Deep Vision Crowd Monitor: AI for Density Estimation and Overcrowding Detection.

SUBMITTED BY,

Avanthikalakshmi Vinod

TABLE OF CONTENT

ABSTRACT

The Deep Vision Crowd Monitor system is designed to accurately estimate crowd density and identify overcrowded environments using advanced deep learning techniques. Large gatherings in public spaces such as airports, stadiums, transport hubs, religious events, and festivals often present risks related to stampedes, congestion, and emergency mismanagement. Manual human monitoring is limited by reaction speed and visual fatigue, making automated digital surveillance essential.

This project employs an AI-based density estimation pipeline that converts raw visual crowd inputs into density maps using Gaussian-based head annotations. The ShanghaiTech dataset is utilized due to its wide diversity in crowd count, scene complexity, and real-world representation of both sparse and extremely dense populations. Density generation is accomplished by spreading each head annotation point into a weighted Gaussian area, forming continuous heat representations.

A CSRNet-based architecture, known for its dilated convolution backbone, is used to predict pixel-wise density intensities from input images. The model is trained using Mean Squared Error (MSE) loss to minimize the difference between predicted and actual density distributions. Through progressive optimization and fine-tuning, the network learns scale variance, multi-head density patterns, and occlusion challenges, all of which are major barriers in real-life crowd analysis.

The final system outputs both numerical estimation and visual congestion indication, enabling automated overcrowding alerts. This supports public safety management, assists in evacuation planning, and enhances real-time monitoring efficiency. Thus, Deep Vision Crowd Monitor demonstrates a practical AI-driven approach to smart urban surveillance and crowd risk prevention.

INTRODUCTION

Crowd monitoring has become an essential requirement in modern public environments due to increasing gatherings in locations such as airports, stadiums, metro stations, public festivals, and event venues. Overcrowding, if not tracked properly, can lead to safety hazards like panic spread, stampedes, and delayed evacuation responses.

Manual CCTV observation alone is often unreliable, as operators may miss critical congestion points due to fatigue and continuous visual load. To overcome this, AI-based automated surveillance systems provide real-time crowd assessment without human bias or delay.

The **Deep Vision Crowd Monitor** project uses deep learning–driven density estimation to identify how populated a scene is and where crowd build-up is occurring. Instead of detecting individual persons, the model generates density maps showing the intensity of human presence, making it highly effective even in dense, overlapping, or occluded gatherings.

Using the ShanghaiTech dataset and CSRNet architecture, the model learns spatial crowd patterns and predicts congestion zones accurately. This helps authorities detect overcrowded regions early and take preventive actions to maintain safety and smooth movement in public spaces.

DATASET USED

The dataset used in this project is the ShanghaiTech Crowd Counting Dataset, one of the most widely adopted benchmarks for crowd density estimation. It contains large variations in crowd distribution, scene perspective, scale, illumination, and viewing angles, making it suitable for real-world congestion monitoring research.

Dataset Components

- Part A
    - Consists of extremely dense public gatherings
    - Images collected from online sources such as events, rallies, parades, and large city gatherings
    - High crowd overlap and heavy occlusion
- Part B
    - Represents moderately crowded street-level images from Shanghai locations
    - Includes sidewalks, shopping streets, campus routes, and commercial pedestrian areas
    - Offers cleaner visibility with lower but realistic crowd spread

Annotations

- Each image is paired with a .mat annotation file
- These contain head coordinate points indicating the exact positions of individuals
- These points are later converted into Gaussian-based density maps to represent crowd intensity instead of individual detection

Purpose and Suitability

- The dataset supports density regression learning, which is more accurate than head detection in extremely dense conditions
- It simulates genuine crowd conditions found in:
    - Metro stations
    - Stadiums
    - Public events
    - Street gatherings
- It provides balanced diversity for both sparse-to-dense environments, ensuring the model learns scalable crowd estimation patterns

DATA PREPROCESSING

1. **Image Loading**

   All images loaded and converted RGB format.

2. **Normalization**

   ImageNet mean & variance applied.

3. **Density Map Generation**

   Gaussian kernels applied on each annotated head coordinate.

4. **Downsampling**

   Density maps reduced by factor ×8 and scaled to preserve total count.

5. **Tensor Conversion**

   Images & density maps converted to PyTorch tensors.

6. **Dataloader Setup**

   Structured batching for model training input.

MODEL TRAINING

Model used: CSRNet (dilated CNN for density regression)

Objective: Learn pixel-wise density distribution and estimate total crowd count from input images.

Training Configuration

- Loss Function: Mean Squared Error (MSE)
  Used to measure the difference between predicted density maps and actual ground-truth maps.
- Optimizer: Adam
  Selected for stable convergence and adaptive learning rate handling.
- Epochs: Trained over multiple cycles until training and validation loss plateaued.
- Input Format:
  - Preprocessed RGB crowd images
  - Gaussian-generated density maps corresponding to annotated head locations

Training Process Explanation

During training, the network is fed with batches of images and their density map targets. By comparing predicted density values with ground-truth maps, the loss function penalizes inaccurate estimations. Through continuous backpropagation, CSRNet learns:

- Scale variations (small heads vs large heads)
- Dense cluster representation
- Occlusion-heavy regions
- Perspective distortions in crowd scenes

The use of dilated convolutions enables the model to capture wide spatial context without losing resolution details, which is crucial in extremely dense gatherings.

Training Outcome

As epochs progress:

- Loss gradually decreases, indicating improved learning
- Predicted maps start to visually align with true density distributions
- Total crowd count estimation becomes more stable and accurate

This training phase forms the foundation of the system, enabling it to detect congestion hotspots and predict human accumulation levels effectively.

MODEL TESTING & EVALUATION

Testing was performed using unseen crowd images to measure how well the trained model generalizes beyond the training dataset.

Evaluation Metrics

- MAE (Mean Absolute Error):
  Indicates the average difference between predicted and actual crowd count. Lower MAE reflects better accuracy.
- RMSE (Root Mean Squared Error):
  Penalizes larger errors more strongly, especially useful for extremely dense scenes.

Testing Outputs

- Density Maps:
  High-intensity areas appear brighter, showing where crowd concentration is highest.
- Predicted Count vs Ground Truth:
  The estimated total count is compared against the annotated value to validate accuracy.

Overcrowding Detection

A fixed threshold is applied:

- If predicted density exceeds the set limit → Overcrowding Alert
- Helps quickly identify potential congestion zones in monitoring environments.

Outcome

The model provides reliable crowd estimation and clear visual highlighting of dense regions, making it suitable for real-time public safety monitoring.

FINE-TUNING

Fine-tuning is performed after the initial training to further improve the model's prediction quality and stability, especially in highly dense scenes.

Adjustments Made

- Reduced Learning Rate:
  Helps the model update weights more carefully, avoiding sudden fluctuations in prediction.
- Increased Training Epochs:
  Allows the network to refine feature understanding and improve density estimation precision.
- Adjusted Batch Size:
  Balances memory usage and learning consistency, improving training smoothness.

Impact of Fine-Tuning

- Enhances count accuracy in both sparse and extremely dense areas.
- Reduces prediction noise and unrealistic density spikes.
- Improves model generalization on new crowd images.
- Stabilizes loss curve and produces clearer, more reliable density outputs.

Fine-tuning ensures the system moves from a basic functioning model to a more polished and dependable crowd monitoring solution.

CONCLUSION

The proposed model provides an effective solution for automated overcrowding detection by converting regular crowd images into meaningful density heat maps. With the help of CSRNet and the ShanghaiTech dataset, the system successfully learns crowd distribution patterns and highlights areas where human accumulation is critically high.

By identifying congestion zones early, the system supports faster decision-making, reduces the risk of stampedes, and enables safer movement management in public spaces. The accuracy of density prediction and clear visual outputs make it suitable for deployment in modern surveillance setups, transportation hubs, and large-scale event monitoring.

Overall, the Deep Vision Crowd Monitor delivers a reliable, AI-driven method for crowd understanding and real-time congestion evaluation, strengthening safety mechanisms in smart city environments.