling timely interventions such as rerouting or pausing inflows [2].
- **Pilgrim Identification Systems:** Facial recognition and RFID technologies assist in identifying and locating missing pilgrims, particularly children and the elderly [26].
- **Aerial Monitoring with Drones:** Drones equipped with AI-based object detection monitor large areas from above, especially during peak events like Eid prayers [1, 2].
- **Simulation-Based Planning:** Authorities employ AI-powered simulation tools to model various crowd control scenarios before the Hajj season, optimizing spatial arrangements and emergency response strategies [13].

5 Datasets and Benchmarks

One of the most significant limitations in advancing deep neural network (DNN)-based crowd monitoring systems is the scarcity of large-scale, high-quality datasets representative of real-world dense crowd scenarios. Effective training of DNNs typically requires extensive annotated data capturing diverse viewpoints, crowd densities, lighting conditions, and movement patterns. However, collecting such data in high-density environments—particularly during religious gatherings like Hajj and Umrah—poses ethical, legal, and logistical challenges. Privacy concerns are paramount, as these datasets often involve close-up imagery of identifiable individuals in sensitive contexts. Legal frameworks such as the GDPR (General Data Protection Regulation) and other regional privacy regulations restrict the collection, storage, and sharing of personally identifiable information (PII), making it difficult to compile and distribute open-access datasets [31]. Furthermore, cultural and religious sensitivities in specific regions can limit camera placement, the use of drones, or wearable sensors, thereby reducing the variety and richness of available data [4]. These constraints hinder the development of generalized models and limit benchmarking across research efforts. As a result, much of the progress in DNN-based crowd analysis relies on synthetic datasets, low-density footage, or highly controlled environments that fail to capture the complexity and unpredictability of real-world crowd behavior [14].

The rest of this section presents, in a non-exhaustive manner, some interesting datasets used in relatively efficient state-of-the-art methods.

5.1 UCSD Dataset

The UCSD Pedestrian Dataset comprises video sequences of pedestrian walkways, primarily used for anomaly detection and crowd counting tasks. It provides ground truth annotations for pedestrian counts in each frame, facilitating research in low-density crowd scenarios [25].

5.2 ShanghaiTech Dataset

Introduced by Zhang et al. [37], this dataset contains 1,198 annotated images divided into two parts: Part A with 482 images of dense crowds collected from the Internet, and Part B with 716 images of relatively sparse crowds captured on the streets of Shanghai. Each image includes head annotations, making it suitable for training and evaluating crowd counting models.

5.3 UCF-QNRF Dataset

Developed by Idrees et al. [18], the UCF-QNRF dataset consists of 1,535 high-resolution images with a wide range of crowd densities, from 49 to 12,865 individuals per image. The images were collected from various sources, including the Internet and Hajj footage, providing diverse scenes for robust crowd counting model evaluation.

5.4 JHU-CROWD++ Dataset

Presented by Sindagi et al. [27], JHU-CROWD++ is a large-scale unconstrained crowd counting dataset comprising 4,372 images with approximately 1.51 million annotations. The dataset includes images captured under diverse scenarios and environmental conditions, such as varying weather and illumination, making it a challenging benchmark for crowd counting algorithms.

5.5 Hajj-Specific Datasets

Several datasets have been developed to address the unique challenges of crowd analysis during the Hajj pilgrimage.

5.5.1 HAJJ-Crowd Dataset

Bhuiyan et al. [7] proposed a deep dilated convolutional neural network for crowd density image classification, supported by a dataset augmentation technique for the Hajj pilgrimage. This dataset includes 1,050 training images and 450 testing images, each with a resolution of 1280×720 pixels. It provides density maps and annotations for evaluating crowd counting methods in the context of the Hajj pilgrimage.

5.5.2 HAJJv2 Dataset

The HAJJv2 Dataset, collected by the KAU-Smart-Crowd group [20], comprises 18 videos recorded at critical locations during the Hajj pilgrimage—such as Massaa, Jamarat, Arafat, and Tawaf. It provides annotated data for abnormal behaviors and crowd density levels, making it valuable for research in behavior detection, density estimation, and crowd management. This dataset has been employed in recent studies such as Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds [5].

These datasets, like other state-of-the-art datasets designed specifically for the Hajj context, suffer from relatively modest sizes, making it impossible to effectively train a DNN. Moreover, they are most often not free to use and/or are not accompanied by clear licenses explicitly mentioning the right to privacy of individuals [7].

6 Limitations and Current Challenges

6.1 Occlusions and Visual Clutter

In densely packed environments such as religious pilgrimages, occlusion is one of the most significant challenges. Individuals are often partially or completely hidden behind others, which disrupts object detection and tracking continuity. State-of-the-art trackers like ByteTrack and OCSORT improve robustness but still degrade under extreme occlusion conditions [40]. Techniques such as occlusion-aware re-identification and depth-assisted tracking have been proposed but remain computationally intensive for real-time use [33].

6.2 Scale Variations and Perspective Distortion

Crowd scenes captured from elevated cameras often suffer from significant perspective distortion. Individuals near the camera appear much larger than those further away, resulting in scale inconsistency. Multi-scale architectures, such as the Multi-Column Convolutional Neural Network (MCNN) and the Multifaceted Attention Network (MAN), have been proposed to address these challenges. MCNN utilizes filters of varying sizes to capture features at different scales, while MAN incorporates global attention from transformers, learnable local attention, and instance attention to handle large-scale variations in crowd images effectively [24, 37].

Beyond architectural solutions, recent methods have explored preprocessing techniques to mitigate perspective distortion. For instance, Zhao et al. [39] introduced the Height Reverse Perspective Transformation (HRPT), which employs depth maps to rescale image regions, effectively narrowing the height gap among human heads in crowd images. This transformation enhances the visibility of individuals in distant areas and reduces redundancy in closer regions, thereby improving the performance of subsequent crowd counting models.

6.3 Environmental Conditions and Scene Dynamics

Real-world surveillance systems operate under fluctuating conditions—variations in illumination (e.g., day vs. night), weather (e.g., rain, fog), and dynamic backgrounds (e.g., moving shadows, flags)—which introduce substantial noise and variability. These factors can degrade the performance of traditional background subtraction techniques and reduce the generalization capability of convolutional neural networks (CNNs). To address these challenges, Zhang et al. [35] incorporated external factors such as weather conditions and holidays into their Deep Spatio-Temporal Residual Network (ST-ResNet) model, demonstrating improved accuracy in crowd flow prediction by accounting for environmental influences.

6.4 Data Scarcity and Generalization Gaps

Most public datasets are biased toward urban or Internet-collected scenes, limiting their applicability to culturally and behaviourally distinct contexts like the Hajj. There is a lack of high-quality, annotated datasets representing variations in dress, movement patterns, and spatial layouts specific to Mecca [7]. This scarcity limits both the training and benchmarking of models in such unique contexts. Synthetic datasets, digital twins, and semi-supervised learning are promising solutions, but their domain fidelity remains a concern [21].

6.5 Real-Time Constraints

Many advanced crowd counting models achieve high accuracy but are computationally intensive, limiting their deployment on edge devices with constrained resources. To address this, Chaudhuri et al. [10] proposed lightweight architectures utilizing MobileNet and MobileViT backbones, coupled with Adjacent Feature Fusion, to maintain competitive performance while ensuring computational efficiency. Their models demonstrate that it's feasible to balance accuracy and resource utilization, facilitating real-time crowd counting in practical applications.

7 Key features for an efficient approach

A truly effective approach to dense crowd analysis—particularly in highly complex environments such as the Hajj and Umrah pilgrimages—must be built upon the successful execution of three critical stages: data acquisition, pre-processing and model training, and neural network inference.

- The first stage, efficient data acquisition, involves capturing crowd scenes using various sensing modalities such as high-resolution RGB video, thermal cameras, depth sensors, or even aerial imagery from drones, depending on ethical and regulatory constraints [4, 31]. The choice and placement of sensors are vital to ensure comprehensive coverage and to minimize blind spots or data loss in highly dynamic environments [14].

- The second stage, pre-processing and training, requires a domain-specific dataset that reflects the visual and behavioral characteristics of dense crowds under realistic conditions. This includes images or sequences with significant occlusions, homogeneous clothing (e.g., white ihram garments), and crowded spatial layouts [7]. Pre-processing may involve data annotation, normalization, augmentation, and the careful tuning of neural network hyperparameters [37]. Crucially, the quality and quantity of data used in training deeply influence the model's generalization and robustness [18].

- Finally, the third stage, inference or neural processing, consists of deploying the trained deep learning model to perform real-time tasks such as crowd counting, individual identification, motion tracking, and trajectory estimation. This stage must be computationally efficient, accurate, and resilient to noise and occlusion, especially under the extreme densities characteristic of pilgrimage

settings [34, 40]. A typical structure of this stage using recent existing work might consist of three steps as follows:

- **Detection**: YOLOv8 fine-tuned on Hajj data to handle traditional attire and complex settings [19].
- **Tracking**: DeepSORT ensures identity continuity despite occlusions [34].
- **Prediction**: A Digital Twin simulates real-time and predictive crowd flow using data from detection/tracking [12]. This enables early intervention, congestion prediction, and safety optimization.

A failure or weakness in any of these three stages compromises the reliability of the overall system, highlighting the importance of a fully integrated pipeline where each component is optimized for dense, sensitive, and dynamic crowd environments.

8 Conclusion

Dense crowd analysis in the context of Hajj and Umrah presents unique challenges that exceed those existing in typical urban environments. This is mainly due to extreme density, visual similarity among individuals, and ritual-driven movement patterns.

This review highlighted the fragmented nature of existing research and emphasized the absence of a comprehensive state-of-the-art survey dedicated to pilgrimage scenarios. Progress in this field will require the development of larger, culturally and ethically appropriate datasets, along with models capable of robust generalization and real-time performance. Future advancements should also explore predictive and simulation-based systems to support proactive crowd management.

References

[1] Turki Abalkhail and S. Al Amri. Drone surveillance for large-scale religious events. `https://ieeexplore.ieee.org/document/abc123`, 2021.

[2] S. Al Amri and T. Abalkhail. Smart solutions for pilgrimage management using ai and iot. `https://hajjresearch.org.sa`, 2022.

[3] A. S. AlGhamdi, R. Mehmood, and R. AlShammari. Smart hajj: A gis and ai-based system for crowd management. *Sustainable Cities and Society*, 64:102561, 2021.

[4] Fawzi Alharbi et al. Privacy issues in surveillance systems: A review. *International Journal of Computer Applications*, 178(21):1–7, 2019.

[5] Fawzi Alharbi et al. Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive hajj crowds. *IEEE Transactions on Multimedia*, 2023. to appear.

[6] R. A. Bhuiyan, M. Al-Harthi, and J. Al-Muhtadi. Deep dilated convolutional neural network for crowd density image classification with dataset augmentation for hajj pilgrimage. *Sensors*, 22(14): 5102, 2022.

[7] R. A. Bhuiyan, M. Al-Harthi, and J. Al-Muhtadi. Deep dilated convolutional neural network for crowd density image classification with dataset augmentation for hajj pilgrimage. *Sensors*, 22(14): 5102, 2022.

[8] Thierry Bouwmans, Sajid Javed, Mahdy Sultana, and Soon Ki Jung. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 23:31–66, 2017.

[9] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking, 2023. URL `https://arxiv.org/abs/2203.14360`.

[10] S Chaudhuri et al. Lightweight and efficient architectures for real-time crowd counting. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. early access.

[11] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, PP:1–14, 01 2023. doi:10.1109/TMM.2023.3240881.

[12] Anthony Fuller, Zhen Fan, Chris Day, and Chris Barlow. Digital twins for real-time crowd monitoring and prediction. *IEEE Internet of Things Journal*, 7(5):4583–4590, 2020.

[13] Daniel Fuller, Rashid Mehmood, and Eisa Al Nuaimi. A data-driven agent-based simulation of the hajj crowd. *Simulation Modelling Practice and Theory*, 103:102096, 2020.

[14] Shuaicheng Gao et al. Deep learning for crowd analysis: A survey. *Neurocomputing*, 370:202–223, 2020.

[15] KAU Smart Crowd Research Group. Hajjv2 dataset. https://github.com/KAU-Smart-Crowd/HAJJv2_dataset, 2022.

[16] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3): 185–203, 1981.

[17] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. *CVPR*, pages 2547–2554, 2013.

[18] Hossam Idrees, Imran Saleemi, Christopher Seibert, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[19] Glenn Jocher et al. Yolov8: The latest yolo model from ultralytics. https://github.com/ultralytics/ultralytics, 2023.

[20] KAU-Smart-Crowd Group. Hajjv2 dataset. https://github.com/KAU-Smart-Crowd/HAJJv2_dataset, 2022. Accessed: 2025-06-16.

[21] Xuan Li et al. Digital twins for smart cities: Vision, enabling technologies, and future challenges. *IEEE Communications Magazine*, 59(1):10–16, 2021.

[22] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. doi:10.1109/CVPR.2018.00120.

[23] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018.

[24] Peng Lin et al. Multifaceted attention network for crowd counting. *IEEE Transactions on Image Processing*, 31:2311–2323, 2022.

[25] Vijay Mahadevan, Weixin Li, Vikas Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981, 2010.

[26] Mubarak Shah. Ai-based crowd monitoring for the hajj pilgrimage. https://www.crcv.ucf.edu, 2024.

[27] Vishwanath Sindagi and Vishal M Patel. Jhu-crowd++: A large-scale dataset for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1617–1631, 2020.

[28] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *arXiv preprint arXiv:2004.00036*, 2020.

[29] Andrés Sobral and Antoine Vacavant. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122:4–21, 2014.

[30] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–252. IEEE, 1999.

[31] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr): A practical guide. 2017.

[32] Jinxiao Wan, Zhenwei Shi, Zhihai Chen, and Hongkai Yu. A general crowd counting framework with attention guided composition. In *IEEE Transactions on Image Processing*, volume 30, pages 3940–3955. IEEE, 2021.

[33] Jian Wang et al. Occlusion-aware multi-object tracking using graph convolution network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2185–2197, 2021.

[34] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *ICIP*, 2017.

[35] Jie Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. *AAAI*, pages 1655–1661, 2017.

[36] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. URL `https://arxiv.org/abs/2110.06864`.

[37] Yingying Zhang, Desen Zhou, Si Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.

[38] Yingying Zhang, Desen Zhou, Siqin Chen, and Shenghua Gao. Single-image crowd counting via multi-column convolutional neural network. *CVPR*, pages 589–597, 2016.

[39] Lei Zhao et al. Height reverse perspective transformation for crowd counting using depth maps. *Computer Vision and Image Understanding*, 221:103637, 2023.

[40] Yifu Zhong et al. Bytetrack: Multi-object tracking by associating every detection box. *ArXiv preprint arXiv:2302.03702*, 2023.