3D Crowd Counting via Multi-View Fusion with 3D Gaussian Kernels
Qi ZHANG, Antoni B. CHAN (City University of Hong Kong)

Introduction

Problems with single-view counting:
- Limited field-of-view of single camera;
- Low resolution in the farther place;
- Severe occlusion, e.g., humans, buildings.

Solution: Multi-view counting
- Previous work fuse multiples views to estimate ground-plane 2D density maps.
- We propose counting with 3D density maps.

3D Crowd Counting

- Assume fixed camera, calibrated and synchronized;
- Pipeline: 3D projection, 3D fusion and 3D density map prediction;
- Losses: Camera prediction, 3D prediction, and 3D-2D projection consistency measure.

3D Crowd Counting Pipeline

1) 3D projection (multi-height projection)
- The multi-height projection puts feature of the person’s body along z-dimension.
- Body features align to form a 3D representation for a person.

2) 3D-2D projection consistency loss PCM:
- Encourages consistency between 3D prediction and 2D ground-truth.
- Example: There are 4 people in the 3D prediction, while only Persons 2 and 3 are visible in the 2D view i. Since Person 1 is occluded by Person 2, and Person 4 is totally occluded in view i, they are masked out in the PCM calculation.

Loss Function

\[ l_{all} = l_{3d} + \beta l_{2d} + \gamma l_{3d,2d}, \]

where \( \beta = 0.01 \) and \( \gamma \) are hyperparameters for weighting the contributions of each term.

Datasets and Results

Datasets: PETS2009, DukeMTMC and CityStreet

Visualization from different view angles: CityStreet

Experiments and Ablation Study

1) Counting performance (mean absolute error; lower is better):

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PETS2009</th>
<th>DukeMTMC</th>
<th>CityStreet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dmap weighted</td>
<td>8.32</td>
<td>2.12</td>
<td>9.36</td>
</tr>
<tr>
<td>Detection-ReID</td>
<td>9.41</td>
<td>2.20</td>
<td>27.60</td>
</tr>
<tr>
<td>Late fusion (Zhang and Chan 2019)</td>
<td>3.92</td>
<td>1.27</td>
<td>8.12</td>
</tr>
<tr>
<td>Naive early fusion (Zhang and Chan 2019)</td>
<td>5.45</td>
<td>1.25</td>
<td>8.10</td>
</tr>
<tr>
<td>MVMS (Zhang and Chan 2019)</td>
<td>3.49</td>
<td>1.03</td>
<td>8.01</td>
</tr>
<tr>
<td>3D counting (ours)</td>
<td>3.15</td>
<td>1.37</td>
<td>7.54</td>
</tr>
</tbody>
</table>

We perform best on two datasets and achieve comparable result on one dataset.

2) Ablation study on loss weights and ground-truth setting:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PET2009</th>
<th>DukeMTMC</th>
<th>CityStreet</th>
</tr>
</thead>
<tbody>
<tr>
<td>74×100cm</td>
<td>4.12</td>
<td>3.20</td>
<td>3.18 (n = 100)</td>
</tr>
<tr>
<td>148×200cm</td>
<td>4.88</td>
<td>4.57</td>
<td>4.24 (n = 10)</td>
</tr>
<tr>
<td>28×180cm</td>
<td>5.34</td>
<td>4.27</td>
<td>4.21 (n = 1)</td>
</tr>
</tbody>
</table>

\( \gamma \) is the hyperparameter for PCM loss. \( n \) is the number of voxels in the z-dimension (height), and \( h \) is the voxel height in the 3D world. The number of voxels in z-dimension is slightly larger (9, 18, 36) for DukeMTMC.

Future Work

- A more practical dataset for 3D counting;
- Cross-scene ability of the model.