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

MahaLakshmi Tetala

Table of Contents

1. Project Overview

The DeepVision Crowd Monitor project focuses on creating a preprocessing pipeline that prepares crowd images for deep learning–based density estimation. Instead of directly counting people in an image, the system transforms visual inputs into numerical representations that a model can learn from.

The goal is not to train the model, but to prepare accurate inputs—images, ground-truth points, density maps, and tensors. These processed inputs will later be used by architectures like CSRNet or MCNN.

This preprocessing pipeline is crucial in real-world applications such as security monitoring, event management, and congestion analysis, where accurate density information is valuable.

2. Dataset Information

The dataset used is the **Shanghai Tech Crowd Counting Dataset** consisting of real images paired with annotations (.mat files).

Part A – Highly congested scenes

Part B – Medium-density outdoor images

Each .mat file contains pixel coordinates for each person's head with helps in counting and generating density maps

3. Environment Setup

The project uses Python in VS Code or Jupyter Notebook with a virtual environment. Main tools and libraries:

- **PyTorch** (torch, torchvision) – Used to build and train the deep learning model.

- **OpenCV** (opencv-python) and **Pillow** (pillow) – For loading, resizing, converting, and manipulating images.

- **NumPy and Pandas** – For efficient numerical operations and tabular data handling.

- **SciPy and Scikit-learn** – For mathematical processing such as loading .mat files and performing basic analysis.

- **Matplotlib and Plotly** – Used for visualizing images, density maps, and data insights during exploration.

- **TQDM** – Provides progress bars while processing large datasets.

- **Flask and Streamlit** – Used for API creation and building the user interface for displaying predictions.

- **Requests and Twilio** – Used for integrating external services such as notifications.

- **PyYAML** – For reading structured configuration files

GPU acceleration is optional but preprocessing is fast even on CPU.

4. Data Exploration

Before converting images into training-ready data, the dataset must be explored to ensure everything is present and correctly arranged.

Exploration steps performed:

1. **Folder Inspection**
   Verified the path structure, file counts, and that each image has a corresponding. mat annotation.

2. **Sample Image Viewing**
   A few images were displayed to confirm resolution, clarity, and the general spread of people.

3. **Annotation Inspection**
   Using the RobustMatReader, annotation coordinates were extracted and plotted over sample images.

This helps confirm that the points align perfectly with visible heads, ensuring the dataset is correctly annotated.

5. Data Preprocessing

The preprocessing workflow converts raw dataset inputs into model-ready formats. Steps include image loading, annotation extraction, resizing, density map generation, and tensor conversion.

5.1 Image Loading

Images are loaded using cv2.imread (BGR format), then:

- converted to RGB for consistency
- optionally transformed to floating-point values
- resized to a uniform working size

This ensures consistency and prepares images for normalization.

5.2 Annotation Extraction

Using RobustMatReader .mat annotation files are opened safely. Head coordinates are extracted from the 'image_info' structure and converted into Python lists/NumPy arrays. This ensures compatibility even if annotation formats differ.

5.3 Resizing and Downscaling

- To standardize data:
  Initial resize - The image is resized to a larger fixed resolution
  (e.g.- 1024×1024) to standardize shape
  Downscaling
  The image is then reduced to 512×512 or your preferred working size.

- This reduces memory usage and speeds up further processing.

5.4 Image Loading

Images are loaded using cv2.imread (BGR format), then:

- converted to RGB for consistency
- optionally transformed to floating-point values
- resized to a uniform working size

This ensures consistency and prepares images for normalization.

5.5 Annotation Extraction

Using RobustMatReader, .mat annotation files are opened safely. Head coordinates are extracted from the 'image_info' structure and converted into Python lists/NumPy arrays. This ensures compatibility even if annotation formats differ.

5.6 Resizing and Downscaling

To standardize data:

- **Initial resize** - The image is resized to a larger fixed resolution (e.g., 1024×1024) to standardize shape
- **Downscaling**
  The image is then reduced to 512×512 or your preferred working size.

This reduces memory usage and speeds up further processing.

5.7 Implementing Robust mat Reader

A RobustMatReader (SciPy-based safe loader) is used to ensure error-free extraction, even if annotation structures vary. These points serve as the foundation for generating Gaussian density maps.

5.8 Density Map Generation

Density maps convert sparse head points into smooth distributions.

Steps:

- Create a blank map filled with zeros.
- For each head coordinate, place a value of 1 at that location (using integer index rounding).
- Apply a Gaussian filter over the entire grid to convert sharp dots into smooth clusters.

Gaussian sigma defines the spread of each density blob. The sum of the density map approximates the crowd count.

5.9 Tensor Conversion & Normalization

- Images are converted to tensors in shape C × H × W.
- Normalized using ImageNet
- statistics: mean = [0.485, 0.456, 0.406]
- std = [0.229, 0.224, 0.225]
- Density maps become 1 × H × W tensors.

This ensures compatibility with PyTorch-based crowd counting models.

6. Training & Testing pipeline

Shanghai Tech Dataset – Part A & Part B

PartA

- Contains very dense crowds with many overlapping people.
- Harder for the model → lower accuracy, higher MAE/RMSE.
- More training is needed because the heads are small and tightly packed.

PartB

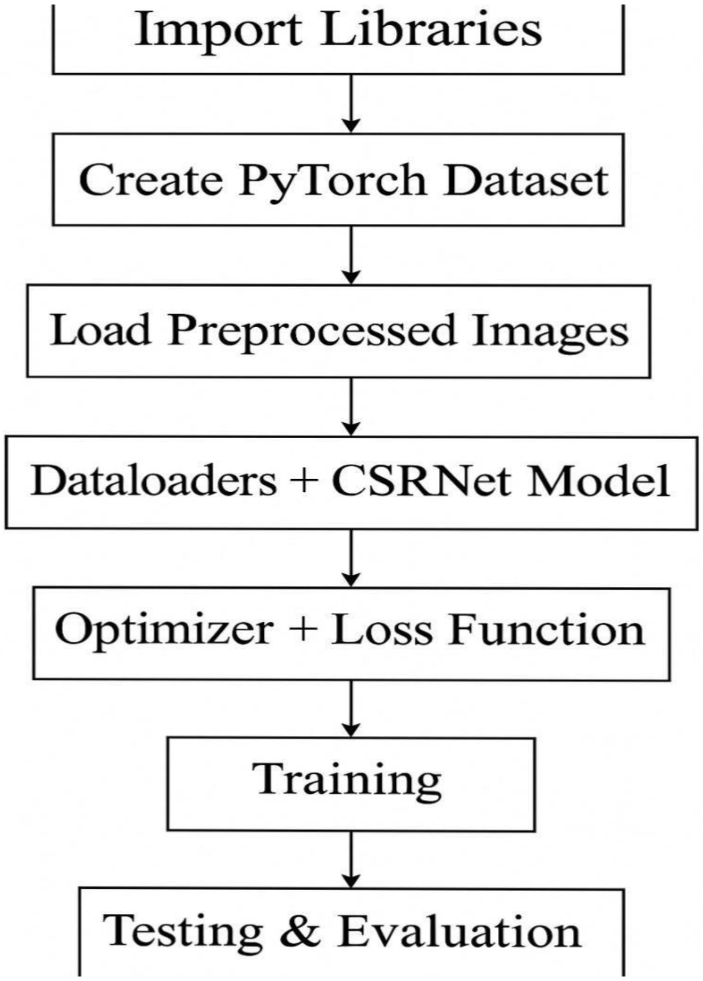
- Contains less dense, clearer outdoor scenes.
- Easier for the model → higher accuracy, lower MAE/RMSE.
- People are more visible and spaced out, so predictions are more reliable.

What I did during training

- Loaded preprocessed images and density maps.
- Created PyTorch Dataset & Dataloaders for batching and efficient training.
- Loaded the CSRNet model for density map prediction.

Why CSRNet is used

- CSRNet uses a VGG-16 frontend for strong feature extraction.
- Uses dilated convolution backend to capture large-scale crowd information.
- Designed specifically for crowd counting, making it well-suited for Part A & B.



Flow diagram

Additional training details

- Used MSE Loss and Adam Optimizer.
- Trained for multiple epochs until the loss stabilized.
- Saved the best checkpoint based on validation accuracy.

Testing Process

- Loaded test images from Part A and Part B.
- Generated predicted density maps and summed them to get counts.
- Calculated MAE and RMSE to measure performance.

Observed that:

1. Part A → lower accuracy (high crowd density).

Results —

- MAE: 78.89
- RMSE: 122.78
- Counting Accuracy: 80.47 %

2. Part B → higher accuracy (compared to Part A - less crowd density).

Results--

- MAE: 11.59
- RMSE: 19.57
- Counting Accuracy: 91.12%

7. Real-Time Crowd Monitoring Using Webcam

The objective of this task is to perform **real-time crowd monitoring using a webcam feed** by applying a trained CSRNet model for crowd density estimation. This helps in observing crowd behavior dynamically and estimating the number of people present in live scenes.

Processing Pipeline

1. **Webcam Input Capture**
   a. Live frames are captured using OpenCV.
   b. Each frame is processed independently.
2. **Frame Preprocessing**
   a. Frames are converted from BGR to RGB format.
   b. Resized to match the CSRNet input size.
   c. Normalized using ImageNet means and standard deviation.
3. **CSRNet Inference**
   a. Each frame is passed through the trained CSRNet model.
   b. The model outputs a density map representing crowd distribution.
4. **Crowd Count Estimation**
   a. The density map values are summed to estimate the total crowd count.
   b. The count is updated continuously for each frame.
5. **Output Display**
   a. Live video feed with crowd count displayed on the frame.
   b. Used for continuous monitoring.

Output

- Real-time webcam video, Live crowd count displayed on screen
- Smooth density-based estimation without explicit person detection

8. Crowd Density Estimation on a Single Video File

This task focuses on estimating crowd density and total crowd count from a **pre-recorded video file** using the CSRNet model. This helps evaluate model performance in offline scenarios.

Methodology

1. **Video Input**
    a. A single `.mp4` video file is loaded using OpenCV.
    b. Video is processed frame-by-frame.
2. **Frame Preprocessing**
    a. Each frame is resized and normalized.
    b. Converted into a PyTorch tensor before inference.
3. **Density Map Generation**
    a. CSRNet generates a density map for each frame.
    b. Density map highlights crowded and sparse regions.
4. **Crowd Count Calculation**
    a. Total crowd count is obtained by summing density map values.
5. **Visualization**
    a. Density heatmap is overlaid on original video frames.
    b. Crowd count is displayed on each frame.

Output

- Annotated output video
- Density heatmap visualization
- Frame-wise crowd count display

9. Batch Video Testing and Output Analysis

This task extends video-based crowd estimation to **multiple video files**, enabling batch processing and consistent analysis across datasets.

9.1 Batch Video Processing Strategy

- A folder containing multiple video files is used as an input.
- Each video is processed sequentially.
- Frames are extracted and passed through the CSRNet model.

9.2 Crowd Count Evaluation

- The crowd count is calculated for every frame.
- Helps compare crowd density variations across videos.

9.3 Folder-Based Video Testing

- Supports automated processing of multiple videos.
- Reduces manual effort for testing individual files.
- Output videos are saved with annotations.

9.4 Output Features

- Processed video files with:
    - Crowd count displayed
    - Density heatmap overlay
- Frame-wise analysis
- Suitable for large-scale monitoring scenarios

10. Conclusion

In this project, an end-to-end crowd density estimation and monitoring pipeline was developed using deep learning techniques. The work involved exploring and preprocessing the Shanghai Tech Crowd Counting Dataset by extracting head annotations, generating Gaussian-based density maps, and converting images and labels into normalized PyTorch tensors. The CSRNet model was used for crowd density estimation and evaluated on both dense (Part A) and medium-density (Part B) datasets, showing better performance on less crowded scenes. The trained model was further applied to real-time webcam input, single video files, and batch video testing, where density maps, crowd counts, and annotated outputs were generated for each frame. Overall, the project demonstrates a practical and effective approach for crowd analysis that can be used in real-world surveillance and crowd monitoring applications.