Design and Implementation of Intelligent Crowd Monitoring System

Akshitha Rai, Amrith M S, Ankur Sinha, Anuj Kumar Mishra, Manjunath R. Kounte
School of Electronics and Communication Engineering,
REVA University, Bengaluru-560064.
akshirai123@gmail.com, amrithms@gmail.com,
ankursinha99@outlook.com, anujkumar123123@gmail.com, manjunath.kounte@gmail.com

ABSTRACT:

One of the key objectives of crowd management is to plan and design a suitable strategy to equalize the crowd flow, prioritizing safety of the public. People tend to behave differently in a crowded scenario. For instance, if there is an altercation between angry fans or a fire in a building and in several other unforeseen circumstances panic and riot-like behaviour can break out. Therefore, an effective crowd management system is required to provide real time analysis of the crowd to law enforcement agencies for faster and efficient decision making. In order to achieve this, a network for Congested Scene Recognition known as CSRNet is introduced to deliver a data oriented and deep learning method that interprets densely-packed areas in order to obtain a precise crowd-count estimate. It generates high resolution and accurate density maps for a given input image. In this paper, CSRNet is implemented on the Shanghai-Tech data set and the MAE (Mean Absolute Error) obtained is significantly about 25% lower than some of the traditional approaches used previously.

Keywords – Crowd management; Congested Scene Recognition; MAE (Mean Absolute Error).

1. INTRODUCTION:

As the global population is growing at an exponential rate, the need for monitoring and managing a crowd is pivotal for public safety. Prior estimation of crowd-density of a particular area can assist in formulating guidelines for designing public spaces. If the estimated density value is above the threshold value for a particular area, then a suitable exit strategy can be deployed to prevent stampedes and other crowd related incidents. Hence, determining crowd-density is imperative for crowd analytics. Crowd-dispersion maps can be contrasting for the same number of individuals, therefore estimating the crowd count value is not sufficient. The density map provides accurate and comprehensive information which is utilized for making appropriate decisions to help people in unsafe environments.

Generating accurate distributive maps is a challenging task. Density values are generated for every pixel in the image, spatial coherence must be incorporated to show a clear distinction between similar looking pixels. If the crowd is distributed unevenly or the camera is at awkward angles, it is difficult to use traditional approaches without DNNs (Deep Neural Network)

Website: http://www.modern-journals.com/
An Intelligent Crowd Monitoring System operates as a continuous looping algorithm as shown in the block diagram given above. The first stage involves ‘planning’ based on historical records of crowd densities of a specific location, simulation results and based on different scenarios such as a parade, music concert, marathon or a protest march. Crowd behaviour will vary in each scenario. The next stage involves ‘decision-making’. Quick decisions are made considering the crowd-behaviour-prediction and assessment of the present scenario. Appropriate actions are taken to guide the law enforcement or event-management authorities to manage the crowd [10]. Continuous feedback is given to help them change strategies as the crowd density changes with time. Cameras or sensors installed at the points of interest are used to detect the crowd density, weather and location conditions etc. It acts as the input to the designed crowd control system.

2. LITERATURE REVIEW:

2.1 Foreground-based methods:

The paper [1] proposes a model which operates as four modules. First the region of interest is selected, from which a density map is generated to account for the camera’s perspective distortion. Then the foreground pixels are identified in the present frame to detect objects in motion. Blob-based feature descriptors are used to detect static objects. Finally, a collective analysis of the scene is carried out. However, the model gives high congestion values even if there is a medium-sized crowd in underground train platforms.

2.2 Feature-based methods:

Feature-based head detection [2, 7] and integral channel feature-based head detection [3] detect and estimate the
2.3 Edge-detection based methods:

The gravitational edge detection method [4] first separates foreground pixels from the background noise. The gravitational approach is used to detect edges of the foreground and based on the length of edge and number of grids within the edge, density of the crowd is obtained. The method works fairly well for large crowds however, if the crowd is sparse, it wrongly identifies the edges giving inaccurate count values.

2.4 Regression-based methods:

Regression-based solutions [5, 8, 11] are best suited for highly congested scenes. The video feed is broken down to frames and it is segmented based on the motion of people walking on the pedestrian crossing. Feature extraction is performed on these segments and finally ‘Bayesian Poisson Regression’ is performed on each segment to find the number of people.

2.5 CNN-based methods:

The proposed system [6] designs a congested scene recognition network known as CSRNet which performs accurate crowd count analysis. Unlike the traditional regression approach which takes segments of the image, CNN-methods [9] generate density-maps from a whole image. The estimated count value is then obtained from the density maps.

3. DESIGN AND IMPLEMENTATION:

3.1 CSR-NET ARCHITECTURE:

VGG-16 is one of the most sought-after models used for image-classification. After some fine-tuning, it is used as the front-end layer in the CSRNet architecture. The size of the output from VGG is $\frac{1}{4}^{th}$ of the original input size. A 3 x 3 convolutional filter size is used to reduce the complexity of the single-column neural network. The first 10 layers of the architecture are incorporated from the VGG-16 model to act as the front-end and dilated convolution layers are deployed as the back-end to draw out profound features from an image retaining its original resolution and to maximize the receptive fields. The last layer is 1 x 1 convolutional layer producing the density map.

VGG, expanded as Visual Geometry Group- was founded and developed by a group of Oxford University researchers. It was trained on a data-set of over fifteen-million images. Each block in the VGG-16 model consists of:

- 2D Convolutional Layer
- Max-Pooling Layer
### Configuration of CSR-Net

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>input (unfixed-resolution color image)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>front-end (fine-tuned from VGG-16)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv3-64-1</td>
<td>conv3-64-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>max-pooling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv3-128-1</td>
<td>conv3-128-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>max-pooling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv3-256-1</td>
<td>conv3-256-1</td>
<td>conv3-256-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>max-pooling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv3-512-1</td>
<td>conv3-512-1</td>
<td>conv3-512-1</td>
<td>conv3-512-1</td>
</tr>
<tr>
<td></td>
<td>back-end (four different configurations)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>conv1-512-1</td>
<td>conv3-512-2</td>
<td>conv3-512-2</td>
<td>conv3-512-4</td>
</tr>
<tr>
<td></td>
<td>conv3-512-1</td>
<td>conv3-512-2</td>
<td>conv3-512-2</td>
<td>conv3-512-4</td>
</tr>
<tr>
<td></td>
<td>conv3-512-1</td>
<td>conv3-512-2</td>
<td>conv3-512-2</td>
<td>conv3-512-4</td>
</tr>
<tr>
<td></td>
<td>conv1-256-1</td>
<td>conv3-256-2</td>
<td>conv3-256-4</td>
<td>conv3-256-4</td>
</tr>
<tr>
<td></td>
<td>conv1-128-1</td>
<td>conv3-128-2</td>
<td>conv3-128-4</td>
<td>conv3-128-4</td>
</tr>
<tr>
<td></td>
<td>conv1-64-1</td>
<td>conv3-64-2</td>
<td>conv3-64-4</td>
<td>conv3-64-4</td>
</tr>
<tr>
<td></td>
<td>conv1-1=1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 2 Single column architecture of CSRNet**

The convolutional parameters are denoted as “conv-(kernel-size)-(number of filters)-(dilation rate)”. Max-pooling layers use 2 x 2-pixel image with strides of 2. In max-pooling, a group of usually 4 pixels is mapped to a single pixel with the highest pixel value among the four.

Dilated convolution layers at the backend is used to increase the field of the kernel, without expanding the parameters. Pooling layers can be replaced by this layer.

The kernel is convolved over the whole image when the dilation rate=1. When the dilation rate=2, the kernel field expands as depicted in the image below:
CSRNet is a single-column architecture crafted from two significant elements:

- Convolutional Neural Network dedicated for 2D feature extraction at the front-end of the architecture.
- Dilated Convolutional Neural Network dedicated for increasing the receptive field at the back-end of the architecture.

The CSRNet is trained on the Shanghai dataset and the ground truth values are obtained for each image in the dataset by making use of geometry adaptive kernels. This method is well-suited for dense crowds. Mean Absolute Error and Root Mean Square Error abbreviated as MAE and RMSE respectively, are the evaluation metrics used. They can be represented as follows:
Where $N$ represents the number of test images, $C_{I_i}^{\text{pred}}$ and $C_{I_i}^{\text{gt}}$ represent the prediction count results and the ground truth values respectively. Precision of the obtained crowd-count estimates is represented by the MAE value. Its validity and reliability is determined by the RMSE value. The predicted count value given by:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| C_{I_i}^{\text{pred}} - C_{I_i}^{\text{gt}} \right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left| C_{I_i}^{\text{pred}} - C_{I_i}^{\text{gt}} \right|^2},$$

$L$ and $W$ represent the predicted density map width.

The model is then tested on real-time captured images of a crowd. The corresponding density map and predicted count values are obtained. During generation of density map, a pre-defined value is used if a pixel maps to a person in the input image. But if it does not map to a person, then the pixel value used is zero.

The model implements CSRNet on the Shanghai Tech dataset. 1198 annotated images of a combined total of 330,165 people is present in the dataset. An MAE value of 75.69 is obtained for the designed system. The model is loaded onto the NVIDIA Jetson Nano Developer Kit for processing.
4. RESULTS AND DISCUSSION:

- The output displayed consists of the density map and the input image. These images are the frames obtained from a sample 15 second video of a moving crowd. The density map shows areas over which the...
crowd concentration is distinctively high. 

- In Fig.5 it is observed that the density of the crowd is the highest in comparison to the other results. Fig.6 has moderate crowd flow and in Fig.7 the crowd is scattered and the density is less.
- The Shanghai Tech data set used to train the model contains Part-A and Part-B, which includes ‘train’ and ‘test’ images. Part-A contains 300 train images and 182 test images of average resolution 589x868. While part-B contains 400 train images and 316 test images of average resolution 768X1024.
- If the resolution of the input image is high, the quality of the density map increases because it is generated for every pixel in the image. However, if the image resolution is too large, memory required to store the image is high and a strong computation to build the model on large-sized images is needed.

5. CONCLUSION AND FUTUREWORKS:

Various crowd counting mechanisms have been developed, each having its fair share of pros and cons. CSRNet model has a strong feature learning ability because it uses a network-prior and being a single column architecture, it reduces the computational cost. Quality of the density map is emphasized using two measures: peak signal to noise ratio (PSNR) and structure index similarity (SSIM) [18]. CSRNet has a PSNR=23.79 and SSIM=0.76, higher the value, higher the quality of the density map [19]. The model has achieved high PSNR and SSIM values in comparison to other models.

However, being a lightweight model, there is a slight dip in accuracy which is a challenge to accomplish in the future. A robust model will need to cope up with diverse scenery, lighting variation of a certain location and weather changes.

6. REFERENCES:


Website: http://www.modern-journals.com/


