Introduction

Problem: To produce accurate person counting given only sparse labelled data in crowded scenes.

State-of-the-art methods:
- Take a regression approach.
- Require exhaustive frame-wise labelling.
- Given a new scene, a model must be learned from scratch, repeating the laborious annotation process.

Contributions:
- Develop a unified active and semi-supervised crowd counting regression model using only a handful of annotations & lots of unlabelled data, to eliminate exhaustive data labelling.
- Formulate a transfer learning model based on crowd data manifold regularisation to utilise labelled crowd data from other crowded scenes.

Our Solution

Features:
- A set of perspective normalised low-level features similar to [1], such as foreground region and edge map, from each frame.

Training data:
- Some of them are labelled $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^l$ but most of them are unlabelled $\mathcal{U} = \{(x_j, y_j)\}_{j=1}^u$, where $l$ and $u$ are the number of labelled and unlabelled samples.

Key components:
- Active point selection: Select automatically the most informative frames for count annotation.
- Semi-supervised counting: Employ the underlying geometric structure of abundant unlabelled data and temporal continuity of crowd patterns.
- Transfer counting: Exploit labelled source data for counting in novel target scenes.

Active Point Selection

Given a fixed number of labelling budget, the most representative frames (in the sense of covering different crowd densities and counts) are the most useful ones to label.

Step 1: Construct an affinity matrix

$$A = \exp(D^{-1/2}AD^{-1/2})$$

where $D$ is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$ and perform spectral clustering.

Step 2: Construct normalised Laplacian $\mathcal{L} = D^{-1/2}AD^{-1/2}$

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Semi-supervised counting:

Step 1: Loss function

$$f^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda_1 \|\nabla f\|_2^2 + \lambda_2 \|\nabla f\|_b^2$$

1. Imposes smoothness to the possible solutions
2. Intrinsically structure of the crowd patterns
3. A penalty term to enforce temporal smoothness

where $\lambda_1, \lambda_2$ control the function complexity in the ambient space, intrinsic geometry of $\mathcal{F}$ and temporal space, respectively. $\beta$ represents the Hessian energy, and $L$ is the normalised Laplacian of temporal space, which is estimated with affinity matrix whose elements are $A_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$.

Step 2: Crowd density is estimated as

$$\hat{f}(x^*) = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \mathcal{L}^N(x^*, x_i) + \ell$$

where $x^*$ is the unseen point and $\mathcal{L} = [\alpha_1, \ldots, \alpha_{N-1}]^T$ is the expansion coefficient vector and $\ell$ is the bias term.

Transfer counting:

Step 1: Feature level alignment

Learn a function to project source data to a target scene $g: \mathcal{F}^{source} \rightarrow \mathcal{F}^{target} \in \mathbb{R}^{d}$

Step 2: Cross domain manifold estimation

Use the enlarged training set $\{(X^{source}\beta_i), X^{target} \}$ to [1] estimate a shared manifold. (2) learn a regression by the loss function above.

Advantages of cross domain manifold:
- To constrain the smoothness of solution w.r.t. intrinsic geometry of the cross domain data-space.
- To make the aligned source data less susceptible to noise.

Evaluation

Effect of # labelled and # unlabelled data

Comparison between the KRR (kernel ridge regression) baseline regression and the proposed semi-supervised regression (SSR) method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRR</td>
<td>8.040 ± 1.10</td>
</tr>
<tr>
<td>SSR (manifold)</td>
<td>7.945 ± 0.86</td>
</tr>
<tr>
<td>SSR (temporal)</td>
<td>7.296 ± 0.75</td>
</tr>
<tr>
<td>SSR (manifold+temporal)</td>
<td>7.329 ± 0.72</td>
</tr>
<tr>
<td>SSR (manifold+temporal+selection)</td>
<td>7.060 ± 0.62</td>
</tr>
</tbody>
</table>

Comparison vs. the state-of-the-arts:
- Consistently outperforms existing methods given sparse labelled samples

Transfer counting comparison:
- Transferring data without cross domain manifold (i.e. KRR) gives worse results.
- With cross domain manifold, SSR reduces the MSE further (in comparison to without transfer)

Examples

<table>
<thead>
<tr>
<th>Method</th>
<th>People Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRR</td>
<td>20%</td>
</tr>
<tr>
<td>SSR (temporal)</td>
<td>14.80 ± 0.87</td>
</tr>
<tr>
<td>SSR (manifold)</td>
<td>13.00 ± 0.72</td>
</tr>
<tr>
<td>SSR (manifold+temporal)</td>
<td>12.10 ± 0.72</td>
</tr>
<tr>
<td>SSR (manifold+temporal+selection)</td>
<td>11.20 ± 0.72</td>
</tr>
</tbody>
</table>

Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>People Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRR</td>
<td>20%</td>
</tr>
<tr>
<td>SSR (temporal)</td>
<td>14.80 ± 0.87</td>
</tr>
<tr>
<td>SSR (manifold)</td>
<td>13.00 ± 0.72</td>
</tr>
<tr>
<td>SSR (manifold+temporal)</td>
<td>12.10 ± 0.72</td>
</tr>
<tr>
<td>SSR (manifold+temporal+selection)</td>
<td>11.20 ± 0.72</td>
</tr>
</tbody>
</table>