From Semi-Supervised to Transfer Counting of Crowds

Chen Change Loy[1], Shaogang Gong[2], Tao Xiang[2]
[1] Dept. of Information Engineering, The Chinese University of Hong Kong
[2] School of EECS, Queen Mary University of London, UK
ccloy@ie.cuhk.edu.hk, {sgg,txiang}@eecs.qmul.ac.uk

Abstract

Regression-based techniques have shown promising results for people counting in crowded scenes. However, most existing techniques require expensive and laborious data annotation for model training. In this study, we propose to address this problem from three perspectives: (1) Instead of exhaustively annotating every single frame, the most informative frames are selected for annotation automatically and actively. (2) Rather than learning from only labelled data, the abundant unlabelled data are exploited. (3) Labelled data from other scenes are employed to further alleviate the burden for data annotation. All three ideas are implemented in a unified active and semi-supervised regression framework with ability to perform transfer learning, by exploiting the underlying geometric structure of crowd patterns via manifold analysis. Extensive experiments validate the effectiveness of our approach.

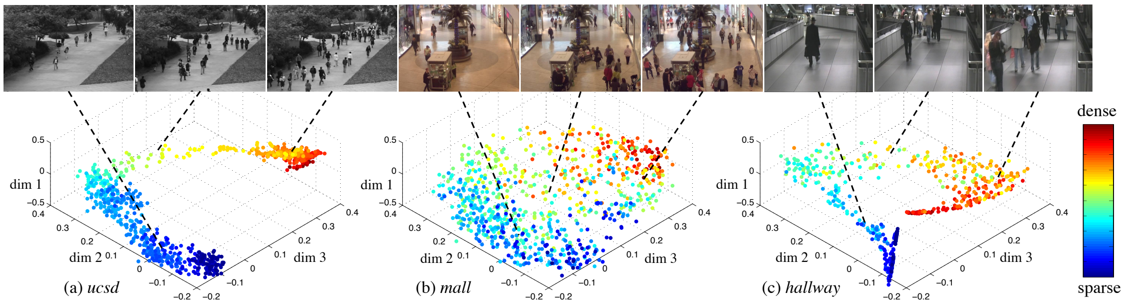
1. Introduction

Video-imagery based crowd counting [21] is important for profiling the population movement over time across spaces for establishing global situational awareness. Counting in crowded public spaces is non-trivial due to severe inter-object occlusion, scene perspective distortion, and visual ambiguity introduced by challenging lighting condition and complex human activities. State-of-the-art methods [9, 10, 19, 7] typically adopt regression-based techniques to learn a mapping between low-level features and people count, so as to circumvent explicit object segmentation and detection in crowded scenes. However, these techniques generally require exhaustive frame-wise labelling or even exact head-position annotations [19] to train a regression model. Furthermore, given a new scene or changed scene layout, a model has to be learned from scratch by repeating the laborious annotation process.

In this study, we aim to learn a regression model for crowd counting by annotating only a handful of frames (dozens rather than hundreds), so as to significantly reduce the amount of manual annotation and make the model much more applicable in practice. To achieve this goal, we wish to explore three ideas with different underlying assumptions. (1) Instead of exhaustively annotating every single frame, we design a model to select automatically and actively the most informative image frames for count annotation. The underlying assumption is that if the selected samples are informative and representative, this should have a minimal effect on the learned regression model as compared to learning from all exhaustively labelled frames. (2) For video-based crowd counting, potentially unlimited amount of data can be readily collected. Rather than learning from only labelled data, the abundant unlabelled data are to be exploited. We assume that the intrinsic distribution structure of those unlabelled data can be computed to facilitate both the learning of a regression counting model using only a handful of labelled data, and the selection of more informative image frames to label therefore further reinforcing the first idea above. (3) Instead of learning a regression model from scratch in every new scene, the labelled data from other scenes should also be exploited to compensate for the lack of labelled data in the new scene. The assumption for this idea is that there is transferrable knowledge in other scenes which can be employed to further alleviate the burden for data annotation. Although different scenes can be visually very different, the crowd patterns share some common grounds (e.g. larger crowd leads to large foreground areas) which correspond to transferrable knowledge.

In order to realise these three ideas for crowd counting with only a handful of labelled frames in one scene and generalising to other scenes, we develop a unified framework for active and semi-supervised learning of a regression model with transfer learning capability. The framework is formulated based on exploiting the underlying manifold structure of unlabelled crowd data to facilitate counting when the labelled samples are sparse. Many real-world data is supported on a low-dimensional manifold [4]. We observe that crowd pattern data often form a well structured manifold due to the inherent imaging process for generat-

[*] Most of the work was done when the first author was at Vision Semantics Ltd, London, UK.



(a) *ucsd* (b) *mall* (c) *hallway*

Figure 1. Three-dimensional embedding of crowd patterns obtained using multi-dimensional scaling. Each point corresponds to a global feature vector of crowd pattern of a video frame. Every point is encoded by colour so that points with higher crowd density are red and points with fewer people are blue. The details of the datasets, *ucsd*, *mall*, and *hallway* are provided in Sec. 4.

ing crowd patterns from shared physical spaces subject to social behavioural constraints [15]. Figure 1 shows different examples of manifold embedding of crowd patterns extracted from three different public scenes. It is evident that typically the crowd density (*e.g*. number of people) varies smoothly within the manifold space. To exploit the existence of such underlying geometric structures of crowd patterns for learning a regression model without exhaustively labelling the data, we develop a semi-supervised regression model with manifold regularisation to assimilate the count estimation of two nearby crowd pattern points in the manifold. This formulation builds on the Laplacian regularised least squares concept [25], but is reformulated carefully to employ Hessian energy [18, 24] for manifold regularisation due to the latter's superior extrapolation potential for semi-supervised learning of a regression function. Modelling the underlying crowd pattern structure also provides a solution to active regression learning. That is, it can help to select fewer and more informative data points to be labelled, given limited labelling budget[1]. This active sample selection for annotation is integrated seamlessly into our semi-supervised model training.

In addition to exploiting intrinsic structures of unlabelled data collected from the same scene for active and semi-supervised regression modelling, we further develop a transfer learning capability to utilise available labelled data from other scenes. Transferring the shared common information from training data across different visual domains requires a process to filter out changes caused by different camera viewing angles or activity patterns. We wish to explore the following consideration: If two different scenes share similar imaging generating processes, they may also share a similar underlying manifold structure from the sampled data, suggesting that the knowledge available from one scene can be re-used in another scene. In this study, we investigate in particular how manifold regularisation would help in learning a crowd counting model with labelled data collected from a different scene. We call this *transfer counting*.

2. Related Work

Crowd counting: Various approaches to crowd counting have been proposed [21], including counting-by-detection [20, 39, 12], counting-by-clustering [6, 29], and counting-by-regression [9, 10, 19, 7]. The latter is favoured by most recent studies due to its robustness against occlusion. The regression-based techniques are fundamentally supervised methods, which often assume the availability of large amount of labelled data for training. Tan *et al*. [31] relax this assumption by presenting a semi-supervised learning framework, which utilises sequential information in the unlabelled frames to penalise sudden prediction change. This method relies on the assumption that the temporal space is dense, *i.e*. high enough video frame rate is required to capture the smoothness in crowd pattern change over time. This assumption can be too stringent for many real-world scenarios when data bandwidth and storage space is limited, or continuous high frame-rate video recording is not available [23]. Our approach relaxes this assumption since our method explores smoothness in intrinsic crowd pattern distribution structure, not only in the video stream temporal space, leading to a more generic/scalable and robust approach to crowd counting estimation (see comparative experiments in Sec. 4.1). Importantly, in the same framework the model is capable of transfer counting.

Semi-supervised and transfer learning: Manifold learning has been widely explored in computer vision, such as face expression [8] and age estimation [13]. The intuition of incorporating manifold regularisation in semi-supervised learning has also been studied [1, 4, 38, 18], whilst manifold-based transfer learning has been proposed in [34] to transfer knowledge across domains via an aligned manifold. However, no crowd counting studies have attempted manifold regularisation for achieving semi-supervised and transfer counting. Although existing work on manifold learning are relevant for our problem, applying them directly for active and semi-supervised regression modelling of crowd count is non-trivial and has not been attempted before. Note that the term 'tranfer counting' has been first

[1] A 'budget' is the funds (or time) available to spend on annotation [32].

used in [35]; but it refers to transferring knowledge across overlapping camera views for the *same* crowd from the *same* scene whilst we are concerned with transferring between completely unrelated scenes – a much more generic and realistic setting.

Active learning for regression: The problem of how to select data points for labelling is addressed by active learning [30, 22], mostly for classification rather than regression. Recently, a few studies have been devoted to regression-based active data selection, including D-optimality [14] and E-optimality [24] designs. These methods stem from the idea of optimal experimental design [2], which either aims to minimise a model's prediction error, output variance, or parameter variance by selecting informative samples. For instance, a data selection method based on optimal experimental design is proposed by He [14]. However, it may suffer from sensitivity problem during the evaluation of determinant of Hessian matrix, due to the difficulty in determining small eigenvalues [5]. To circumvent the sensitivity problem, we take an approach similar to the robust experimental design method of [11] to identify supporting points via clustering. Our clustering-based data selection method can be considered as a degenerate case of the active learning approach [14] in that our algorithm still selects a set of informative points for human to label but in a batch manner without updating the sampling strategy sequentially, in exchange for a more stable behaviour in data selection.

Our contributions are three-fold: (1) To eliminate exhaustive data labelling for learning a regression based crowd counting model, this is the first study to systematically develop a unified active and semi-supervised crowd counting regression model using only a handful of annotations. (2) A concept of transfer counting with practical potential is proposed and a transfer learning model based on crowd data manifold regularisation is formulated to utilise labelled crowd data from other crowd scenes. (3) Extensive comparative evaluations are conducted using two publicly available crowd datasets and a new dataset extracted from the i-LIDS dataset [16] to demonstrate the effectiveness of the proposed approach.

3. Learning Inherent Constraints for Counting

3.1. Semi-supervised Crowd Counting

Counting by regression: Taking a regression approach to crowd counting, one typically extracts a set of perspective normalised low-level features $\mathbf{x}$ from each frame, *e.g.* foreground segments or an edge map, and subsequently learns a model to predict the crowd density given the low-level features. Ridge Regression (RR) or its kernelised version, Kernel Ridge Regression (KRR) have shown promising performance for crowd counting regression [10]; it is thus chosen as the regression baseline model in our framework. Formally, given a set of l labelled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, of samples $\mathbf{x}_i$ from $X \subseteq \mathbb{R}^d$ with corresponding labels y_i in $Y \subseteq \mathbb{R}$, KRR estimates the unknown regression function as

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum\nolimits_{i=1}^{l} V\left(\mathbf{x}_i, y_i, f\right) + \lambda \|f\|_K^2, \quad (1)$$

where V is a loss function, typically the squared loss $[y_i - f(\mathbf{x}_i)]^2$ for Regularised Least Squares (RLS) regression problem. The kernel K is a positive definite Mercer kernel $K : X \times X \to \mathbb{R}$, and there is an associated Reproducing Kernel Hilbert Space (RKHS) of functions $X \to \mathbb{R}$. Penalising the RKHS norm $\|f\|_K^2$ imposes smoothness to the possible solutions.

Semi-supervised regression: A semi-supervised regression method is specifically formulated here to produce accurate person counting given only sparse labelled data. This is made possible by exploiting the underlying geometric structure of abundant unlabelled data and temporal continuity of crowd pattern. More precisely, given a set of training data, we assume 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, respectively. A user shall only label a few data points and the rest of the unlabelled training data will be annotated automatically by inference using the model. A set of regression coefficients is estimated at the end for inductive inference given unseen data.

Our goal is to perform semi-supervised learning to assimilate the vast majority of unlabelled data points $\mathcal{U}$ by the labels of the small minority $\mathcal{L}$. This is computed by a joint regularisation through learning the crowd pattern intrinsic distribution (geometric) structure ($p(\mathbf{x})$) and imposing temporal smoothness of activity patterns in the scene. In other words, we would like to ensure that the solution is optimal with respect to three considerations: (1) regression in a reduced kernel space (RKHS), (2) the marginal distribution of unlabelled data points $p(\mathbf{x})$, and (3) temporal continuity in the physical space. To achieve this, we introduce the following additional regularisers to Eqn. (1):

$$\begin{aligned} f^* &= \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum\nolimits_{i=1}^{l} [y_i - f(\mathbf{x}_i)]^2 \\ &\quad + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2 + \lambda_T \|f\|_T^2, \end{aligned} \quad (2)$$

where $\|f\|_I^2$ is a regularisation term to reflect the intrinsic structure of the crowd patterns, whilst $\|f\|_T^2$ is a penalty term to enforce temporal smoothness. Here 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. We now explain each regularisation terms in detail.

Distribution structure regularisation: The underlying distribution structure (geometrical) of crowd patterns can be modelled using a crowd manifold. We assume that the marginal distribution $p(\mathbf{x})$ is supported on a low-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$. In particular, if two samples $\mathbf{x}_i$, $\mathbf{x}_j$ are close in the intrinsic geometry of $p(\mathbf{x})$, then the conditional distributions $p(y|\mathbf{x}_i)$ and $p(y|\mathbf{x}_j)$, *i.e.* the crowd density are similar.

Several choices of $\|\cdot\|_I$ exist. We adopt the Hessian regularisation function introduced in [18], which has a noticeable difference in comparison to the more commonly used Laplacian regularisation [4]. Specifically, Hessian regularisation prefer functions that vary linearly with respect to the geodesics on the data manifold [18]. This property is particularly critical for enabling better extrapolation behaviour in solving a semi-supervised regression problem.

The Hessian regulariser is the squared norm of the second covariant derivative, $\|\nabla_a \nabla_b f\|^2$, corresponding to the Frobenius norm of the Hessian of f in normal coordinates. An estimate of $\|\nabla_a \nabla_b f\|^2$ of $\mathbf{x}_i$ is given as

$$\|\nabla_a \nabla_b f\|^2 \approx \sum\nolimits_{\gamma,\,\beta=1}^{k} \mathbf{f}_\gamma \mathbf{f}_\beta B^i_{\gamma\beta}, \tag{3}$$

where $\mathbf{f}_j = f(\mathbf{x}_j)$, and $\mathbf{x}_\gamma, \mathbf{x}_\beta$ are the set of k nearest neighbour, $N_k(\mathbf{x}_i)$, of point $\mathbf{x}_i$ in a k-NN graph. Here $B^i_{\gamma\beta}$ represents the local Hessian energy of $\mathbf{x}_i$ estimated through second-order polynomial fitting in normal coordinates [18]. The total estimated Hessian energy is a sum over all $(l+u)$ labelled and unlabelled points

$$\hat{S}_{\text{Hess}}(f) = \sum_{i=1}^{l+u} \sum_{\gamma \in N_k(\mathbf{x}_i)} \sum_{\beta \in N_k(\mathbf{x}_i)} \mathbf{f}_\gamma \mathbf{f}_\beta B^i_{\gamma\beta} = \mathbf{f}^\mathsf{T} B \mathbf{f}. \tag{4}$$

The regression loss function of Eqn. (2) is now re-written as

$$\begin{aligned} f^* \;\; = \;\; & \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum\nolimits_{i=1}^{l} [y_i - f(\mathbf{x}_i)]^2 \\ & + \lambda_A \|f\|_K^2 + \lambda_I \mathbf{f}^\mathsf{T} B \mathbf{f} + \lambda_T \|f\|_T^2. \end{aligned} \tag{5}$$

Temporal regularisation: The temporal constraint can be incorporated easily into our framework by assuming that if two observations $\mathbf{x}_i$ and $\mathbf{x}_j$ occur close in time, then the crowd density should not differ significantly. Again, several choices of $\|\cdot\|_T$ exist. Empirically we found that Laplacian yields better performance than Hessian for temporal regularisation. To estimate a normalised Laplacian, we first construct an affinity matrix $A \in \mathbb{R}^{(l+u)\times(l+u)}$ defined by $A_{ij} = \exp\left(\left(-\|t_i - t_j\|^2\right)/2\sigma^2\right)$, for $i \neq j$ and $A_{ii} = 0$, $t \in \{1, 2, \dots\}$ is the time index of each frame, and the scale parameter σ is automatically inferred using the self-tuning approach [36]. Intuitively, A_{ij} has a high value for neighbouring samples in time and a low value if the samples are far apart temporally. The normalised Laplacian L is computed as

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{6}$$

where D is a diagonal matrix with $D_{ii} = \sum_j^{l+u} A_{ij}$.

Our final loss function to be minimised is defined as

$$\begin{aligned} f^* \;\; = \;\; & \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum\nolimits_{i=1}^{l} [y_i - f(\mathbf{x}_i)]^2 \\ & + \lambda_A \|f\|_K^2 + \lambda_I \mathbf{f}^\mathsf{T} B \mathbf{f} + \lambda_T \mathbf{f}^\mathsf{T} L \mathbf{f}. \end{aligned} \tag{7}$$

This is solved efficiently using either the Newton's method [17] or preconditioned conjugate gradient [25] to obtain the $(l+u)$-dimensional expansion coefficient vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_{l+u}]^\mathsf{T}$ coefficients and the optimal bias b. By the representer theorem, given an unseen low-level feature vector $\mathbf{x}^*$, the crowd density is estimated as

$$f^*(\mathbf{x}^*) = \sum\nolimits_i^{l+u} \alpha_i K(\mathbf{x}^*, \mathbf{x}_i) + b. \tag{8}$$

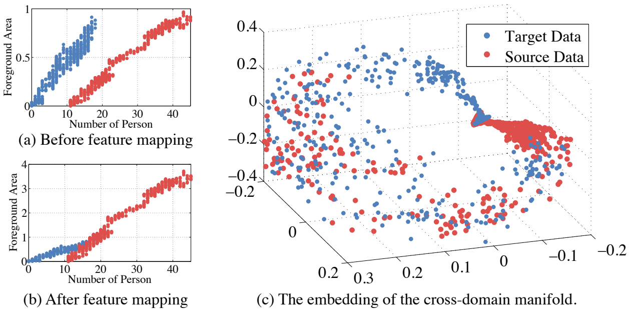
3.2. Active Learning for Regression

Having formulated the model above for learning crowd counting using only a handful of labelled data supported by a large number of unlabelled data, we now address the problem of how to actively select the optimal handful of labelled data so that they have a maximal impact on learning the model. Our intuition is that 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. This brings in a chicken-and-egg problem – without labelling all frames, how does one know which ones are representative? To solve this problem, we propose to discover these representative points ("supporting points") through clustering in the crowd marginal distribution structure (manifold).

Specifically, given a crowd manifold learned from a set of unlabelled data, we perform spectral clustering [26] on the data projected onto the manifold. That is, we treat the problem of actively picking data points for labelling as of partitioning a weighted graph. Each node in the graph corresponds to frame-level global crowd patterns, connected by edges whose weights are defined by the affinity between the patterns. More precisely, we construct an affinity matrix $A^s \in \mathbb{R}^{(l+u)\times(l+u)}$ with $A_{ij} = \exp\left(\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)/2\sigma^2\right)$, where σ is determined via self-tuning method [36]. Similar to Eqn. (6), a normalised Laplacian L^s is then constructed. Given a designated annotation budget of K_A, we find the K_A largest eigenvectors of L^s, forming a matrix $E \in \mathbb{R}^{(l+u)\times K_A}$. Finally, we cluster E (with its rows unit-length normalised) into K_A clusters using k-means algorithm. The K_A supporting points are estimated as the cluster centres.

3.3. Transfer Counting

For transfer learning in general, one considers a given sparse set of labelled target training instances $\mathcal{L}^{\text{target}} = \{(\mathbf{x}^{\text{target}}, y^{\text{target}})\}$. In addition, a set of labelled training instances collected from a related source $\mathcal{L}^{\text{source}} =$



(a) Before feature mapping

(b) After feature mapping

(c) The embedding of the cross-domain manifold.

Figure 2. (a)-(b) Performing feature mapping using the corresponding points to align the feature range of *ucsd* and *hallway* datasets. (c) The embedding of the cross-domain manifold using the source data *ucsd* (red dots) and target data *hallway* (blue dots).

$\{(\mathbf{x}^{\text{source}}, y^{\text{source}})\}$ are also made available. The objective is to transfer the knowledge in $\mathcal{L}^{\text{source}}$ in order to facilitate learning of the target model.

In the context of transfer crowd counting, we consider that the most straightforward approach to transferring labelled data from one scene to another is feature-representation transfer [28]. More specifically, we wish to first obtain perspective normalised features [7], followed by feature-level alignment [33] to ensure the features extracted from disparate scenes lie within the same space.

To perform feature-level alignment, we assume n pairs of 'corresponding samples', $\{\hat{\mathbf{x}}^{\text{source}}, \hat{\mathbf{x}}^{\text{target}}\}$ with identical count labels, that is $y^{\text{target}} = y^{\text{source}}$, are available. A possible way to align the features is by estimating a mapping $g : \hat{\mathbf{x}}^{\text{source}} \rightarrow \hat{\mathbf{x}}^{\text{target}} \in \mathbb{R}^d$, and align the remaining source data to the target domain as $g(\mathbf{x}^{\text{source}})$. We choose a simple linear form for the function g, that is $g(X^{\text{source}}) = X^{\text{source}}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a diagonal matrix with $\boldsymbol{\beta}_{ii} = (\beta_1, \ldots, \beta_n)$. The values of $\boldsymbol{\beta}$ are estimated in a least-squares sense. An example of feature-level alignment is shown in Fig. 2(a)-(b).

After aligning X^{source} to the target domain, we combine them together with the target data to form an expanded training set, $g(X^{\text{source}}) \cup X^{\text{target}}$. This new enlarged training set is then employed to estimate a shared manifold (Fig. 2(c)) and to learn a regression function following the steps described in Sec. 3.1.

Note, using the aligned source data in its *original* high-dimensional form may lead to poor result due to the likely suboptimal feature alignment caused by poor corresponding points selection or imperfect perspective normalisation. Therefore, the step for learning a shared manifold is rather critical in that it allows one to constrain the smoothness of our solution with respect to the intrinsic geometry of the *cross-domain* data space. In particular, the locality-preserving character in learning a manifold with dominant eigenvectors makes the solution less susceptible to noise or small deviations in the aligned source data [3].

Data	$\mathbf{N_f}$	**R**	**D**	**Tp**	**Tr/Te**
ucsd [7]	2000	238 × 158	11–46	49885	800/1200
mall [10]	2000	320 × 240	13–53	62325	800/1200
hallway	2200	360 × 288	0–30	14707	500/1700

Table 1. Dataset properties: N_f = number of frames, R = Resolution, D = Density (minimum and maximum number of people in the ROI), Tp = total number of pedestrian instances, Tr/Te = number of frames in the training (Tr) and testing (Te) partitions.

4. Experiments

Datasets: Apart from the established UCSD pedestrian dataset (*ucsd*) [7] and a more recent shopping mall dataset (*mall*) [10, 9, 21], we introduce a new dataset in this study for comparative evaluation, referred to as the i-LIDs hallway dataset (*hallway*). A unique characteristics of this new dataset is its severe perspective distortion and occlusion. The *hallway* dataset is composed of 2200 frames extracted at 3 frames per second (fps) from the sequence ABTEN201c of the i-LIDS dataset [16]. We annotate the data by labelling the head position of every pedestrian in all frames[2]. Example frames of these datasets are shown in Fig. 1, with their details given in Table 1. Both the *hallway* and *mall* datasets are challenging. In particular, the perspective distortion, especially in the *hallway* dataset, is heavier than that in the *ucsd* dataset, resulting in more severe inter-object occlusion, and larger change in object size and appearance at different depths of the scene. In addition, the *mall* dataset is challenging in that it covers crowd densities from sparse to crowded, as well as diverse activity patterns (static and moving crowds) under large range of illumination conditions at different time of the day.

Features: For each dataset, we set a region of interest (ROI) to exclude the non-corridor/non-pathway regions in the scene. From the ROI, we extract segment-based and structural-based features following the methods described in [10]. For local texture features, we adopt uniform Local Binary Patterns (LBP) [27], which frequently corresponds to primitive micro-features such as edges and corners. For both the *ucsd* and the *hallway* datasets, scene lighting is stable so we employ a static background subtraction method to extract the foreground segments. For the *mall* dataset, gradual illumination change is present, we therefore adopt a GMM-based dynamic background modelling method. All features are perspective normalised [7].

Evaluation metric: We employ *mean squared error* (MSE) in performance evaluation, $\epsilon_{\text{sqr}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where N is the total number of test frames, y_i is the actual count, and $\hat{y}_i$ is the estimated count of ith frame. All results are averaged over 20 trials unless specified.

[2] The ground truth, together with the extracted features, and the train/test partitions can be downloaded at http://www.ie.cuhk.edu.hk/~ccloy/.

Method	Transductive	Inductive
KRR	2.780 ± 0.46	8.040 ± 1.10
SSR (manifold)	1.751 ± 0.25	7.943 ± 0.86
SSR (temporal)	1.213 ± 0.29	7.296 ± 0.75
SSR (manifold+temporal)	1.224 ± 0.29	7.329 ± 0.72
SSR (manifold+temporal+selection)	**1.005** ± 0.05	**7.060** ± 0.62

(a) *uscd*

Method	Transductive	Inductive
KRR	12.871 ± 1.51	19.282 ± 3.83
SSR (manifold)	13.088 ± 1.81	18.417 ± 3.35
SSR (temporal)	11.921 ± 1.46	18.791 ± 3.53
SSR (manifold+temporal)	11.726 ± 1.41	18.112 ± 3.38
SSR (manifold+temporal+selection)	**11.437** ± 0.88	**17.853** ± 2.38

(b) *mall*

Method	Transductive	Inductive
KRR	2.559 ± 0.46	7.971 ± 1.00
SSR (manifold)	2.770 ± 0.41	7.389 ± 1.18
SSR (temporal)	2.189 ± 0.22	6.828 ± 0.72
SSR (manifold+temporal)	1.774 ± 0.09	5.546 ± 0.30
SSR (manifold+temporal+selection)	**1.634** ± 0.03	**5.342** ± 0.16

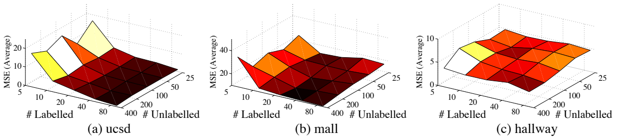
(c) *hallway*

Table 2. Performance comparison between the KRR baseline regression and the proposed semi-supervised regression (SSR) method: with manifold regularisation, temporal regularisation, a combination of two, and finally the automatic labelled data selection. The performance is measured in mean squared error (MSE), averaged over 20 trials. A smaller MSE value is better.

Parameter settings: The proposed method has a few free parameters, including the number of neighbours k to build the k-NN graph, the dimensionality m of PCA subspace during the determination of normal coordinates, and the regularisation parameters λ_A, λ_I, and λ_T. The Kernel Ridge Regression (KRR) has two free parameters, the bandwidth of Gaussian kernel and the regularisation parameter λ. All the above free parameters for each method were optimally estimated by cross validation on the labelled samples.

4.1. Semi-Supervised Crowd Counting

Semi-supervised learning: The goal of this experiment is to evaluate the effectiveness of exploiting unlabelled data distribution structure and temporal regularisations in the semi-supervised regression (SSR) learning framework. All datasets are partitioned into training/test sets in accordance to the numbers given in the last column of Table 1. Note that we follow [7] and [10] in partitioning the *ucsd* and *mall* datasets. A total of 50 samples in the training partition are randomly selected as labelled samples, while the rest of the samples in the training partition (750 in both *ucsd* and *mall*, and 450 in *hallway*) remain unlabelled. We evaluate the transductive learning (tested with unlabelled data in the training partition) and inductive inference (tested with unlabelled data in the test partition) performances of the proposed SSR method with different regularisation terms. The results obtained from KRR without semi-supervised learning is also reported as a baseline.



(a) ucsd (b) mall (c) hallway

Figure 3. The effect of labelled and unlabelled data.

It is evident from Table 2 that semi-supervised learning improves remarkably the crowd counting performance with the help of unlabelled data, *i.e.* an average of 18% reduction in MSE over KRR when we apply labelled data selection. Interestingly, given datasets of better frame rates, *e.g.* the *ucsd* with 10 fps and the *hallway* with 3fps, slightly better results were obtained using the temporal smoothness constraint in comparison to manifold regularisation. On the *mall* dataset with (∼1-2 fps), the effect of temporal smoothness decreased notably. In general, combining both regularisation terms yielded better and more reliable performance.

We further examine the effect of labelled and unlabelled data, by measuring the MSE performances on labelled set $\{5, 10, 20, 40, 80\}$ given unlabelled set $\{25, 50, 100, 200, 400\}$. Figure 3 shows clearly that adding more unlabelled data improved the counting performance. For instance, given 80 labelled data, the MSE in the *ucsd*, *mall*, and *hallway* datasets were reduced by nearly 7%, 22%, and 19% respectively, when we increased the unlabelled data size from 25 to 400.

Active learning for labelled points selection: In this experiment we compare our manifold-based "supporting point" selection method (m-landmark) (see Sec. 3.2) with k-means landmark discovery [31] and the random selection baseline (RAND). Figure 4 shows that in general, both the method in [31] and our method outperform the random selection. For instance, our method constantly outperforms RAND by around 7%-9% reduction in MSE on the *ucsd* and *hallway* datasets. This is not surprising since the latter blindly selects instances that may not contribute towards the regression model learning. The result also shows that compared to [31], our method gains better performance on the *ucsd* and *mall* datasets, and more stable performance overall (see the standard deviation plots in Fig. 4).

Active semi-supervised learning: Figure 5 shows a comparison of the actual counting performance between KRR (without semi-supervised learning) and our full active semi-supervised regression method. As compared to KRR that employed all 500 data for training, our method only requires 10% of training data whilst simultaneously achieving a remarkable 20% reduction in MSE. Note that both methods fail to estimate the spike around frame 1500 because the range of counts beyond 20 are not captured in the training set. One can resolve this problem by simply collecting more diverse training data with a wider-range of counts.

Table 3 shows the quantitative comparison of our SSR

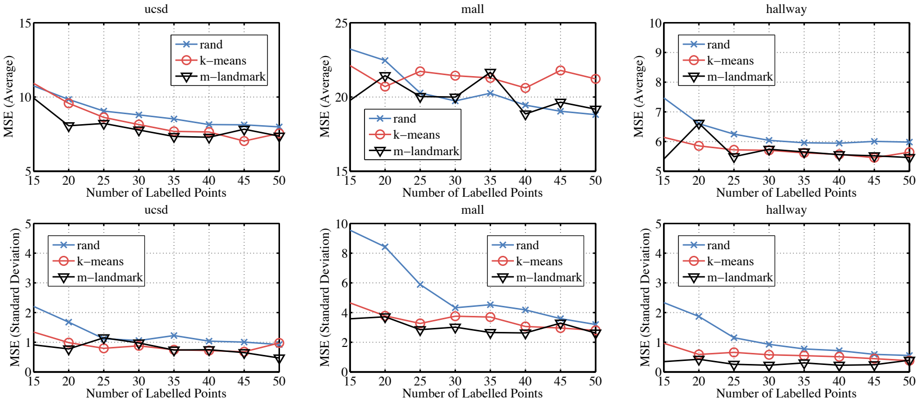


Figure 4. Count estimation performance using three different labelled data selection methods. The first row reports the average MSE whilst the second row shows the associated standard deviation plots. The 'm-landmark' is the proposed active selection method.

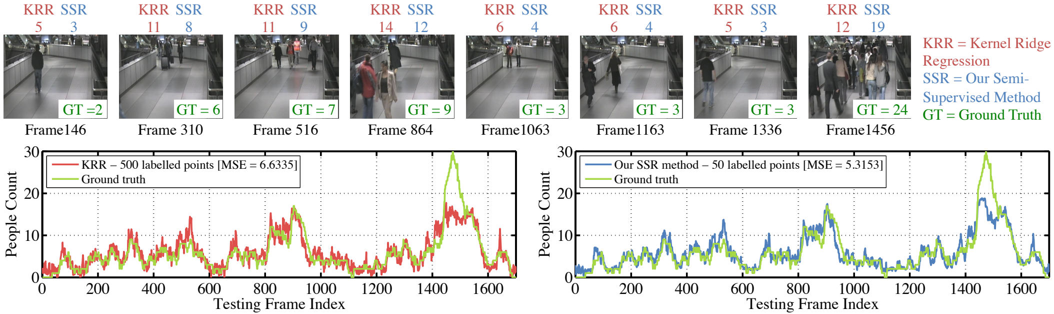


Figure 5. Comparison of counting performance between the KRR and our semi-supervised method SSR on the *hallway* dataset. Note that our method achieved 20% reduction in mean squared error with just 10% of labelled samples as compared to the KRR.

method against two recently proposed models [7, 9], all of which use the same feature representations (Sec. 4). The proposed SSR approach not only consistently outperforms existing methods given sparse labelled samples (50 samples), but also performs comparatively to GPR and CA-RR that learn from full training set.

Method	# train samples	*ucsd* [7]	*mall* [10]	*hallway*
GPR [7]	50	11.10	49.83	27.56
GPR [7]	Full	7.68	14.88	5.60
CA-RR [9]	50	9.27	22.19	5.53
CA-RR [9]	Full	7.19	14.80	5.00
SSR	50	**7.06**	**17.85**	**5.34**

Table 3. Comparison against the state-of-the-art methods: GPR = Gaussian Processes Regression, CA-RR = Cumulative Attribute Ridge Regression, SSR = the proposed Semi-Supervised Regression method. The performance is measured in mean squared error.

4.2. Transfer Crowd Counting

In this experiment we evaluate the proposed transfer counting method (Sec. 3.3). We randomly selected 100 random labelled samples from the source data to be transferred for target model learning. In addition, a total of 50 random labelled data in the target scene are chosen for bootstrapping, 25 of which have corresponding labels with the source labelled set. Those 25 pairs of corresponding samples are employed to learn a mapping function for aligning the source labelled set.

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

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

Table 4. Transfer counting results.

The *ucsd* and *hallway* datasets are selected in this experiment. Table 4 summarises the transfer counting results averaged over 10 trials. The top half of Table 4 shows the results on using KRR and SSR without transfer learning, *i.e*. using the 50 labelled data in the target scene for model learning. In the bottom half of the table, we show the transfer learning results on both models, of which training are conducted using the target scene data as well as 100 labelled data from the source domain. It is evident that transferring the data without learning a cross-domain manifold (*i.e*. using KRR) results in worse results in comparison to training with just target data alone. On the other hand, our transferring method helps in reducing the MSE further (in comparison to without transfer) with the use of cross-domain manifold.

The above results suggest that poor results may be obtained if the suboptimal aligned source samples are employed directly in the target model training. However, when those source data are embedded in a shared cross-domain manifold together with the target data, they can effectively help in filling the 'gap' not captured in the target labelled data, leading to a more accurate estimation.

5. Conclusion

In contrast to most existing crowd counting studies that rely on exhaustive annotations for model training, a unified active and semi-supervised regression approach is formulated to enable crowd counting with just a few labelled sample images through exploiting the underlying distribution structure of crowd patterns given readily available vast quantity of unlabelled data. In addition, we proposed a novel concept of and a model for transfer counting. We demonstrated that the lack of labelled data in a new scene can be helped by knowledge transferred from other scenes in minimising the effort required for bootstrapping crowd counting at the new scene. This has significant practical value. In the current transfer counting method, we imposed an assumption that the source and target data sharing a similar manifold representation. Future work will explore ways to relax this assumption through automatic estimation of source-target relevance [37].

References

[1] A. Argyriou, M. Herbster, and M. Pontil. Combining graph laplacians for semi-supervised learning. In *NIPS*, pages 67–74, 2005. 2

[2] A. Atkinson, A. Donev, and R. Tobias. *Optimum experimental designs*. Oxford University Press, 2007. 3

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *NC*, 15(6):1373–1396, 2003. 5

[4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006. 1, 2, 4

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2007. 3

[6] G. J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In *CVPR*, pages 594–601, 2006. 2

[7] A. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. *TIP*, 21(4):2160–2177, 2012. 1, 2, 5, 6, 7

[8] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *IVC*, 24(6):605–614, 2006. 2

[9] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013. 1, 2, 5, 7

[10] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012. 1, 2, 3, 5, 6, 7

[11] H. Dror and D. Steinberg. Robust experimental design for multivariate generalized linear models. *Technometrics*, 48(4):520–529, 2006. 3

[12] W. Ge and R. Collins. Marked point processes for crowd counting. In *CVPR*, pages 2913–2920, 2009. 2

[13] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *TIP*, 17(7):1178–1188, 2008. 2

[14] X. He. Laplacian regularized D-optimal design for active learning and its application to image retrieval. *TIP*, 19(1):254–263, 2010. 3

[15] D. Helbing, A. Johansson, and E. T. Hochschule. *Pedestrian, crowd and evacuation dynamics*. Swiss Federal Institute of Technology, 2009. 2

[16] i-LIDS Team. Imagery library for intelligent detection systems (i-LIDS); a standard for testing video based detection systems. In *IEEE International Carnahan Conferences Security Technology*, pages 75–80, 2006. 3, 5

[17] S. Keerthi and D. DeCoste. A modified finite Newton method for fast solution of large scale linear SVMs. *JMLR*, 6(1):341–361, 2005. 4

[18] K. Kim, F. Steinke, and M. Hein. Semi-supervised regression using Hessian energy with an application to semi-supervised dimensionality reduction. *NIPS*, 22:979–987, 2009. 2, 4

[19] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *NIPS*, 2010. 1, 2

[20] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *ICPR*, pages 1–4, 2008. 2

[21] C. C. Loy, K. Chen, S. Gong, and T. Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation, and Visual Analysis of Large Crowds*. Springer, 2013. 1, 2, 5

[22] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong. Stream-based joint exploration-exploitation active learning. In *CVPR*, 2012. 3

[23] C. C. Loy, T. Xiang, and S. Gong. Incremental activity modelling in multiple disjoint cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(9):1799–1813, 2012. 2

[24] K. Lu, J. Zhao, and Y. Wu. Hessian optimal design for image retrieval. *PR*, 44(6):1155–1161, 2011. 2, 3

[25] S. Melacci and M. Belkin. Laplacian support vector machines trained in the primal. *JMLR*, pages 1149–1184, 2011. 2, 4

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001. 4

[27] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002. 5

[28] S. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010. 5

[29] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, pages 705–711, 2006. 2

[30] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. 3

[31] B. Tan, J. Zhang, and L. Wang. Semi-supervised elastic net for pedestrian counting. *PR*, 44(10):2297–2304, 2011. 2, 6

[32] S. Vijayanarasimhan, P. Jain, and K. Grauman. Far-sighted active learning on a budget for image and video recognition. In *CVPR*, pages 3035–3042, 2010. 2

[33] C. Wang. *A geometric framework for transfer learning using manifold alignment*. PhD thesis, University of Massachusetts Amherst, 2010. 5

[34] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *ICML*, pages 1120–1127, 2008. 2

[35] M.-F. Weng, Y.-Y. Lin, N. C. Tang, and H.-Y. M. Liao. Visual knowledge transfer among multiple cameras for people counting with occlusion handling. In *ACM-MM*, 2012. 3

[36] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004. 4

[37] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *UAI*, pages 733–742, 2010. 8

[38] Z. Zhang, H. Zha, and M. Zhang. Spectral methods for semi-supervised manifold learning. In *CVPR*, pages 1–6, 2008. 2

[39] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *TPAMI*, 30(7):1198–1211, 2008. 2