From Semi-Supervised to Transfer Counting of Crowds

香港中文大學 The Chinese University of Hong Kong

Chen Change Loy
The Chinese University of Hong Kong
ccloy@ie.cuhk.edu.hk

Shaogang Gong
Queen Mary University of London
sgg@eecs.qmul.ac.uk

Tao Xiang
Queen Mary University of London
txiang@eecs.qmul.ac.uk

Queen Mary University of London

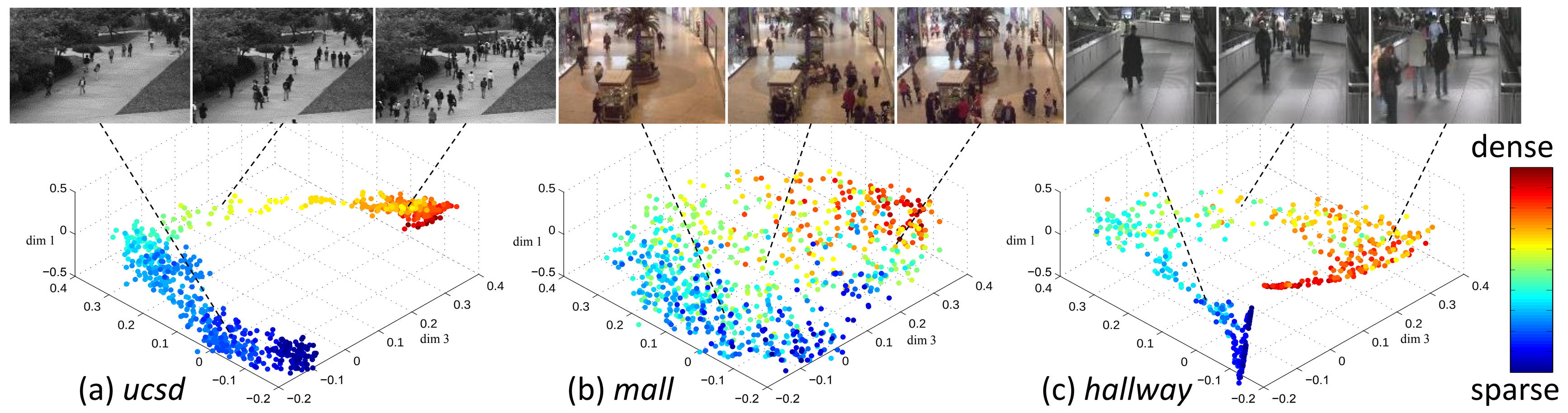
1 Introduction

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 crowd scenes.

2 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 $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ but most of them are unlabelled $\mathcal{U} = \{\mathbf{x}_j\}_{j=l+1}^{j=l+u}$ where l and u are the number of labelled and unlabelled samples.

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

3 Active Point Selection

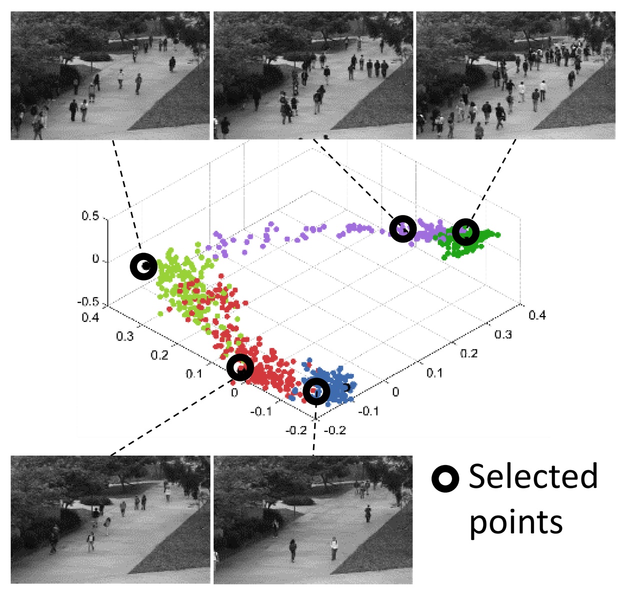
Given a fixed number of labelling budget, the most representative frames (in the sense of covering different crowd densities/counts) are the most useful ones to label.

Step-1: Construct an affinity matrix

$A \in \mathbb{R}^{(l+u)\times(l+u)}$

$A_{ij} = \exp\left(\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)/2\sigma^2\right)$

Step-2: Construct normalised Laplacian $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ where D is a diagonal matrix with $D_{ii} = \sum_j^{l+u} A_{ij}$ and perform spectral clustering.



Project page: http://personal.ie.cuhk.edu.hk/~ccloy/

[1] A. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. TIP, 21(4):2160–2177, 2012
[2] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In CVPR, 2013

4 Semi-supervised & Transfer Counting

Semi-supervised counting:

Step-1: Loss function

$$f^* = \operatorname{argmin}_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}[y_i - f(\mathbf{x}_i)]^2 + \lambda_A\|f\|_K^2 + \lambda_I \mathbf{f}^\top B \mathbf{f} + \lambda_T \mathbf{f}^\top L \mathbf{f}$$

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

where λ_A, λ_I and λ_T control the function complexity in the ambient space, intrinsic geometry of $p(\mathbf{x})$, and temporal space, respectively. B represents the Hessian energy, and L is the normalised Laplacian of temporal space, which is estimated with affinity matrix whose elements are $A_{ij} = \exp\left(\left(-\|t_i - t_j\|^2\right)/2\sigma^2\right)$

Step-2: Crowd density is estimated as

$$f^*(\mathbf{x}^*) = \sum_i^{l+u} \alpha_i K(\mathbf{x}^*, \mathbf{x}_i) + b$$

where $\mathbf{x}^*$ is the unseen point and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{l+u}]^\top$ is the expansion coefficient vector and b is the bias term.

Transfer counting:

Step-1: Feature level alignment

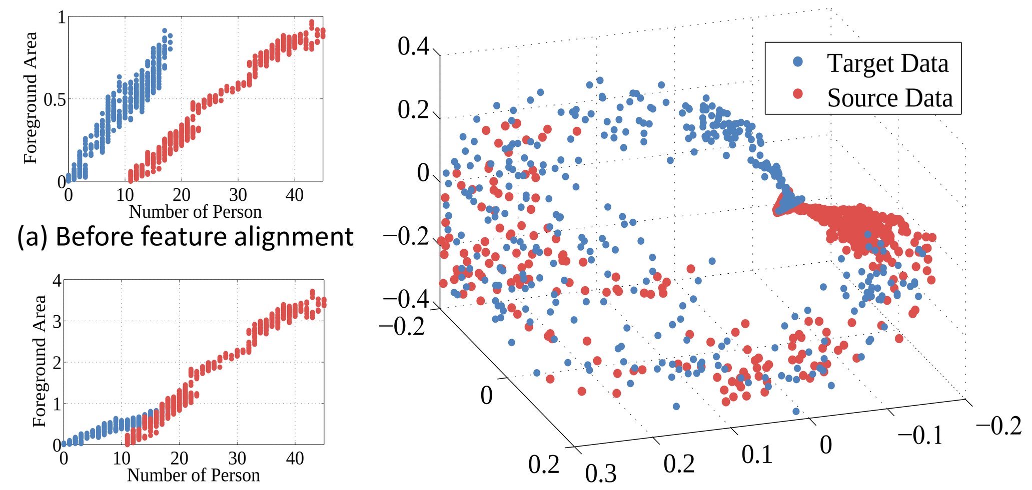
Learn a function to project source data to a target scene

$g : \hat{\mathbf{x}}^{\text{source}} \to \hat{\mathbf{x}}^{\text{target}} \in \mathbb{R}^d$

Step-2: Cross domain manifold estimation

Use the enlarged training set $g(X^{\text{source}}) \cup X^{\text{target}}$ 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.

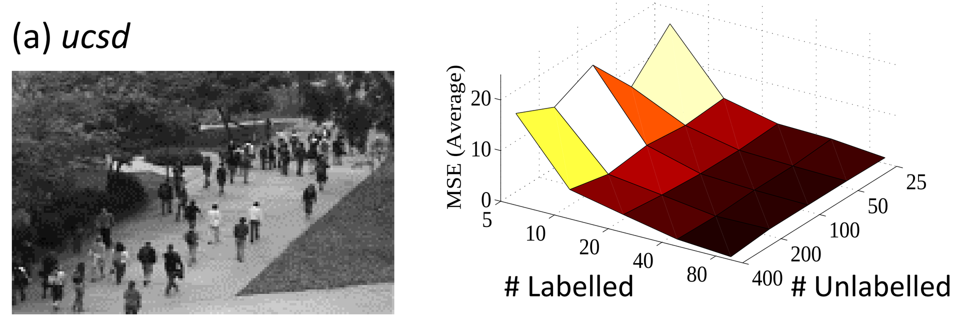


(a) Before feature alignment
(b) After feature alignment
(c) The embedding of the cross-domain manifold.

5 Evaluations

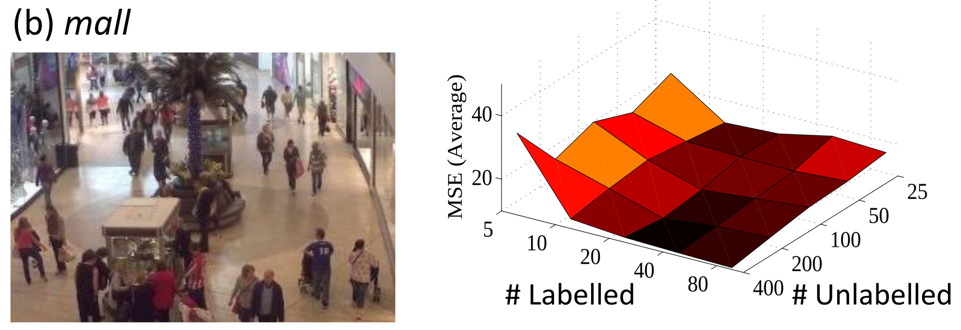
Datasets | **Effect of # labelled and # unlabelled data** | **Comparison between the KRR (kernel ridge regression) baseline regression and the proposed semi-supervised regression (SSR) method.**

(a) *ucsd*



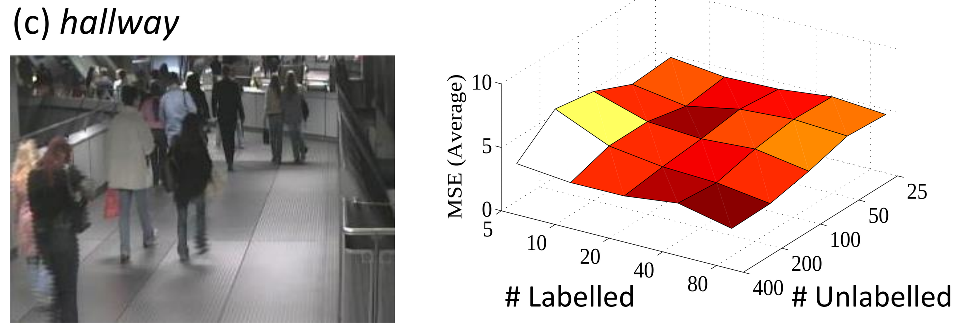
Method	Mean Squared Error
KRR	8.040 ± 1.10
SSR (manifold)	7.943 ± 0.86
SSR (temporal)	7.296 ± 0.75
SSR (manifold+temporal)	7.329 ± 0.72
SSR (manifold+temporal+selection)	**7.060 ± 0.62**

(b) *mall*



Method	Mean Squared Error
KRR	19.282 ± 3.83
SSR (manifold)	18.417 ± 3.35
SSR (temporal)	18.791 ± 3.53
SSR (manifold+temporal)	18.112 ± 3.38
SSR (manifold+temporal+selection)	**17.853 ± 2.38**

(c) *hallway*



Method	Mean Squared Error
KRR	7.971 ± 1.00
SSR (manifold)	7.389 ± 1.18
SSR (temporal)	6.828 ± 0.72
SSR (manifold+temporal)	5.546 ± 0.30
SSR (manifold+temporal+selection)	**5.342 ± 0.16**

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

Method	# train samples	ucsd	mall	hallway
Gaussian Processes Regression [1]	50	11.10	49.83	27.56
	Full	7.68	14.88	5.60
Cumulative Attribute Ridge Regression [2]	50	9.27	22.19	5.53
	Full	7.19	14.80	5.00
SSR	50	**7.06**	**17.85**	**5.34**

Measured in mean squared error (MSE)

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)

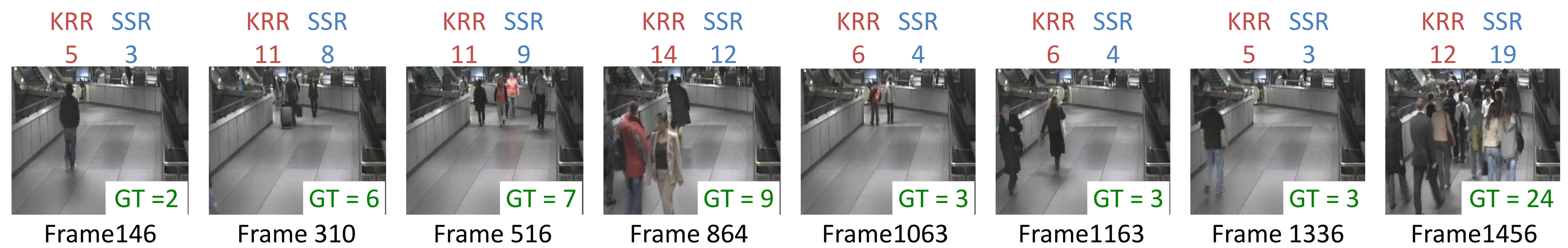
Source	Target	Without Transfer Counting	
		KRR	SSR
--	hallway	8.356 ± 0.70	6.285 ± 0.54
--	ucsd	8.538 ± 1.22	7.732 ± 0.93

Source	Target	With Transfer Counting	
		KRR	SSR
ucsd	hallway	16.848 ± 3.27	**5.984 ± 0.40**
hallway	ucsd	23.010 ± 5.66	**7.321 ± 1.86**

Measured in mean squared error (MSE)

6 Examples

- Compare counting accuracy between the KRR and our semi-supervised method SSR.
- SSR achieves 20% reduction in mean squared error with just 10% of labelled samples as compared to the KRR.



KRR = Kernel Ridge Regression
SSR = Our Semi-Supervised Method
GT = Ground Truth