SINGLE-IMAGE CROWD COUNTING VIA MULTI-COLUMN CONVOLUTIONAL NEURAL NETWORK

Literature Review

Detection-based Framework: Early approaches to crowd counting relied on a detection-style framework, which involves deploying detectors over consecutive frames of video sequences to estimate the number of pedestrians or individuals present. These methods primarily relied on boosting appearance and motion features to enhance detection accuracy. However, they faced significant challenges, particularly in densely populated or clustered environments, where occlusion among individuals could severely affect the performance of the detector and consequently the accuracy of crowd count estimations. [1], [2], and [3] used a detection-based Framework.

Trajectory Clustering: Another set of methodologies focused on clustering trajectories of tracked visual features to estimate crowd counts in videos. This approach, used by [4] and [5], involved tracking the movement of individuals using different methods and grouping them into clusters that represent independently moving entities within the crowd. While this method showed promise in video-based crowd counting scenarios, it was found to be impractical for estimating crowd counts from individual still images, limiting its applicability in certain contexts.

Feature-based Regression: Feature-based regression emerged as one of the most widely used methods for crowd counting,used by [6], [7] and [8]. This approach typically involves several steps, starting with foreground segmentation to isolate individuals or the crowd from the background. Following segmentation, various features are extracted from the foreground region, such as the area of the crowd mask, edge count, or texture features, which capture important visual characteristics of the crowd. These features are then utilized as inputs to a regression function, which predicts the crowd count based on the extracted features. Regression models employed in this approach range from simple linear or piecewise linear functions to more sophisticated techniques such as ridge regression, Gaussian process regression, and neural networks, each offering varying degrees of accuracy and complexity in crowd count estimation.

Still Image Crowd Counting: In addition to video-based crowd counting, there has been a growing interest in estimating crowd counts from individual still images ([9]). While many of the methodologies and techniques used in video-based crowd counting can be adapted for still image analysis, specific approaches have been developed to address the unique challenges posed by static images. These approaches often leverage different algorithms and methodologies tailored to the characteristics of still images, allowing for accurate crowd count estimates even in the absence of temporal information provided by video sequences.

Mathematical Foundations

1.Density map via Gaussian kernel

To estimate crowd density from an image, we start with labeled head positions represented as delta functions. Hence an image with N heads labeled can be represented as a function -

$$H(\boldsymbol{x}) = \sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i).$$

Converting this to continuous density involves convolving these deltas with a Gaussian kernel.

More precisely, the formula for density becomes a sum of delta functions convolved with these geometry-adaptive kernels for some parameter β. In other words, we convolve the labels H with a density kernel.

$$F(\boldsymbol{x}) = \sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i) * G_{\sigma_i}(\boldsymbol{x}), \quad \text{with} \quad \sigma_i = \beta \bar{d}^i$$

2. Multi-column CNN for density map estimation

The proposed Multi-column Convolutional Neural Network (MCNN) addresses the challenge of capturing crowd density at various scales due to perspective distortion. The MCNN consists of three parallel CNNs, each with filters of different sizes to model density maps corresponding to heads of different scales. Larger receptive fields are used to capture characteristics of larger heads.

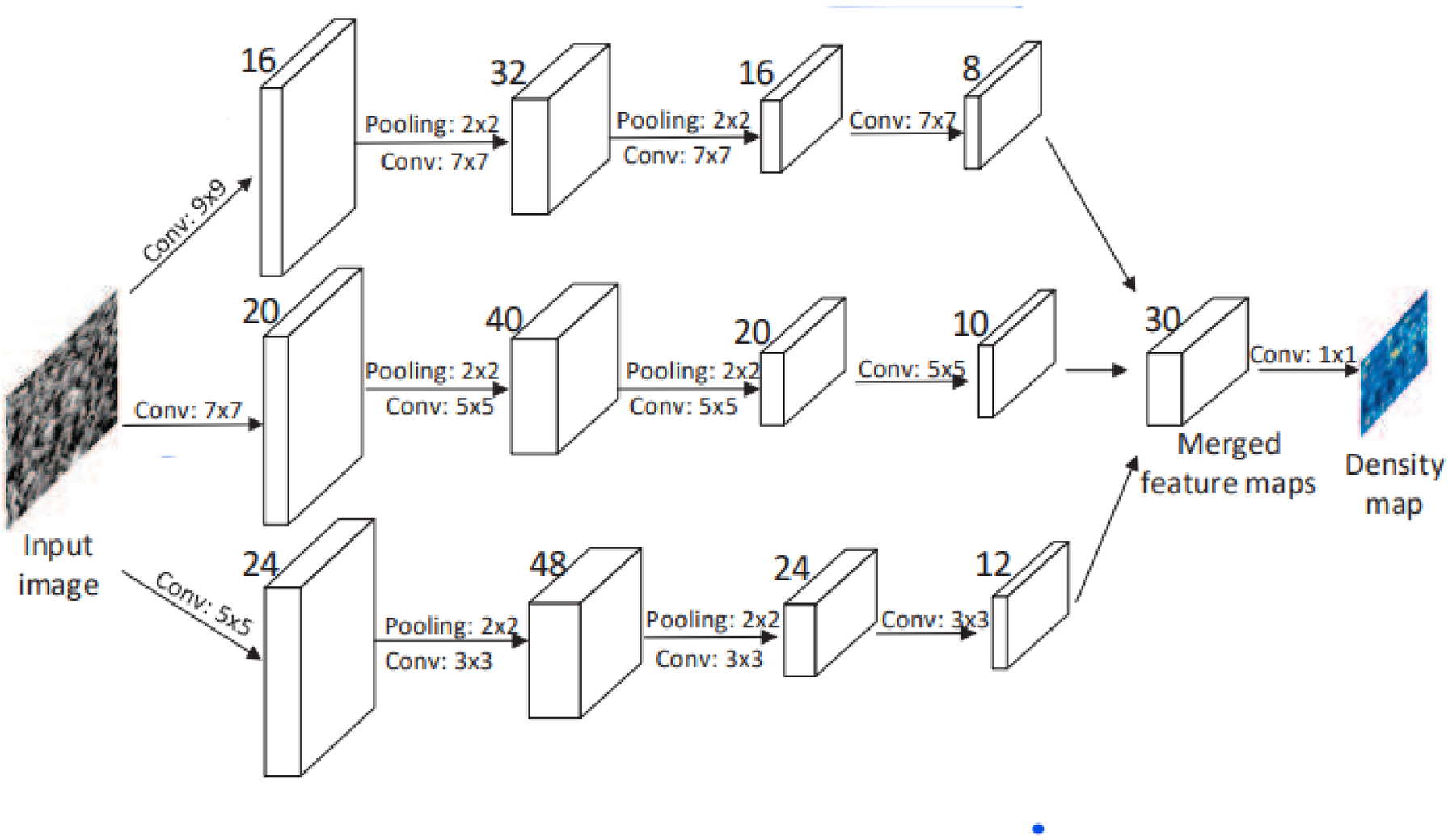
The structure of the MCNN involves parallel CNNs with similar architectures (conv-pooling-conv-pooling), but with varying sizes and numbers of filters to handle different scales. Max pooling is applied for each 2×2 region, and ReLU activation functions are utilized for their effectiveness in CNNs.

To manage computational complexity, fewer filters are used for CNNs with larger filters. The output feature maps from all CNNs are stacked and mapped to a density map using 1×1 filters. The Euclidean distance measures the difference between the estimated density map and the ground truth, serving as the loss function during training which is as follows

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|F(X_i; \Theta) - F_i\|_2^2,$$

Methodology

Following is the architecture of the Multi-column CNN -



There are three branches, each with different kernel sizes. Also since there are two layers of max pooling, the resolution of the images becomes one-forth, which is taken care of by downscaling the images.

Since there are so many parameters, instead of training the model directly, first the branches are trained seperately on the data. The final MCNN is then initialized with the parameters of these pre-trained branches.

The optimizer used is Adam and the loss function is MSE.

Results

MAE - 200.60246065687664
MSE - 349.7032351139194
RELATIVE ERROR - 31%

References

[1] - . Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision, 63(2):153-161, 2005.

[2] - Z. Lin and L. S. Davis. Shape-based human detection and segmentation via hierarchical part-template matching. Pattern Analysis and Machine Intelligence, 32(4):604-618, 2010.

[3] - M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In CVPR, pages 3401-3408. IEEE, 2011.

[4] - V. Rabaud and S. Belongie. Counting crowded moving objects. In CVPR, volume 1, pages 705-711. IEEE, 2006.

[5] - G. J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In CVPR, volume 1, pages 594-601. IEEE, 2006.

[6] - . B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In CVPR, pages 1-7. IEEE, 2008.

[7] - A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In ICCV, pages 545-551. IEEE, 2009.

[8] - K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In BMVC, volume 1, page 3, 2012.

[9] - C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In CVPR, 2015.