Crowd Counting With Limited Labeling Through Submodular Frame Selection

Qi Zhou, Junping Zhang, *Member, IEEE*, Lingfu Che, Hongming Shan, and James Z. Wang

***Abstract*— Automated crowd counting is valuable for intelligent transportation systems, as it can help to improve the emergency planning and prevent congestion in transit hubs such as train stations and airports. Semi-supervised crowd counting aims to estimate the number of pedestrians in an ongoing scene using a combination of a small number of labeled frames and a large number of unlabeled ones. However, existing methods do not incorporate ways to effectively select informative frames as labeled training samples, resulting in low accuracy on unseen crowd scenes. We propose a submodular method to select the most informative frames from the image sequences of crowds. Specifically, the method selects the most representative images to guarantee the information coverage, by maximizing the similarities between the group of selected images and the image sequence. In addition, these frames are chosen to avoid redundancies and preserve diversity. Finally, our semi-supervised method incorporates graph Laplacian regularization and spatiotemporal constraints. Extensive experiments on three benchmark data sets demonstrate that our proposed approach achieves higher accuracy compared with the state-of-the-art regression methods and competitive performance with deep convolutional models, especially when the number of labeled data is exceptionally small.**

***Index Terms*— Intelligent transportation hubs, crowd counting, submodular subset selection, semi-supervised learning.**

I. Introduction

COUNTING the number of pedestrians in images and videos has broad applications in intelligent transportation systems [1], [2]. For instance, in a public event, the event organizers or the police often need to closely monitor the number of people using the public transportation or showing up at a public location to prevent stampede accidents. Businesses count the crowd in shopping areas to estimate the potential revenue. Example crowd scenes captured from video surveillance and included in benchmark datasets, the UCSD [3], the Fudan [4] and the Mall datasets [5], are shown in Fig. 1.

Fig. 1. Left to right: Example crowd scenes in the UCSD [3], the Fudan [4] and the Mall datasets [5], respectively.

Roughly speaking, crowd counting techniques can be categorized into three types: detection-based counting [6]–[11], regression-based counting [2]–[5], [12]–[22], and deep learning based density estimation [23]–[29]. Detection-based methods scan every single person in the images via appearance or motion detection [6]–[11]. The performance of such methods degrades in complex environments, such as scenes with pedestrian occlusion or high density. Feature-based regression methods, on the other hand, aim at learning a mapping relationship between low-level image features and the actual crowd count [2]–[5], [12]–[22]. In genaral, these methods are computationally efficient and able to overcome the aforementioned drawbacks. A critical weakness is that training an accurate regression model requires a large number of labeled crowd images, which are time-consuming to produce. Recent deep learning based methods [23]–[29] first predict a density map of the crowd via Convolutional Neural Networks (CNNs), then the prediction is obtained through integration of the whole predicted density map. These methods are usually trained on large-scale training image datasets to achieve satisfactory performance, because there are millions of parameters in the deep model. However, existing datasets contain at most several thousand frames. Training data shortage is a challenge even when data augmentation and other tricks [23], [24] are employed.

Among the regression-based methods, semi-supervised crowd counting necessitates less human labeling [2], [4], [16], making them more useful for cost- or time-sensititve applications. The application can be developed and deployed more quickly to an ongoing situation. Nevertheless, regression-based crowd counting has two key issues yet to be addressed.

Manuscript received September 7, 2017; revised March 8, 2018; accepted April 7, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61673118 and in part by the Science and Technology Commission of Shanghai Municipality under Grant 16PJD009. The work of J. Z. Wang was supported by The Pennsylvania State University. The Associate Editor for this paper was Z. Duric. *(Corresponding author: Junping Zhang.)*

Q. Zhou, J. Zhang, and L. Che are with the Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: qizhou15@fudan.edu.cn; jpzhang@fudan.edu.cn; lfche16@fudan.edu.cn).

H. Shan is with the Center for Biotechnology and Interdisciplinary Studies, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: shanh@rpi.edu).

J. Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: jwang@ist.psu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2829987

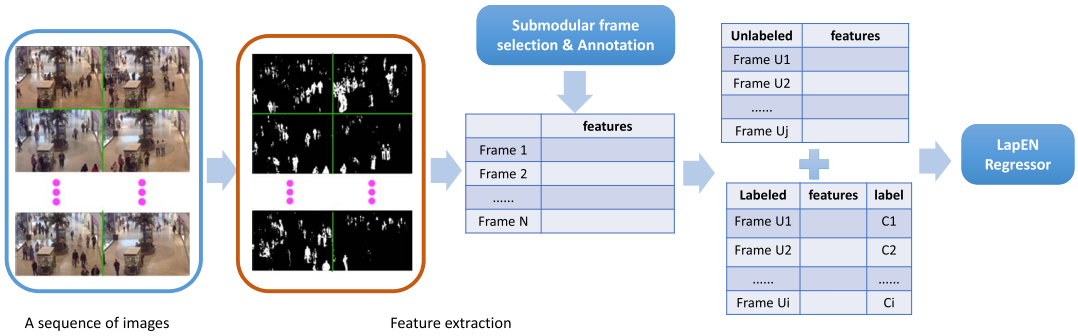


Fig. 2. The flow of the proposed crowd counting method.

1) The performance is highly sensitive to the quality of the set of training images. Through exploiting the most informative crowd images for manual annotation, we attempt to address this issue.
2) The methods emphasize the utilization of limited labeled images, while there can be abundant redundancies within the training set itself. For example, adjacent or nearby frames within a video sequence often have near-identical features as well as counts. Removal of redundancies can ensure that limited human labor or computational resources are devoted to only necessary tasks.

Because the crowd monitoring videos contain high similarity and redundancy in short periods, we propose a submodular strategy to select the most informative images as the training set. Intuitively, the most informative images should be the best representatives of the input image sequence. That is, the selected images, as a group, should be closest visually to the whole image sequence. Certain images, such as occasional severe occlusions in pedestrians and high-speed passage of bicycles, cannot represent the whole sequence and thus should not be selected. Further, preserving diversity is an important criterion for selecting the informative subset of frames. For instance, the selected image subset should include both sparse and dense crowd scenes. The selection of adjacent frames should be avoided because of their high similarity.

Consequently, the most informative subset of frames should be both *representative* and *diverse*. We propose two submodular functions to measure the representativeness and the diversity so that the original crowd image selection issue is transformed into a submodular maximization problem [30]. Finally, these selected images are annotated for the crowd counting task. Besides the limited labeled images, the abundant remaining unlabeled images are utilized to learn a semi-supervised model, which exploits both spatial and temporal structure of the crowd images. The pipeline of our framework is illustrated in Fig. 2.

The *contributions* of this work are as follows. First, this work introduces the idea of employing submodularity to crowd counting. We propose a submodular objective so that the image selection task is formulated as a submodular maximization problem. Second, our proposed semi-supervised method effectively incorporates the spatial and temporal regularizations into the elastic net [31] formulation, and further improves the performance by utilizing unlabeled images. Third, we effectively integrate the submodular frame selection method with semi-supervised regression, so that the model can achieve high prediction accuracy with small-sized labeled frame datasets. We also compare our proposed method with the state-of-the-art deep CNN models, and demonstrate the competitiveness of our method when exceptionally limited training frames are available. Moreover, the submodular frame selection method is not designed for a specific regression method. Experiments have shown that it can be integrated with other regression methods and has the potential to be incorporated into other applications in intelligent transportation systems.

The remainder of this paper is organized as follows. Section II surveys related work of crowd counting and submodular learning. In Section III we introduce our proposed submodular maximization method for crowd image selection. Section IV presents the semi-supervised crowd counting algorithm. We report and analyze the experimental results in Section V and conclude our work in Section VI.

II. Related Work

We survey the development of crowd counting by regression, crowd images selection and submodularity techniques.

Generally, most regression-based crowd counting methods consist of three steps: 1) extracting foreground from background in the region of interest (ROI); 2) extracting features of the foreground pixels, *e.g.* number of pixels, shape, edge and texture; and 3) estimating the crowd count or density by a regression function [2]–[5], [12], [13], [16]–[19]. There are several state-of-the-art methods, including Gaussian process regression (GPR) [3], [17], ridge regression (RR) [5], [13], [16], [21], elastic net (EN) regression [2], [4], support vector regressor (SVR) [14], and Bayesian regression [12], [18], [19]. Although these methods have achieved promising accuracy, most require a large number of labeled images to train a regression model. By penalizing label change among adjacent frames, Tan *et al.* [4] introduced semi-supervised elastic net regression so that only a small-sized labeled dataset is required. However, this method relies on the assumption that the images are sampled at a high frame rate. Loy *et al.* [16] introduced semi-supervised ridge regression by utilizing both spatial and temporal regularization. This method is effective, but ignores the impact of feature dimensions on performance with different sample sizes. Xia *et al.* [2] proposed an elastic net based semi-supervised method by incorporating

spatial, temporal, and activity consistencies. However, this method depends on the effectiveness of extracted local features, *i.e*. detailed spatial and temporal information of each blob or subgroup of pedestrians. As a result, it involves more human annotations.

When only a small number of labeled crowd images are available, the quality of the images is crucial to the prediction performance. However, limited prior research has addressed the problem of crowd images selection. Tan *et al*. [4] used k-means to perform pre-clustering, then randomly selected samples from each cluster. This method could select diverse frames to avoid redundancy to some extent, but it is incapable of finding the most informative images in each cluster and avoiding the outliers because of its intrinsic Gaussian assumption. Loy *et al*. [16] proposed an active learning style technique, which selects informative points (m-landmark) through clustering in the crowd marginal distribution structure, followed by estimating the most informative points as the cluster centers. However, this method does not adequately explore the representativeness and diversity of the selected data, and the number of clusters are fixed to the number of labeled data, which may risk selecting the outliers.

Selecting optimal subset via submodular optimization methods have received increasing attention, which is widely investigated and appied in many domains, including intelligent transportation, data sampling, speech recognition, and sensor placement [32]–[36]. Submodular optimization methods give near-optimal solution to challenging combinatorial optimization tasks, which are often NP-complete. These methods first formulate the task of subset selection as submodular functions which exploiting sub-modularity of information-theoretic measures such as mutual information and entropy. Then these functions can be optimized to find the optimal subsets. In particular, del Arco *et al*. [32] found sparse selection of wireless sensor in traffic dynamics reconstruction. Ranieri *et al*. [35] and Krause *et al*. [36] investigated near-optimal sensor placement via submodular analysis. Wei *et al*. [33] studied the connection of submodularity with the likelihood functions of Naïve Bayes and Nearest Neighbor classifiers, then optimized the submodular functions to find the most informative subset for these classifiers. Shinohara [34] modeled the utility of subset for training speech recognition system through submodular functions. None of these submodular methods have been employed for crowd counting.

III. Submodularity and Frame Selection

We now describe our method for extracting the most informative crowd images to improve counting accuracy.

Given a crowd image sequence $\mathcal{V} = \{\boldsymbol{x}_1, \boldsymbol{x}_2..., \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i$ contains low-level features of an image. Formally, our target is to discover the optimal subset of images $\mathcal{S}$ from $\mathcal{V}$ with the constraint $|\mathcal{S}| = T$, where T denotes the annotation budget (*i.e*. the number of labeled images). We formulate the task of frame selection as a submodular maximization problem. Then the proposed semi-supervised regression algorithm in Section IV is applied for the learning process. The submodular-based frame selection method is useful for building a practical crowd counting system because it can help select a collection of informative frames from a large volume of videos to train the prediction model.

Submodularity: Suppose we are given a set of arbitrary objects $\mathcal{V}$, and a function $F : 2^{\mathcal{V}} \to \mathbb{R}$ returns a real value for any subset $\mathcal{S} \subseteq \mathcal{V}$. F is *submodular* if it satisfies:

$$F(\mathcal{A} \cup \{p\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{p\}) - F(\mathcal{B}), \quad \forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}, \quad p \in \mathcal{V} \setminus B. \quad (1)$$

If this is satisfied everywhere with equality, F is called *modular* function. This property is called as *diminishing returns* [30], stating that the marginal gain of adding an item p to a set $\mathcal{A}$ is greater than to its superset $\mathcal{B}$.

Submodular functions have several important properties: 1) given a set of submodular functions $\{F_1, F_2, ...\}$, their non-negative weighted sum is also a submodular [30], *i.e*. $F = \sum_i \alpha_i F_i$, $\alpha_i \geq 0$ is a submodular; and 2) if f is non-decreasing concave and F is a non-decreasing submodular, $F'(\mathcal{S}) = f(F(\mathcal{S}))$ is a non-decreasing submodular [33]. In the following, we define several submodular terms, each of which captures the quality of subset $\mathcal{S}$ from a specific aspect. Then we can naturally combine all of these terms and keep our objective submodular.

A. Unsupervised Data Subset Selection

Before defining our submodular functions, we first consider an unsupervised data subset selection problem. Given a k-nearest neighbor graph G, the i-th node corresponds to a crowd image $\boldsymbol{x}_i$. The weighted adjacency matrix of graph G is the normalized similarity matrix $W = (w_{ij})_{i,j=1,...,n}$, and each node $\boldsymbol{x}_i$ only connects its k-nearest neighborhood set $\mathcal{N}(\boldsymbol{x}_i)$ by the nonnegative weight w_{ij}, which reflects the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The graph G is then cut into K groups $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_K$ via spectral clustering [37]. Each node $\boldsymbol{x}_i$ is given a scalar value g_i which indicates the cluster index, *i.e*. $g_i = 1, 2...K$, thus these nodes in the same cluster should have the same index. Finding the best subset $\mathcal{S}$ to represent the whole image set $\mathcal{V}$ is equivalent to optimizing the following *data log-likelihood function* [33] (for the sake of simplicity, we use i to represent the sample $\boldsymbol{x}_i$ in $\mathcal{V}$):

$$\begin{aligned} \ell(\mathcal{S}) &= \sum_{i \in \mathcal{V}} \log p(\boldsymbol{x}_i, g_i) \\ &= \sum_{i \in \mathcal{V}} \log p(\boldsymbol{x}_i | g_i) + \sum_{i \in \mathcal{V}} \log p(g_i), \end{aligned} \quad (2)$$

where $p(\boldsymbol{x}_i, g_i)$ is the likelihood of sample $\boldsymbol{x}_i$, and $p(\boldsymbol{x}_i|g_i)$ and $p(g_i)$ are generative likelihood and prior likelihood of sample $\boldsymbol{x}_i$, respectively. The notation *data log-likelihood function* $\ell : 2^{\mathcal{V}} \to \mathbb{R}$ was first introduced in [33], which maps each subset $\mathcal{S}$ of training set to a log-likelihood evaluated on the whole data set $\mathcal{V}$. To construct the generative likelihood and prior likelihood, we extend the assumptions discussed in [33] as follows:

Assumption 1: If a sample $\boldsymbol{x}_i$ belongs to the cluster $\mathcal{G}_k$, *i.e*. $g_i = k$, then the prior probability is estimated as $p(g_i = k) = \frac{m_k(\mathcal{S})}{|\mathcal{S}|}$, where $m_k(\mathcal{S})$ counts the number of samples in both subset $\mathcal{S}$ and cluster $\mathcal{G}_k$, *i.e*. $m_k(\mathcal{S}) = |\mathcal{S} \cap \mathcal{G}_k|$, $|\mathcal{S}| =$

$m_1(\mathcal{S}) + ... + m_K(\mathcal{S})$, and obviously each $m_k(\mathcal{S})$ is a modular set function.

Assumption 2: The generative likelihood $p(\boldsymbol{x}_i|g_i = k)$ is determined only by the closest sample v in the same cluster $\mathcal{G}_k$, *i.e.* $v \leftarrow \arg\max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij}$. It is thus expressed as $p(\boldsymbol{x}_i|g_i = k) = c \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij}$, where c is a constant.

Under these two assumptions, we express $\ell(\mathcal{S})$ in (2) as

$$\begin{aligned}\ell(\mathcal{S}) &= \sum_{i\in\mathcal{V}} \log p(\boldsymbol{x}_i|g_i) + \sum_{i\in\mathcal{V}} \log p(g_i) \\ &= \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log p(\boldsymbol{x}_i|g_i = k) + \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log p(g_i = k) \\ &= \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log c \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij} + \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log \frac{m_k(\mathcal{S})}{|\mathcal{S}|} \\ &= \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij} + \sum_{k=1}^{K} |\mathcal{G}_k| \log m_k(\mathcal{S}) \\ &\quad - |\mathcal{V}| \log |\mathcal{S}| + C, \end{aligned} \tag{3}$$

where $C = \sum_{i\in\mathcal{V}} \log c$ is a constant. Because $|\mathcal{V}|$ is independent of $\mathcal{S}$, given the equality constraint $|\mathcal{S}| = T$, the third term $|\mathcal{V}| \log |\mathcal{S}|$ in $\ell(\mathcal{S})$ then becomes a constant. Therefore, the problem of data log-likelihood function maximization can be reformulated as a constraint optimization problem as follows:

$$\begin{aligned}\max_{\mathcal{S}:|\mathcal{S}|=T} \ell(\mathcal{S}) = \max_{\mathcal{S}:|\mathcal{S}|=T} &\sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \log \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij} \\ &+ \sum_{k=1}^{K} |\mathcal{G}_k| \log m_k(\mathcal{S}). \end{aligned} \tag{4}$$

Intuitively, the first term measures the similarity or relevance of subset $\mathcal{S}$ to the whole set $\mathcal{V}$, and the second term is correlated to the number of samples in each cluster. Next, we propose our submodular functions, such that the objective in (4) is equivalent to a submodular maximization problem.

B. Representativeness Term

As stated, the first term measures the similarity of the subset frames $\mathcal{S}$ to the whole set $\mathcal{V}$. A representative subset of crowd images should be most similar to the whole set, in order to preserve as much information as possible. Furthermore, selecting the most representative images could help eliminate the outliers, because they are less similar to the whole image set. Therefore, we propose a *localized facility location function*, f_{lfac}, to measure the representativeness of subset $\mathcal{S}$ to the whole set $\mathcal{V}$:

$$f_{\text{lfac}}(\mathcal{S}) = \sum_{k=1}^{K}\sum_{i\in\mathcal{G}_k} \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij}, \tag{5}$$

where $\sum_{i\in\mathcal{G}_k} \max_{j\in\mathcal{S}\cap\mathcal{G}_k} w_{ij}$ has a similar form as the facility location function [38], which measures the similarity between $\mathcal{S}$ and the cluster $\mathcal{G}_k$. Because it is monotonic submodular [38], by property 1, the representative term is also submodular. This function is similar to the first term in (4) derived from the data log-likelihood function. Assuming the second term is a constant, optimizing the objective in (4) is equivalent to maximize the representativeness term.

For our crowd image selection task, these images with similar features should connect with large weights in the graph, *i.e.* in a same cluster. Function f_{lfac} selects the most representative samples in each cluster, and it is maximized only if all the similarities between $\mathcal{S}$ and clusters $\mathcal{G}_1, ..., \mathcal{G}_k$ are maximized. Therefore, the most representative images in each cluster can be selected via f_{lfac} optimization.

C. Diversity Term

An ideal subset of frames for crowd counting should contain both sparse and dense crowd density. Therefore, selecting frames from a short range of image sequence or single cluster is unfavorable for learning a robust regression model, as a small region of intensive samples where many temporally adjacent frames may be redundant cannot well represent the distribution of the whole image set. To formulate this intuition, we adopt the commonly used *diversity reward function* [33] to measure the diversity in the crowd image sequences:

$$f_{\text{rew}}(\mathcal{S}) = \sum_{k=1}^{K} \sqrt{r(\mathcal{S}\cap\mathcal{G}_k)}, \tag{6}$$

where the function $r(\cdot) \geq 0$ indicates the rewards of selected frames in a cluster. The sum of square root scores the quantity of total rewards given the solution $\mathcal{S}$, which could effectively find solutions that are more uniformly distributed over the clusters.

Suppose a collection of frames are available, a direct assumption is that the quantity of total rewards is correlated to the number of frames. We thus define the reward function as a simple formula

$$r(\mathcal{S}\cap\mathcal{G}_k) = \frac{|\mathcal{S}\cap\mathcal{G}_k|}{|\mathcal{G}_k|} = \frac{m_k(\mathcal{S})}{|\mathcal{G}_k|}, \tag{7}$$

which indicates the proportion of selected frames in cluster $\mathcal{G}_k$. This function is simple, effective, and modular ($|\mathcal{G}_k|$ is independent of $\mathcal{S}$, thus constant). Furthermore, this function brings another advantage. For any optimal solution $\mathcal{S}^*$, its reward $f_{\text{rew}}(\mathcal{S}^*)$ requires the uniformity of rewards $r(\mathcal{S}\cap\mathcal{G}_1), ...r(\mathcal{S}\cap\mathcal{G}_K)$, such that the proportions of selected frames in each cluster are roughly equal. That is, $\frac{m_1(\mathcal{S})}{|\mathcal{G}_1|} \approx \frac{m_2(\mathcal{S})}{|\mathcal{G}_2|} \approx ... \approx \frac{m_k(\mathcal{S})}{|\mathcal{G}_K|}$, then all $m_k(\mathcal{S}), k = 1, 2...K$, are forced to be constant. As a result, the second term in (4) is reduced to a constant. Optimizing the likelihood function in (2) is then equivalent to finding the solution $\mathcal{S}$ such that the local facility location function f_{lfac} and diversity reward function f_{rew} are maximized.

D. Optimization

By now the frame selection problem is transformed into the optimization of submodular functions, including the representativeness term f_{lfac} and the diversity term f_{rew}. We incorporate these two submodular terms into a unified objective

function:

$$\mathcal{S}^* = \underset{\mathcal{S}:|\mathcal{S}|=T}{\arg\max}(f_{\text{lfac}}(\mathcal{S}),\, f_{\text{rew}}(\mathcal{S})) \tag{8}$$

Direct maximization of f_{lfac} or f_{rew}, however, is an NP-hard problem. Fortunately, submodular functions could be efficiently solved via the greedy algorithm, which gives a $(1-1/e)$ approximation to the optimal solution [38]. We propose a two-step greedy algorithm to optimize our submodular objective functions, including f_{lfac} optimization stage and f_{rew} optimization stage in each iteration. First, the localized facility location function is optimized, which could be realized by performing facility location function optimization in each cluster $\mathcal{G}_k$. Each cluster maintains a solution $\mathcal{A}_k$ and the most representative frame a_k^*, then a candidate set $\mathcal{R}$ is obtained which contains frame a_k^* of each cluster, *i.e.* $|\mathcal{R}|$ is a constant of K. Next, the diversity reward function is optimized by scoring each frame from $\mathcal{R}$, and a frame $p^* \in \mathcal{R}$ that leads to the largest marginal gain is added to the final solution $\mathcal{S}$. Then the next representative frame in cluster $\mathcal{G}_d$ which p^* belongs to is selected and used to update the candidate set. After T iterations, we get a subset $\mathcal{S}^*$ which best represents and uniformly scatters around the whole set $\mathcal{V}$. Finally, these frames in subset $\mathcal{S}^*$ are annotated for regression learning. The pseudocode of our unified algorithm is given in Algorithm 1.

Algorithm 1 Submodular Frame Selection for Crowd Counting

Input: crowd image sequence $\mathcal{V}$, number of clusters K, rounds T
Output: $\mathcal{S}, \boldsymbol{\beta}$
Init: $\mathcal{S} = \mathcal{A}_1 = ... = \mathcal{A}_K = \emptyset$, candidate set $\mathcal{R} = \emptyset$
Dividing the whole set $\mathcal{V}$ into K clusters $\mathcal{G}_1, \mathcal{G}_2 ... \mathcal{G}_k$ via spectral clustering
for each cluster $k = 1, 2...K$ **do**
  $a_k^* \leftarrow \underset{a \in \mathcal{G}_k \setminus \mathcal{A}_k}{\arg\max}\, f_{\text{fac}}(\mathcal{A}_k \cup \{a\}) - f_{\text{fac}}(\mathcal{A}_k)$
  $\mathcal{A}_k \leftarrow \{a_k^*\}$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{a_k^*\}$
end for
for $t = 1, 2...T$ **do**
  $p^* \leftarrow \underset{p \in \mathcal{R}}{\arg\max}\, f_{\text{rew}}(\mathcal{S} \cup \{p\}) - f_{\text{rew}}(\mathcal{S})$
  $d \leftarrow$ the index of cluster which p^* belongs to
  $\mathcal{S} \leftarrow \{\mathcal{S} \cup p^*\}$
  $\mathcal{R} \leftarrow \mathcal{R} \setminus \{p^*\}$
  $a_d^* \leftarrow \underset{a \in \mathcal{G}_d \setminus \mathcal{A}_d}{\arg\max}\, f_{\text{fac}}(\mathcal{A}_d \cup \{a\}) - f_{\text{fac}}(\mathcal{A}_d)$
  $\mathcal{A}_d \leftarrow \mathcal{A}_d \cup \{a_d^*\}$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{a_d^*\}$
end for
Annotating frame set $\mathcal{S}$, then $\mathcal{V}$ is divided into labeled training set $\mathcal{L}$ and unlabeled training set $\mathcal{U}$
Performing semi-supervised regression on training set.

E. Discussion

The proposed representativeness term in (5) is similar to the existing *facility location function* f_{fac} [38] and *Nearest Neighbor submodular function* f_{NN} [33] in (9). Here we analyze the relations and differences among these three functions.

$$\begin{aligned} f_{\text{fac}}(\mathcal{S}) &= \sum_{i \in \mathcal{V}} \max_{j \in \mathcal{S}} w_{ij}, \\ f_{\text{NN}}(\mathcal{S}) &= \sum_{y \in \mathcal{Y}} \sum_{i \in \mathcal{V}^y} \max_{j \in \mathcal{S} \cap \mathcal{V}^y} w_{ij}. \end{aligned} \tag{9}$$

Facility location function f_{fac} is often applied to identify representative instances from a collection of items. By comparing (5) and (9), it can be seen that function f_{fac} views $\mathcal{V}$ as only one cluster, whereas f_{lfac} regards $\mathcal{V}$ as multiple clusters. f_{fac} is thus a special case of f_{lfac}.

Nearest Neighbor submodular function f_{NN} is designed for subset selection in k-Nearest Neighbor classifier conditions [33]. f_{lfac} can be viewed as an equivalent of f_{NN} regardless of the conditions. However, f_{NN} is only applied to classification tasks, and it categorizes each class of instances as a cluster and maximizes the f_{fac} value for each class. Moreover, to divide the whole set $\mathcal{V}$ into $|\mathcal{Y}|$ partitions, this function needs a labeled training set to estimate the posterior probability of each sample. Unlike this function, we optimize the localized facility location function f_{lfac} to estimate $\mathcal{S}$ in an unsupervised manner. Furthermore, it can be smoothly employed in both classification and regression tasks.

IV. Semi-Supervised Regression for Crowd Counting

In this section, we focus on semi-supervised crowd counting because it requires less labeled images, which is consistent with the goal of our proposed submodular strategy. Once the most representative and diverse frames are selected, we can use these frames to train a regression model. For the task of crowd counting by regression, the purpose is to find a solution $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ of the linear model $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$, such that the squared loss is minimized. However, when the number of labeled frames is small or the dimension of low-level feature $\boldsymbol{x}$ is too large, it is difficult to learn a good mapping. To address these problems, the elastic net regularized regression [31] was employed in many tasks and it has shown promising performance in the recent literature [2], [4], [39], [40]. It is thus chosen as the base model of our method.

Our goal is to improve the counting accuracy via performing semi-supervised regression, by fully utilizing the abundant unlabeled frames. Assume we have a set of annotated frames $\mathcal{L} = (\mathbf{X}_\mathcal{L}, \mathbf{y}_\mathcal{L}) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{X}_\mathcal{L}$ is a $n \times p$ matrix where each row $\boldsymbol{x}_i$ is a p-dimensional vector of low-level features, $\mathbf{y}_\mathcal{L}$ is a vector of ground truth, and each entry y_i is the corresponding label of $\boldsymbol{x}_i$. Besides the labeled frames $\mathcal{L}$, a large number of unlabeled images are available $\mathcal{U} = (\mathbf{X}_\mathcal{U}) = \{(\boldsymbol{x}_i)\}_{i=1}^m$, and all $m+n$ images $\mathbf{X} = [\mathbf{X}_\mathcal{L}{}^T, \mathbf{X}_\mathcal{U}{}^T]^T$ are obtained from a collection of image sequences in a video. To formulate our semi-supervised learning objective function, we introduce an elastic net regression setting with the following regularization:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\arg\min} \|\mathbf{y}_\mathcal{L} - \mathbf{X}_\mathcal{L}\boldsymbol{\beta}\|_2^2 + \lambda_I \|f\|_I^2 + \lambda_A P_\alpha(\boldsymbol{\beta}), \tag{10}$$

where $P_\alpha(\boldsymbol{\beta})$ is the elastic net penalty [31], which linearly combines ℓ_1-norm $\|\boldsymbol{\beta}\|_1$ and ℓ_2-norm $\|\boldsymbol{\beta}\|_2^2$ of the lasso and ridge methods:

$$P_\alpha(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2, \tag{11}$$

and the coefficient $\alpha \in (0,1]$ determines the influence of the ℓ_1 penalty relative to the ℓ_2 penalty. Here, $\|f\|_I^2$ is a regularization term which reflects the intrinsic structure and temporal smoothness of data points, and λ_I is a coefficient of the regularization.

To learn the intrinsic distribution of training data, we adopt the widely used graph Laplacian regularization [41]. Fully characterizing the manifold structure of data points, graph Laplacian encourages the smoothness of target function f with respect to the distribution of both labeled and unlabeled data. In order to compute the graph Laplacian, a directed graph G is first constructed. For each node $\boldsymbol{x}_i$, it only connects with the points in its nearest neighborhoods. Then we construct an affinity matrix $\mathbf{W} \in \mathbb{R}^{(m+n)\times(m+n)}$ defined by $w_{ij} = (1-\epsilon)w_{ij}^s + \epsilon w_{ij}^t$, where w_{ij}^s captures the similarity in feature space and w_{ij}^t represents temporal similarity, and ϵ is the proportion of temporal similarity relative to spatial similarity. Here, we use the Gaussian kernel function with constant bandwidth to define these similarities:

$$\begin{cases} w_{ij}^s = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) & \boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i) \\ w_{ij}^s = 0 & \text{otherwise} \\ w_{ij}^t = \exp(-\|t_i - t_j\|^2) & t_j \in [t_i - b, t_i + b] \\ w_{ij}^t = 0 & \text{otherwise,} \end{cases} \tag{12}$$

where $\mathcal{N}_k(\boldsymbol{x}_i)$ denotes k nearest neighbors of $\boldsymbol{x}_i$ in feature space, $t \in \{1, 2, ...\}$ is the image index, and b is equivalent to the number of nearest neighbors in temporal sequence. The Laplacian matrix $\mathbf{L}$ is then computed as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j^{m+n} w_{ij}$. The regularization is then expressed as

$$\|f\|_I^2 = \mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{i,j} w_{ij}\|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\|_2^2 = 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\boldsymbol{\beta}. \tag{13}$$

This regularization is reduced to spatial regularization when $\epsilon = 0$, or temporal regularization when $\epsilon = 1$. To simplify the following computation, we assume the number of temporal and spatial neighbors are equal, *i.e.* $k = 2b$. Then, (13) can be incorporated into (10) as

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_A P_\alpha(\boldsymbol{\beta}), \tag{14}$$

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_\mathcal{L} \\ \sqrt{2\lambda_I}\mathbf{L}^{1/2}\mathbf{X} \end{bmatrix}, \quad \tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_\mathcal{L} \\ \mathbf{0} \end{bmatrix}. \tag{15}$$

Because the affinity matrix $\mathbf{W}$ may be assymmetrical, we first formulate $\mathbf{W}^* = (\mathbf{W} + \mathbf{W}^T)/2$, so that the Laplacian matrix $\mathbf{L}$ is positive semidefinite. (14) has the same form as an elastic net objective function, which can be solved efficiently via the least angle regression (LARS) algorithm [42].

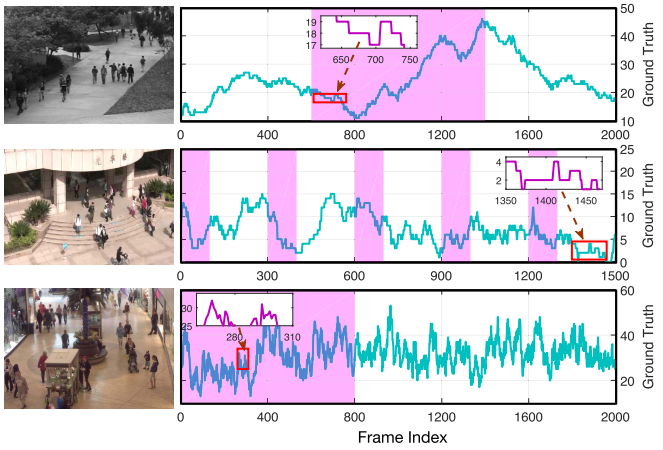


Fig. 3. Example data and the corresponding number of pedestrians in each frame. Those partitions with pink background were selected as training set, and the rest testing. The three are from the UCSD, the Fudan and the Mall pedestrian datasets, respectively. As shown in the mini line graphs, frames from 661 to 690 of the UCSD dataset and frames 1,374 to 1,408 of the Fudan dataset have identical ground truth.

TABLE I

DATASET DETAILS. **R** = RESOLUTION, **FPS** = FRAME PER SECOND, **S** = CROWD SIZE, $\mathbf{N_{tr}}$ / $\mathbf{N_{te}}$ = NUMBER OF TRAINING / TESTING FRAMES, $\mathbf{N_P}$ = NUMBER OF ADJACENT FRAME PAIRS WITH SAME CROWD SIZE.

Dataset	R	fps	S	N_{tr} / N_{te}	N_p
UCSD [3]	158×238	10	11 - 46	800/1200	1684
Fudan [4]	240×320	10	0 - 15	500/1500	1175
Mall [5]	480×640	<2	13 - 53	800/1200	389

V. EXPERIMENTS

We compare our proposed submodular frame selection for semi-supervised crowd counting in three benchmark datasets. We also perform a comprehensive comparison with other state-of-the-art crowd counting algorithms. Finally, we study the influence of parameters in our proposed algorithm to the accuracy of crowd counting.

A. Experimental Setting

Three benchmark datasets were used – the UCSD [3], the Fudan [4], and the Mall dataset [5]. Details of them are shown in Table I. Fig. 3 shows three example images and the ground truth of each frame index. The UCSD and the Fudan datasets were sampled at a high frame rate of 10 frames-per-second (fps). We observe that they have more than 1,000 pairs of adjacent frames with identical crowd size. It means that these two datasets are highly redundant. For example, frames from 661 to 690 of the UCSD dataset and frames 1,374 to 1,408 of the Fudan dataset have identical ground truth, as pedestrians in these frames simply walked through the scene without moving into or out of the region of interest (ROI). Meanwhile, the Mall dataset is sampled at a relative low fps, thus has less redundancy. Furthermore, the Mall dataset is more challenging as the illumination condition changes greatly and inter-object occlusion is more severe [16]. Each dataset is split into training and testing sets without overlapping similar

TABLE II
DESCRIPTION OF EACH LOW LEVEL FEATURE

	Feature	Dimension	Description
Segment	Area	1	total number of pixels in the segment
	Perimeter	1	total number of pixels on the perimeter
	Perimeter orientation	6	orientation histogram of the perimeter, orientations: 0°, 30°, 60°, 90°, 120°, 150°
	Perimeter-area ratio	1	ratio between the segment perimeter and area
Edge	Edge	1	total number of edge pixels
	Edge orientation	6	histogram of the edge orientations in the segment, orientations: 0°, 30°, 60°, 90°, 120°, 150°
	Minkowski dimension	1	the Minkowski fractal dimension of the edges
Texure	Homogeneity	4	a smoothness measure, orientations: 0°, 45°, 90°, 135°
	Energy	4	total sum-squared energy, orientations: 0°, 45°, 90°, 135°
	Entropy	4	a randomness measure, orientations: 0°, 45°, 90°, 135°

to other published work [3], [4], [16]. For the UCSD and the Mall datasets, 800 of the 2,000 frames were selected to form the training set, and the rest frames for testing. The Fudan dataset contains five discontinuous image sequences. The first 100 frames of each sequence were selected to form the training set. The details and the partition of three crowd datasets are illustrated in Fig. 3.

For each dataset, 29 features are extracted from the ROI of each image, including area, perimeter, edge, texture features, and Minkowski dimension [3]. The description of these features are shown in Table II. All features are perspective normalized to compensate for the perspective distortion [3], [12]. All compared methods adopt the same features.

Our proposed semi-supervised regression algorithm includes a few free parameters and regularization parameters. The number of neighbors for the graph Laplacian is fixed to 10. The proportion of ℓ_1 penalty α is automatically selected by LARS algorithm [42]. For the remaining parameters, λ_A and λ_I, we employ a *grid search* and the optimal estimations are selected via *10-fold* cross validation on the training set.

The error rate is measured by the *mean squared error* (MSE), defined by:

$$R = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{16}$$

where N is the number of test frames, $\hat{y}_i$ is the count prediction, and y_i is the ground truth. All results are averaged over 50 trials.

B. Comparison Among Different Regression Methods

We evaluate the accuracy of various crowd counting methods including blind and selective counting methods, based on whether the labeled training sets are selected by some criteria. For the first class of methods, we selected four state-of-the-art methods introduced in [3], [4], [13], [16], and our proposed Laplacian regularized elastic net regression (LapEN). Among these five methods, the Gaussian Process Regression (GPR) and Cumulative Attribute Ridge Regression (CA-RR) are supervised methods, while the Semi-Supervised elastic net (SSEN), Semi-Supervised Regression (SSR), and LapEN train regression models in a semi-supervised manner. For the second class, to our knowledge, there are only two criteria including k-means selection [4] and m-landmark [16]. We compare these methods with our submodular frame selection algorithm.

For the first subcategory of counting methods, all methods use the same training-testing split. A total of 50 frames in the training partition are randomly selected as labeled samples, and the rest samples in training partition (750 in both the UCSD and the Mall datasets, 450 in the Fudan dataset) remain unlabeled. Table III shows that our proposed LapEN achieves the best performance on the UCSD and the Fudan dataset. For the Mall dataset, it obtains the second smallest error and has comparable MSE to the best performance.

For the second class of counting methods, we first randomly extract 700, 400, and 700 samples from the UCSD, the Fudan, and the Mall datasets, respectively. Next, each method selects 50 frames for annotation based on its selection criterion, and the rest samples in *training partition* remain unlabeled (750, 450, and 750 unlabeled frames for these datasets, respectively). This strategy keeps the obtained labeled frames in each trial different. For our submodular frame selection algorithm, we fix the number of clusters, K, to be 5 and the proportion of temporal regularization, ϵ, to be 0.5 for all datasets based on our experience. We further examine the influence of these parameters later. Table III shows that our submodular frame selection algorithm outperforms the other methods significantly, because these methods do not sufficiently exploit both representativeness and diversity for frame selection. Moreover, the submodular frame selection algorithm achieves notable improvements on the low redundant Mall dataset, demonstrating the robustness of the algorithm. Fig 4 shows the predictions made by submodular and LapEN against the ground truth for each dataset.

From Table III, it is evident that (1) our proposed Laplacian regularized elastic net regression shows competitive performance against the existing methods, and (2) the integration of submodular frame selection and LapEN can leverage the performance significantly, and our approach achieves the highest accuracy on the three datasets.

C. Comparison With Density Estimation Method

Recently proposed deep learning based methods have shown promising performance by estimating the density maps of input images [23]–[29], which indicate the crowd density

TABLE III

THE MSE COMPARISON AGAINST STATE-OF-THE-ART COUNTING METHODS AND FRAME SELECTION METHODS. GPR = GAUSSIAN PROCESS REGRESSION, SSEN = SEMI-SUPERVISED ELASTIC NET REGRESSION, SSR = SEMI-SUPERVISED REGRESSION, CA-RR = CUMULATIVE ATTRIBUTE RIDGE REGRESSION, LAPEN = THE PROPOSED LAPLACIAN REGULARIZED ELASTIC NET REGRESSION. FRAME SELECTION METHODS CONTAIN: k-MEANS, m-LANDMARK AND THE PROPOSED SUBMODULAR METHOD. A SMALLER MSE VALUE IS BETTER. N/A: THESE RESULTS ARE NOT AVAILABLE.

Dataset	Without Frame Selection					Frame Selection Methods		
	GPR [3]	SSEN [4]	SSR [16]	CA-RR [13]	LapEN	k-means + SSEN [4]	m-landmark + SSR [16]	submodular + LapEN
UCSD [3]	7.39	9.88	7.33	9.27 (from [16])	**6.98**	6.13	7.06	**4.32**
Fudan [4]	2.02	1.96	N/A	N/A	**1.94**	1.83	N/A	**1.49**
Mall [5]	19.61	**16.37**	18.11	22.19 (from [16])	16.52	13.87	17.85	**9.20**

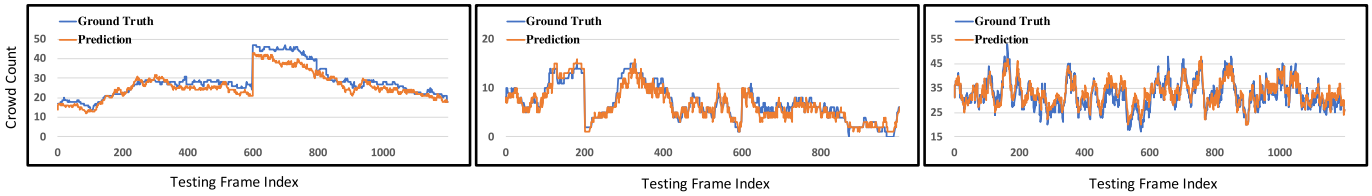


Fig. 4. The ground truth and the estimates predicted by submodular + LapEN. From left to right: the UCSD, the Fudan, and the Mall datasets.

TABLE IV

COMPARISON BETWEEN MCNN AND OUR PROPOSED METHOD. MCNN FAILED ON MALL DATASET WHEN DATA AUGMENTATION TECHNIQUES ARE NOT APPLIED

Method	UCSD	Fudan	Mall
MCNN [24]	4.07	1.87	>20
MCNN + augmentation [24]	**2.26**	**1.32**	9.56
submodular + LapEN	4.32	1.49	**9.20**

TABLE V

PERFORMANCE COMPARISON BETWEEN ELASTIC NET (EN) AND THE PROPOSED LAPEN: WITH SPATIAL REGULARIZATION, TEMPORAL REGULARIZATION, AND BOTH OF THESE REGULARIZATIONS. S: SPATIAL, T: TEMPORAL, S+T: BOTH SPATIAL AND TEMPORAL. IN EACH ELEMENT, MSE PLUS OR MINUS STANDARD DEVIATION IS SHOWN

Method	UCSD	Fudan	Mall
EN	8.43 ± 4.25	2.42 ± 0.79	18.25 ± 6.13
LapEN (S)	**6.91 ± 3.46**	2.05 ± 0.69	16.62 ± 6.26
LapEN (T)	7.07 ± 3.72	**1.89 ± 0.69**	17.09 ± 6.37
LapEN (S + T)	6.98 ± 3.87	1.94 ± 0.71	**16.52 ± 4.78**

in each pixel. The final predictions are obtained through integration of the whole density maps. However, it is nontrivial for these methods to fit the model with only dozens of images. In this subsection, we compare our method with the recently proposed Multi-Column Convolutional Neural Network (MCNN) [24]. By utilizing multi-scale CNNs for crowd counting, MCNN achieved state-of-the-art accuracy. To keep diversity in the training set, 50 frames with equal intervals are selected to train MCNN model as adopted in [2]. For example, indices in {600, 616, 632, ..., 1368, 1384} of the UCSD dataset are training samples. We also test the effect of data augmentation for MCNN, where the horizontal flip and random position crop of each image are added into the training sets. Because MCNN only optimizes pixel-wise loss and the size of the training set is too small, the validation performance is highly unstable in the training process. We report the average MSE of the last 100 iterations.

The comparison results are shown in Table IV. When only 50 images are available, our method attains competitive performance with MCNN. MCNN attains poor performance on the Mall dataset in our trials with different uniform indices, and the MSE on the testing set is difficult to converge throughout the iterations. A possible reason is that this data is much more complex than the UCSD and the Fudan datasets, as aforementioned. And perhaps CNN is much more sensitive to the change of illumination than handcrafted features. Specifically, handcraft features are extracted from foreground of images (crowd segmentation), which is generated from videos by mixture of dynamic texture models [3]. These features are less sensitive to the change of illumination, whereas CNN features suffer from this illumination change. When image augmentation techniques are applied, MCNN achieved lower MSE than merely training on 50 raw images, and outperforms our submodular method on the UCSD and the Fudan datasets.

Although predicting the density maps via deep models attains better performance than regression methods in previous literature, they are considerably more time-consuming in figuring out the optimal parameters. In our experiments, MCNN costs nearly ten hours to fit the model on a single GPU (NVIDIA Titan X, 6GB RAM), and its performance also heavily depends on the number of training samples and data augmentation trick. In contrast, our proposed submodular selection and LapEN regression method requires less human annotations and is fast to train (taking only several minutes on CPUs).

D. The Effects of Spatial and Temporal Regularizations

By design, LapEN allows the regressor to utilize both spatial and temporal label propagation. We conduct an experiment to investigate if the regressor takes advantage of this opportunity.

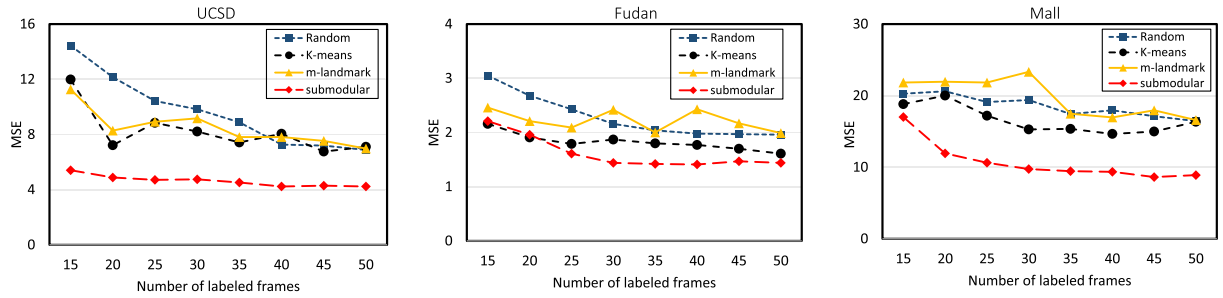


Fig. 5. Counting performance using four different frame selection methods. The symbol 'submodular' is the proposed selection method.

TABLE VI
COMPARISON AMONG DIFFERENT SUBMODULAR TERMS

Method	UCSD	Fudan	Mall
random	6.98 ± 3.87	1.94 ± 0.71	16.52 ± 4.78
f_{fac}	6.73 ± 3.01	1.82 ± 0.85	12.70 ± 2.91
f_{lfac}	6.61 ± 3.12	1.57 ± 0.72	10.51 ± 4.79
f_{rew}	4.98 ± 1.39	1.81 ± 0.59	14.75 ± 5.56
$f_{\text{lfac}} + f_{\text{rew}}$	**4.32 ± 0.87**	**1.49 ± 0.09**	**9.20 ± 2.48**

TABLE VII
COMPARISON AMONG DIFFERENT FRAME SELECTION METHODS AND REGRESSION METHODS

Method	UCSD	Fudan	Mall
SSEN + k-means	6.13 ± 4.98	1.83 ± 0.91	13.87 ± 3.93
SSEN + m-landmark	7.83 ± 3.54	2.08 ± 0.77	**11.15 ± 3.44**
SSEN + submodular	**5.41 ± 2.10**	**1.79 ± 0.83**	13.47 ± 3.25
GPR + k-means	7.44 ± 0.89	1.97 ± 0.23	13.30 ± 2.61
GPR + m-landmark	6.90 ± 1.01	1.91 ± 0.26	**9.18 ± 1.92**
GPR + submodular	**6.25 ± 0.55**	**1.87 ± 0.25**	14.66 ± 3.59

Table V shows the effects of different regularizations of our proposed Laplacian regularized elastic net, and their capability of exploiting unlabeled data distribution and temporal smoothness. It is evident that the performance improved remarkably with the usage of unlabeled images. For instance, with the application of unlabeled data, the MSE in the UCSD, the Fudan, and the Mall datasets are reduced by nearly 17%, 20%, and 9%, respectively, when both spatial and temporal regularization are used. Although different datasets favor different regularizations, the usage of both regularizations yields best or near-optimal performance. It is also more robust than using only spatial or temporal regularization. On average, it achieves the lowest MSE on all three datasets. We conclude that using both yields the best and robust performance.

E. Comparison Among Frame Selection Methods

In this experiment we compare the performance of our proposed submodular frame selection method with k-means landmark selection [4] and m-landmark selection [16], and we also take the random selection as the baseline. We use the LapEN to train the model. The number of labeled data is in {15, 20, 25, 30, 35, 40, 45, 50} and the rest of training data remain unlabeled. Fig. 5 shows that all these methods gain more or less improvements, while our submodular method outperforms its counterparts remarkably. When only 15 labeled data are available, our method performs comparatively or even better than other methods with 50 samples.

F. Effects Of Submodular Terms and Parameters

To examine the effects of the proposed representative term f_{lfac} and diversity term f_{rew}, we compare them with the random selection baseline and the facility location function f_{fac}. As shown in Table VI, the improvements of function f_{fac} against random selection is limited, because it takes the whole data set as only one cluster. Functions f_{lfac} and f_{rew} improve the performance evidently, and the combination of these two submodular terms yields the best performance on all datasets.

We further evaluated our approach when different values of the number of clusters K and the proportion of temporal regularization ϵ are selected. As shown in Fig. 6, for the UCSD and the Mall dataset, the best performance is achieved when $K = 5$. And for the Fudan dataset, the best performance is achieved when $K = 2$, because the crowd size of the Fudan dataset is in a smaller range (0 to 15). It can also be seen that the performance is sensitive to the value of ϵ. If ϵ grows to 1, the performance degrades because only temporal smoothness is considered. On the other hand, if ϵ is set to 0, only spatial regularization does not yield the best performance. In general, the best performance of the three datasets is achieved when $\epsilon = 0.2$, in the range 0.4 to 0.8, and in the range 0.2 to 0.8, respectively. Although the performance is sensitive to ϵ, it is more stable if the appropriate value of K is determined. Moreover, the optimal value of ϵ can be attained via cross validation experimentally, leading to a near-optimal performance.

G. Generalizability

We further examine the generalizability of these frame selection methods. To make the low-level features consistent in frame selection and counting phases, we exclude detection-based and CNN-based methods and only evaluate regression-based methods. Each selection method selects 50 frames for annotation before GPR [3] or SSEN [4] regression model is employed. Table VII shows that the proposed submodular frame selection method still achieves competitive performance on all three datasets, and almost all combinations improve the performance in contrast to GPR and SSEN in Table III.

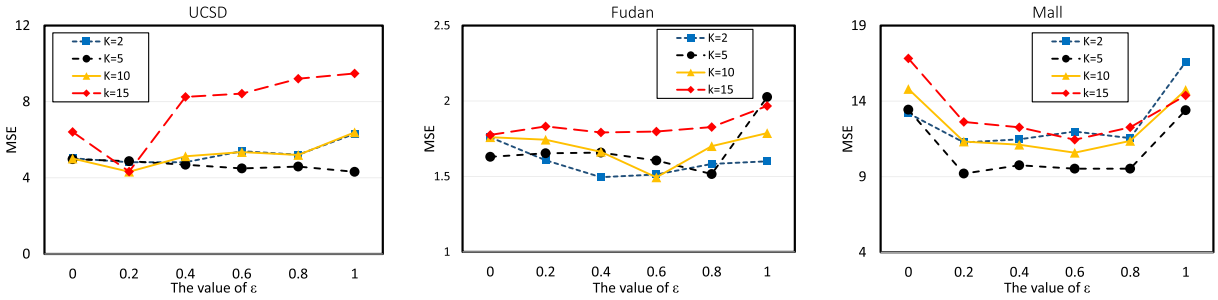


Fig. 6. Effects of parameter selection of K and ϵ. The horizontal axis denotes different values of ϵ, while lines with different colors denote different K values.

VI. Conclusions

We proposed a submodular algorithm to select informative frames from videos in crowd counting task. This method avoids traditional blind and exhaustive annotation by exploiting most representative and diverse images from crowd image sequences. Semi-supervised regression is performed on the selected images and the remaining unlabeled images. Extensive experiments with multiple datasets have demonstrated the effectiveness of the proposed algorithm, and shown the practical application in intelligent transportation systems. Moreover, the proposed submodular method can be integrated with other regression methods, and has the potential to be incorporated into other applications, for intelligent transportation systems and beyond.

Acknowledgment

The authors thank the Associate Editor and the anonymous reviewers for their constructive comments, which helped improve this manuscript.

References

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[2] W. Xia, J. Zhang, and U. Kruger, "Semisupervised pedestrian counting with temporal and spatial consistencies," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1705–1715, Aug. 2015.

[3] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[4] B. Tan, J. Zhang, and L. Wang, "Semi-supervised elastic net for pedestrian counting," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2297–2304, 2011.

[5] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 21.1–21.11.

[6] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[8] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 899–906.

[9] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3198–3205.

[10] M. Wang, W. Li, and X. Wang, "Transferring a generic pedestrian detector towards specific scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3274–3281.

[11] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2913–2920.

[12] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and Bayesian regression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2160–2177, Apr. 2012.

[13] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[14] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, "A method for counting moving people in video surveillance videos," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 231–240, 2010.

[15] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling Simulation and Visual Analysis of Crowds*, Springer, 2013, pp. 347–382.

[16] C. C. Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2256–2263.

[17] J. Zhang, B. Tan, F. Sha, and L. He, "Predicting pedestrian counts in crowded scenes with rich and high-dimensional features," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1037–1046, Dec. 2011.

[18] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 545–551.

[19] B. Liu and N. Vasconcelos, "Bayesian model adaptation for crowd counts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4175–4183.

[20] K. Chen and Z. Zhang, "Pedestrian counting with back-propagated information and target drift remedy," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 639–647, Apr. 2017.

[21] K. Chen and J.-K. Kämäräinen, "Pedestrian density analysis in public scenes with spatiotemporal tensor features," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1968–1977, Jul. 2016.

[22] N. Tang, Y.-Y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Cross-camera knowledge transfer for multiview people counting," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 80–93, Jan. 2015.

[23] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 833–841.

[24] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 589–597.

[25] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 615–629.

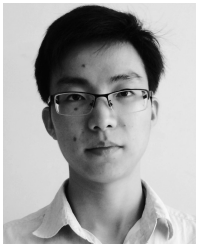
[26] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 660–676.

[27] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, vol. 1. no. 3, pp. 4031–4039.

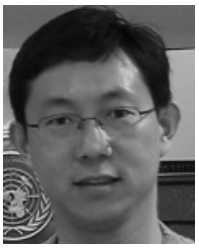
[28] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. ACM Multimedia Conf.*, 2016, pp. 640–644.

[29] J. Li, H. Yang, L. Chen, J. Li, and C. Zhi, "An end-to-end generative adversarial network for crowd counting under complicated scenes," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Jun. 2017, pp. 1–4.

[30] A. Krause and D. Golovin, "Submodular function maximization," *Tractability, Pract. Approaches Hard Problems*, vol. 3, no. 19, pp. 71–104, 2012.
[31] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[32] E. del Arco, E. Morgado, M. I. Chidean, J. Ramiro-Bargueño, I. Mora-iménez, and A. J. Caamaño, "Sparse vehicular sensor networks for traffic dynamics reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2826–2837, Oct. 2015.
[33] K. Wei, R. Iyer, and J. Bilmes, "Submodularity in data subset selection and active learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1954–1963.
[34] Y. Shinohara, "A submodular optimization approach to sentence set selection," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, May 2014, pp. 4112–4115.
[35] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1135–1146, Mar. 2014.
[36] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin, "Simultaneous optimization of sensor placements and balanced schedules," *IEEE Trans. Autom. Control*, vol. 56, no. 10, pp. 2390–2405, Oct. 2011.
[37] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2. 2001, pp. 849–856.
[38] N. Lazic, I. Givoni, B. Frey, and P. Aarabi, "Floss: Facility location for subspace segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 825–832.
[39] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 51–65, Feb. 2015.
[40] N. Arbabzadeh and M. Jafari, "A data-driven approach for driving safety risk prediction using driver behavior and roadway information data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 446–460, Feb. 2018.
[41] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.
[42] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.



Qi Zhou received the bachelor's degree in computer science from East China Normal University, China, in 2015. He is currently pursuing the M.S. degree with the School of Computer Science, Fudan University. His research interests include machine learning, computer vision, intelligent transportation systems, and crowd counting.



Junping Zhang (M'05) received the bachelor's degree in automation from Xiangtan University, China, in 1992, the M.S. degree in control theory and control engineering from Hunan University, Changsha, China, in 2000, and the Ph.D. degree in intelligent systems and pattern recognition from the Institution of Automation, Chinese Academy of Sciences, in 2003. He has been a Professor with the School of Computer Science, Fudan University, since 2006. His research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems. He has been an Associate Editor of IEEE INTELLIGENT SYSTEMS since 2009 and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS since 2010.

Lingfu Che received the bachelor's degree from the School of Mathematical Sciences, Fudan University, China, in 2016, where he is currently pursuing the M.S. degree with the School of Computer Science. His research interests include machine learning, computer vision, intelligent transportation systems, and crowd counting.

Hongming Shan received the bachelor's degree from Shandong University of Technology, China, in 2011 and the Ph.D. degree from Fudan University, China, in 2017. He is currently a Post-Doctoral Scholar with Rensselaer Polytechnic Institute. His research interests include machine/deep learning, computer vision, dimension reduction, and biomedical imaging.

James Z. Wang received the bachelor's degree *(summa cum laude)* in mathematics and computer science from University of Minnesota and the M.S. degree in mathematics, the M.S. degree in computer science, and the Ph.D. degree in medical information sciences from Stanford University. He was a Visiting Professor with Robotics Institute, Carnegie Mellon University, from 2007 to 2008. He is a Professor of information sciences and technology at The Pennsylvania State University. His research interests include image analysis, image modeling, image retrieval, and their applications. He was a recipient of the National Science Foundation Career Award (2004). He was a Lead Special Section Guest Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (2008) and a Program Manager at the Office of the Director of the National Science Foundation (2011–2012).