Estimating Number of People in Crowds

A Review of Prior Research

Ankan Bansal
Advisor: Prof. K S Venkatesh
Motivation

Analysis of crowded scenes has a large number of applications -

- Crowd Management
- Mathematical Models of Crowds
- Simulations
- Surveillance
- Public Space Design
Motivation

- Most of the existing work focuses on non-crowded situations.
- Crowds are complex.
- Difficult to differentiate individuals.
- A crowd acts as a single entity.
- Modeling and synthesis are difficult problems.
Interesting problems

Crowds

Analysis

Synthesis

People Counting

People Tracking

Behaviour Understanding
People Counting/Density Estimation

- Density can be used as a measure of comfort level in public spaces.
- Detect potentially dangerous situations e.g. fighting, rioting, protest, mass panic etc.
- Public space design.
- Crowd control.
- Manual counting of number of people in large crowds is difficult.
The models used can be broadly divided into three categories:

- **Pixel-level analysis**
  - Local features e.g. individual pixel analysis, edge detection after background subtraction, edge orientations etc.
  - Density estimation rather than counting.

- **Texture-based analysis**
  - Analysis of image patches.
  - Estimate number of people rather than identify individuals.

- **Object-level analysis**
  - Identify individual objects in the scene.
  - Mostly feasible in low density crowds.
1. Privacy Preserving Crowd Monitoring

- System for estimating the size of non-homogeneous crowds that does not depend on object detection or feature tracking

- Can be implemented with hardware that does not produce a visual record of the people in the scene
Crowd counting system [1]
Steps

- Crowd segmentation - mixture of dynamic textures
- Perspective normalization of features
- Feature extraction - some features vary linearly with number of people
  - Segment features
    - Area, Perimeter, Perimeter edge orientation, Perimeter-area ratio
  - Internal edge features - Canny edge detector
    - Total edge pixels, Edge orientation, Minkowski dimension
  - Texture features
    - Homogeneity, Energy, Entropy
- Gaussian process regression
2. Analysis of Crowded Scenes using Holistic Properties

- Introduced a new database (PETS 2009) for testing the counting algorithm proposed in [1]
- Showed results on the same
3. A Viewpoint Invariant Approach for Crowd Counting

- Feature normalization to deal with perspective and camera orientations
- Background subtraction
- Edge orientation and blob size histograms as features
- A homography is computed between the ground plane and the image plane coordinates for the ROI
- Linear fitting and neural networks
Steps

● Feature extraction
  ○ Foreground regions - blob size histogram and edge orientation histogram
  ○ Mixture-of-Gaussian based background modeling method to generate foreground mask

● Density estimation - density map associated with the ROI and a scale factor
  ○ Used to normalize the features with respect to object translation and different viewpoints

● Feature normalization

● Linear fitting or neural network
Results for camera orientation 30 degree [7]
Results for camera orientation 70 degree [7]
4. Crowd Counting using Multiple Local Features

- Local features to count the number of people in each foreground blob segment
- Total is the sum of group estimates
- Holistic systems require large training set
- Localised approach to ground truth annotation which reduces the required training data
Steps

- Foreground segmentation into blobs
- Perspective normalization
- Feature extraction from each blob segment
  - Area, Perimeter, Perimeter-Area ratio, Edges, Edge Angle Histogram
- Crowd Counting
  - Using neural networks as classifiers
  - Total crowd estimate is the sum of estimates for all blobs
Accuracy testing results [4]
Downscaling and upscaling testing results using linear classifier [4]
5. Feature Mining for Localised Crowd Counting

- Single regression model
- Estimate people count in spatially localised regions
- Localised feature importance mining and information sharing among regions
- Scalable
Multi-point regression framework for localised crowd counting by feature mining [6]
Steps

● Perspective normalisation
● Extraction of low-level features from each cell region
  Local foreground, edges and texture features
● Local intermediate feature vectors \rightarrow a single ordered
  (location-aware) feature vector
● Multi-output regression model based on multivariate
  ridge regression is trained
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td>maе</td>
<td>mse</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mae</td>
<td>mse</td>
</tr>
<tr>
<td>RR [22]</td>
<td>✓</td>
<td>–</td>
<td>2.25</td>
<td>7.82</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPR [6]</td>
<td>✓</td>
<td>–</td>
<td>2.24</td>
<td>7.97</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR [23]</td>
<td>–</td>
<td>✓</td>
<td>2.60</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MORR</td>
<td>–</td>
<td>✓</td>
<td>2.29</td>
<td>8.08</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance comparison between different methods [6]
6. Density-aware Person Detection and Tracking in Crowds

- Information of global structure of the scene is used and all detections are resolved jointly
- Person detection is formulated as the optimization of a joint energy function combining crowd density estimation and localization of individual people
Person detection performance (Average precision) [12]
Tracking RMS error [12]
7. Estimating Pedestrian Counts in Groups

- Counting and tracking
- Use of prior knowledge from scene and accurate camera calibration
- Groups are tracked in the same manner as individuals using Kalman filtering techniques
- A group of people is treated as a single entity
Average estimated counts of both methods based on the modal estimate for groups of various sizes

<table>
<thead>
<tr>
<th>Group size</th>
<th>No. of groups</th>
<th>Heuristic count</th>
<th>Shape count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>46</td>
<td>2.46</td>
<td>2.60</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>2.80</td>
<td>3.82</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>4.45</td>
<td>4.23</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>5.34</td>
<td>5.36</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>6.55</td>
<td>6.49</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>8.20</td>
<td>7.81</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>9.04</td>
<td>8.85</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>9.7</td>
<td>9.48</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>8.90</td>
<td>10.43</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>12.5</td>
<td>11.63</td>
</tr>
</tbody>
</table>
8. Crowd Counting by Integer Programming with Local Features

- Integer programming method for estimating the instantaneous count of pedestrians crossing a line of interest (LOI) in a video sequence
- Video is converted into a temporal slice image through a line sampling process
- Local features (HOG) are used to estimate the count in the temporal slice
a) temporal ROI counts over time and instantaneous count estimates using b) integer programming, c) least-squares, and d) non-negative least squares [14]
9. Learning to Count Objects in Images

- Supervised learning approach
- Dotted annotations
- Find the density function as a function of the pixels
- The integration of this density function over the image gives the total count
- Specific distance metric between density functions called MESA (Maximum Excess over SubArrays) distance
Framework

- Each pixel in each image is associated with a real-valued feature vector
- Each training image is annotated with a set of 2D points
- The ground truth density for a dotted annotation is defined as a sum of normalized gaussians centered at the dots such that the sum of the ground truth density over the entire image gives, approximately, the count of object in the image
- A linear transformation of the feature representation that approximates the density function at each pixel is to be learned
- This is done by minimizing the sum of mismatches between the ground truth and the estimated density functions (the loss function) under regularization
- The loss function is the MESA distance
<table>
<thead>
<tr>
<th>Method</th>
<th>'maximal'</th>
<th>'downscale'</th>
<th>'upscale'</th>
<th>'minimal'</th>
<th>'dense'</th>
<th>'sparse'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting-by-Regression [17]</td>
<td>2.07</td>
<td>2.66</td>
<td>2.78</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Counting-by-Regression [28]</td>
<td>1.80</td>
<td>2.34</td>
<td>2.52</td>
<td>4.46</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Counting-by-Segmentation [28]</td>
<td><strong>1.53</strong></td>
<td>1.64</td>
<td>1.84</td>
<td><strong>1.31</strong></td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Density learning</td>
<td>1.70</td>
<td><strong>1.28</strong></td>
<td><strong>1.59</strong></td>
<td>2.02</td>
<td>1.78±0.39</td>
<td>2.06±0.59</td>
</tr>
</tbody>
</table>

Mean absolute errors for people counting in surveillance videos [11].
Here [28] -> Crowd Counting Using Multiple Local Features [4]
1. People Counting in a Challenging Situation

- Estimating the number of people and locate each individual in a low resolution image
- Postprocessing steps on background subtracted results to estimate the number of people
- EM-based method to locate individuals
- Number of people is used as a priori for locating individuals based on feature points
Block diagram of the method [9]
Steps

- Background estimation based on a GMM
- Obtain foreground image and binarize to obtain the foreground pixels
- Perspective correction
  - Objects at different locations are brought to the same scale
- Computer number of foreground pixels ($X$)
- Closing operation on foreground pixels ($C$)
- Neural network with foreground pixels and closed foreground pixels as inputs to find person count ($M$)
- Individual detection
Learned relationship between X and M

(a)

Learned relationship between C and M

(b)

People counting results [9]
<table>
<thead>
<tr>
<th>Inputs</th>
<th>Mean error percentage (for the 51 test cases)</th>
<th>Accuracy (% of cases with error percentage less than 10%)</th>
<th>Accuracy (% of cases with error percentage less than 15%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$ vs. $M$</td>
<td>16.36</td>
<td>45.10</td>
<td>60.78</td>
</tr>
<tr>
<td>$C$ vs. $M$</td>
<td>10.68</td>
<td>60.78</td>
<td>80.39</td>
</tr>
<tr>
<td>$X$, $C$ vs. $M$</td>
<td>10.03</td>
<td>68.04</td>
<td>80.39</td>
</tr>
</tbody>
</table>

Results [9]
2. Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images

- Multiple sources - low confidence head detections, repetition of texture elements (using SIFT), frequency-domain analysis - along with the associated confidence
- Global consistency constraints on counts using Markov Random Field - caters for disparity in counts in local neighborhoods and across scales
- Still images
- 94 - 4543 individuals in images in the data set
Framework

- Density of people varies from region to region
- Crowded image cannot be analysed in its entirety
- Counting in small patches uniformly sampled over the image
- But the density varies smoothly and adjacent patches should have similar density
- When counting the density is assumed uniform but it is assumed that the number of people in each patch is independent of adjacent patches
- After estimating the counts/density in each patch, the independence assumption is removed and the patches are placed in multi-scale MRF to model the dependence in counts among nearby patches
Counting in Patches

Given a patch $P$, three different sources are used for estimating the count alongside confidences for those counts:

- **HOG based Head Detections**
  - Deformable Parts Model trained on INRIA Person dataset
  - Only filter corresponding to head applied
  - Detections are accompanied with scale and confidence
  - For each patch, the number of detection $n_H$, mean and variance of scale $m_{H,s}$, $s_{H,s}$ and confidence $m_{H,c}$, $s_{H,c}$. 
Fourier Analysis

- A crowd is inherently repetitive in nature.
- If the crowd density in a patch is uniform, the periodic occurrence of heads shows as peaks in the frequency domain.
- First a low pass filter is applied to remove the high frequency components.
- Then the low amplitude frequencies are discarded.
- This is followed by reconstructing the image through inverse Fourier Transform.
- The number of local maxima in the reconstructed image after alignment, non-maximal suppression and normalization for size of the patch gives the estimate for FT based count, $n_F$.
- Several other measures such as entropy, mean, variance, skewness and kurtosis for both the reconstructed image and the difference image are also calculated.
• Interest Points based Counting
  ○ To estimate counts and to get a confidence whether the patch represents crowd or not
  ○ Support Vector Regression is used on SIFT features to obtain the counts
  ○ These features can also be used to learn whether a crowd is present in the image or not

Fusion of Three Sources
• Overlapping patches are densely sampled from the training images and counts are obtained for the corresponding patches from annotations
• Counts and confidences are computed from the three sources
• Individual features are scaled and epsilon-SVR is used for regression
Counting in Images

- To impose smoothness among counts from different patches, they are placed in a multi-scale MRF framework with grid structure.

Multi-scale MRF for inferring counts for entire images [10]
Results on selected images [10]
Per patch estimates [10]. Means - black, SD - red, GT - olive. AD - Mean Absolute Difference, NAD - Normalized Absolute Difference (obtained by normalizing the AD with the actual count for each image)
<table>
<thead>
<tr>
<th>Method</th>
<th>Per Patch</th>
<th>Per Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD</td>
<td>NAD</td>
</tr>
<tr>
<td>Fourier</td>
<td>13.8 ± 21.3</td>
<td>96.4 ± 200.4</td>
</tr>
<tr>
<td>F+confidence</td>
<td>11.0 ± 19.7</td>
<td>58.7 ± 74.9</td>
</tr>
<tr>
<td>Fc+Head</td>
<td>11.1 ± 19.3</td>
<td>63.3 ± 84.0</td>
</tr>
<tr>
<td>FHc+SIFT</td>
<td>10.2 ± 18.9</td>
<td>53.3 ± 69.5</td>
</tr>
<tr>
<td>FHSc+MRF (Proposed)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rodriguez et al.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lempitsky et al.</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Quantitative Results [10]


