Learning to Count with Back-Propagated Information

Ke Chen and Joni-Kristian Kämäräinen
Department of Signal Processing
Tampere University of Technology
http://vision.cs.tut.fi/

***Abstract*—Error back-propagation is one of the principled learning strategies widely used in pattern recognition and machine learning, e.g. neural networks. The existing frameworks employed back-propagated error as a performance criteria (or termed, object function) aiming for supervising model-learning. Inspired by the recent success achieved by learning with the privileged information (LPI), we propose a novel regression-based framework by extending the concept of back-propagation in supervised learning methods to high-level guiding the model learning, so the proposed model is able to mine the importance of samples contributed to the fitting performance, which is missed in the existing regression techniques. To verify the effectiveness of the proposed learning paradigm, both low-level imagery features and intermediary semantic attributes are adopted in this paper. Extensive evaluations on pedestrian counting with public UCSD and Mall benchmarks demonstrate the effectiveness of the proposed framework.**

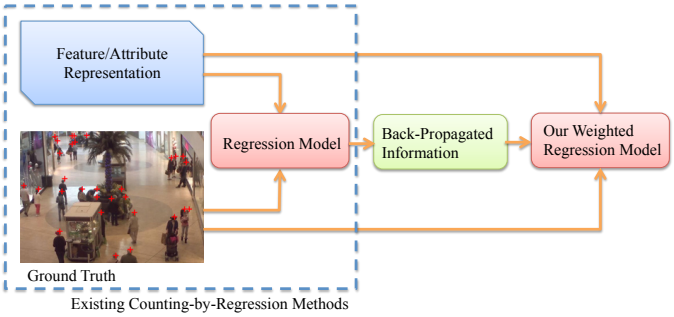


Figure 1: A flow chart illustrating the pipeline of learning to count approaches with and without the back-propagated information.

I. Introduction

Learning to count pedestrians in the video frames is a hot yet challenging problem in computer vision, which has its significance in public security and customers profiling. The problem is made difficult due to the dynamic spatially-distributed crowd pattern in the depths of the scene, frequent inter-object occlusion caused by high crowdedness, and perspective distortion.

A number of algorithms has been proposed to address such a problem, which can be categorised into counting-by-regression [1], [2], [3], [4], [5], counting-by-detection [6], [7], [8] and counting-by-clustering [9], [10]. Different from counting-by-detection and counting-by-clustering algorithms dependent on either the explicit object segmentation or temporal motion tracking, counting-by-regression methods [1], [2], [3], [4], [5], [11], [12], [13], [14] are more favourable for real-time surveillance owing to 1) more efficiency in the aspect of computation and 2) more effectiveness to cope with frequent inter-object occlusion (i.e. more crowded environment).

In the domain of counting-by-regression, the existing algorithms learn the regression mapping between low-level imagery feature and object count in a global manner (i.e. in the whole image) [1], [2], [11], in the local fashion (i.e. within the localised regions) [5], [12], [13], [14], or in the way of hybrid of the two [4]. The aforementioned counting-by-regression methods focused on mitigating the suffering of large feature variation by either proposing better regressors [1], [2], [4] or more robust features [5], [11], [12]. Recently, Chen et al [3] introduced the concept of cumulative attribute to the regression problems, which can significantly improve the counting performance because of capturing the cumulative dependent nature of regression labels. However, the existing counting-by-regression methods overlook an important question: *which training samples can contribute more than the rest to the performance?* Intuitively, the information to distinguish the *easier* from *harder* training samples will discover the importance of training data, which is useful for regression model learning.

Loy et al [15] attempted to address the aforementioned question in an active-learning framework by automatically selecting the most informative samples for label annotation. Nevertheless, the assumption in [15] that the lack of labelled training data is beyond the scope of the discussion in this paper. Inspired by the success achieved by classic BP neural network [16], [17] and learning with privileged information [18], [19], [20], regression learning with back-propagated information is proposed in this paper to enforce the model by weighting each sample with the back-propagated error in the learning procedure. Different from learning with privileged information [18], [19], [20], our proposed method do not need any additional efforts to provide other sources of information, which can be laborious and time-consuming. The pipelines of our method with back-propagated information and the existing counting-by-regression methods are illustrated in Figure 1.

The concept of learning with back-propagated information is dead simple: *the high-level "supervision" is constructed to give weights to each training sample according to the back-propagated error*. The notion of the proposed learning framework loosely corresponds to the observation that all the training samples are not equally informative. Apparently, the *easier* samples with smaller errors approximately capture the consistent part of the imagery representation and should be weighted higher, whereas the *harder* samples having larger variation could be noisy and need to be penalised with lower weights. Consequently, tackling with the *easier* and *harder* samples respectively can contribute to learning a better regression model.

II. Related Work

Counting-by-Regression – Most existing techniques for learning to count persons in a regression framework fall into three categories: global approaches [1], [2], [3], [11], local approaches [5], [12], [13], [14], and hybrid of both approaches [4]. Chan et al [1], [2] firstly proposed to learn a regression mapping between the low-level imagery feature extracted from the whole image and the number of count for the image. To model the dynamic crowd pattern in local regions, the approaches in [5], [12] divided the images into several local regions and then learned an independent regressor for each region. Recently, Lempitsky and Zisserman [13] further extended the region-level local approaches to pixel-level density estimation, which is able to provide the count within any given image region. In [14], a simplified patch-to-patch framework based on structured regression forest is proposed to achieve comparable results to Lempitsky's model [13]. Chen et al [4] proposed a single multi-output model to mine the the importance and discover latent dependency of localised features in a unique framework, which takes both advantages of global and local approaches. They further introduced the concept of cumulative attribute for regression for addressing both challenges of feature inconsistency and sparse & imbalanced data by capturing the cumulative dependent nature of regression labels [3]. Either the feature representation or attribute representation used in the existing supervised counting-by-regression methods miss to discover the importance of samples to favour for learning a better regression model.

Sample-Importance Discovery – Cost sensitive learning algorithms for classification [21], [22], [23] were exploited for feeding the penalised cost to the original classification frameworks aiming to reflect the severity of misclassification problem. Different from the standard classification frameworks utilising the total error as the price of misclassification, cost-sensitive learning techniques is to minimise the total cost, which are usually pre-defined manually. Evidently, the cost definition in cost-sensitive learning frameworks is less flexible and/or laborious because of the need of the efforts by humans. Recently, learning with privileged information (LPI) [18], [19], [20] attracts the wide attention, as the importance of samples can be weighted with the additional privileged information, which is similar to our idea. However, the learning paradigm with privileged information also cannot avoid using the expensive manpower to collect and/or annotate the privileged information. To overcome the limitation of learning with privileged information, our framework with back-propagated information automatically generated by the model is proposed and demonstrated its effectiveness.

Contributions – The main contributions and novelties of this work are three-fold:

- The proposed learning paradigm is generic and can be readily generalised to other supervised regression frameworks.
- To the authors' best knowledge, learning with back-propagated information is a novel approach to seek support from the importance of training sample without any price to pay.
- Extensive experiments on pedestrian counting with the UCSD and Mall benchmarks verify the motivation of our proposed learning framework.

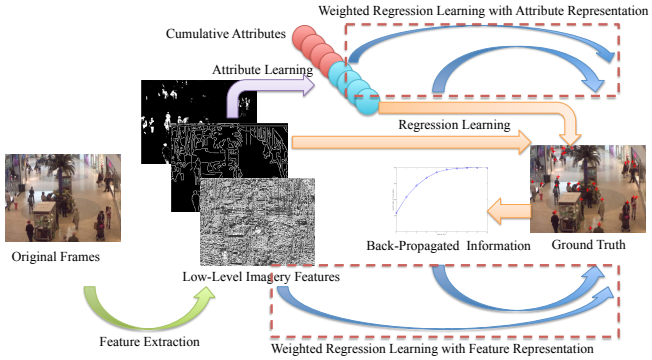


Figure 2: The proposed weighted regression framework incorporating the back-propagated information with low-level image feature or intermediary semantic attribute representations for pedestrian counting.

III. Methodology

As shown in Figure 2, our proposed counting-by-regression framework with the back-propagated information can be formulated as a weighted regression method that uses the back-propagated error measurement as the weights to discover the importance of samples. It is worth pointing out that, in our framework, both low-level feature and intermediate semantic attribute representations can employ such a back-propagated information. During training, the proposed counting-by-regression method consists of the following steps:

Step 1: Given ith training video frame, we extract low-level imagery features $\mathbf{x}_i, i = 1, 2, \cdots, N$ including segment, edge and texture shown as the green arrow in Figure 2, where N denotes the total number of training frames. With annotated count labels y_i, the training pair consists of $\{(\mathbf{x}, y)\}_i$ (Refer to Section III-A).

Step 2: Alternatively, according to annotated count labels y_i, a binary cumulative attribute vector $\mathbf{a}_i$ defined in [3] can be generated automatically. As a result, we can obtain the training pairs consisting of features, cumulative attributes and labels $\{(\mathbf{x}, \mathbf{a}, y)\}_i$ shown as the purple arrow in Figure 2 (Refer to Section III-B).

Step 3: Back-propagated information is then computed by the errors between the estimated output of the learned regressor using feature/attribute representation and count labels as the orange arrows shown in Figure 2 (Refer to Section III-C).

Step 4: With the back-propagated errors as the weights indicating the importance of each sample, a weighted regression model is then learned to map either feature representation $\mathbf{x}_i$ or cumulative attribute representation $\hat{\mathbf{a}}_i$ as the input to the scalar-valued output illustrated by the blue arrows (Refer to Section III-D).

During testing, given an unseen image, the low-level features or intermediary attributes are constructed as the input, and then mapped to the corresponding weighted regression model. It is noted that the back-propagated information is only generated and used during training to enforce sample mining, whereas the learned weights will be used directly for testing.

Table I: Dimensions of the low-level image feature vector.

Features	Dims	Params
Area	1	–
Perimeter	1	–
Perimeter-area ratio	1	–
Perimeter orientation	6	$0^o, 30^o, 60^o, 90^o, 120^o, 150^o$
Blob count	1	the size of blob $>$ 10 pixel
Edge	1	–
Edge orientation	6	$0^o, 30^o, 60^o, 90^o, 120^o, 150^o$
Edge Minkowski	1	–
Texture homogeneity	4	$0^o, 45^o, 90^o, 135^o$
Texture energy	4	$0^o, 45^o, 90^o, 135^o$
Texture entropy	4	$0^o, 45^o, 90^o, 135^o$

A. Feature Representation

Given a training video frame i, three types of features are extracted to form a 30-dimensional feature vector as in [1]:

- Segment-based features: the total number of pixels of foreground area (Area), the total number of pixels of foreground perimeter (Perimeter), the ratio of the number of pixels between perimeter and area (Perimeter-are ratio), histogram bins of perimeter edge orientation (Perimeter orientation), and the count of blob (Blob count);
- Edge-based features: the total number of edge pixels (Edge), histogram bins of edge orientation (Edge orientation), and the Minkowski dimension [24];
- Texture features: the homogeneity, energy and entropy of Gray Level Co-occurrence Matrix (GLCM) [25].

More details are given in Table I. It is worth pointing out that all frames are transformed to gray-scale prior to feature extraction. In addition, features are normalised due to perspective distortion and scaled into $[0, 1]$ [1].

B. Attribute Representation

Cumulative attribute [3] is recently proposed for regression problems, which is more discriminative than low-level feature representation. For verifying the proposed learning paradigm, cumulative attribute is also adopted in our framework as an alternative choice of low-level features. Given the ith training data point, the scalar-valued output y_i is converted into a cumulative attribute vector $\mathbf{a}_i$. The dimensionality m of the vector $\mathbf{a}_i$ usually depends on the value range of y. The mathematical formulation of such a cumulative attribute in the form of binary vector can be written as the following:

$$a_i^j = \begin{cases} 1, & \text{when } \ j \leqslant y_i, \\ 0, & \text{when } \ j > y_i, \end{cases}$$

where $j = 1, 2, \cdots, m$ denotes the jth entry of the cumulative attribute vector. Examples of cumulative attribute $\mathbf{a}_i$ is shown in Figure 3.

With training set represented as $\{(\mathbf{x}, \mathbf{a}, y)\}_i, i = 1, 2, \cdots, N$, joint attribute learning is employed with $\{(\mathbf{x}, \mathbf{a})\}_i$ as the input and output of model [3]. With learned joint attribute model, the intermediary attribute representation $\hat{\mathbf{a}}_i$ can thus be achieved by mapping low-level imagery features to cumulative attribute space.

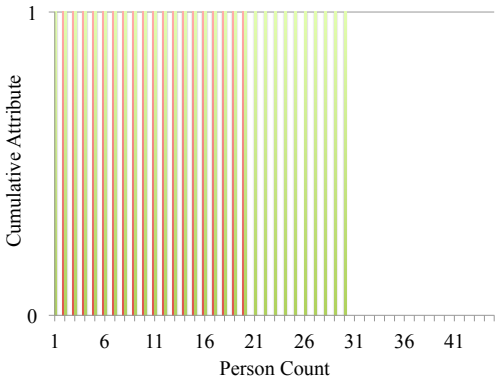


Figure 3: Cumulative attribute with red and green bars denoting examples with 20 and 30 person count respectively. The distance between two examples in cumulative attribute space is proportional to the difference in label space.

C. Back Propagated Information

Before giving more details about the back-propagated information, let us define the input of the model to be $\mathbf{r}_i, i = 1, 2, \cdots, N$ regardless of using either feature or attribute representation, for the convenience of the presentation. Now the training pair is represented by $\{(\mathbf{r}, y)\}_i, i = 1, 2, \cdots, N$. For learning a regression mapping with the training pair $\{(\mathbf{r}, y)\}_i, i = 1, 2, \cdots, N$ to automatic discover the importance of each sample, a ridge regression function is considered here as its proven efficiency for crowd counting problem [3], [4], [15], which can be formulated as the following:

$$\min \quad \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N} \text{loss}(y_i, f(\mathbf{r}_i)), \tag{1}$$

where $\mathbf{w}$ is the weight vector to be optimised, C denotes the trade-off parameter between the regularised term and the loss function $\text{loss}(\cdot)$, and $f(\mathbf{r}_i) = \mathbf{w}^T\mathbf{r}_i + b$.

For simplifying Equation (1) with a quadratic loss function, the following formulation can be derived:

$$\min \quad \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N} ||y_i - (\mathbf{w}^T\mathbf{r}_i + b)||_2^2. \tag{2}$$

Considering the differentiability of Equation (2), the gradient of such an object function is enforced to be zero, which thus has the following close-form solution:

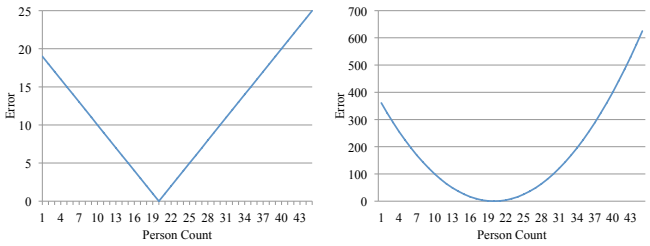
$$\begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = -(Q^TQ)^{-1}Q^T\mathbf{p}, \tag{3}$$

where b is the bias term, and positive semi-definite matrix Q and vector $\mathbf{p}$ are given as

$$Q = \begin{bmatrix} 2C\sum_{i=1}^{N}\mathbf{r}_i\mathbf{r}_i^T + I & 2C\sum_{i=1}^{N}\mathbf{r}_i \\ 2C\sum_{i=1}^{N}\mathbf{r}_i^T & 2CN \end{bmatrix},$$
$$\mathbf{p} = \begin{bmatrix} -2C\sum_{i=1}^{N}\mathbf{r}_i y_i^T \\ -2C\sum_{i=1}^{N} y_i^T \end{bmatrix}.$$

The parameter C is determined by n-fold cross validation.

Now, the back-propagated information based on the error measurement can be generated with learned $\mathbf{w}$ and b, and we



(a) Absolute Error (b) Square Error

Figure 4: Two types of error measurement as back-propagated information for a frame with 20 persons within the region of interest using the UCSD dataset.

will introduce two types of widely-used error measurement presented below, which is also illustrated in Figure 4:

- absolute error: $\epsilon_{abs}(\mathbf{r}_i, y_i) = |\mathbf{w}^T\mathbf{r}_i + b - y_i|$, for $i = 1, 2, \cdots, N$.
- square error: $\epsilon_{sqr} = (\mathbf{r}_i, y_i) = ||\mathbf{w}^T\mathbf{r}_i + b - y_i||$, for $i = 1, 2, \cdots, N$.

In view of the aforementioned error measurement used as weights (i.e. the higher weights corresponding to the lower errors), the back-propagated information is constructed as: the error measurement is first normalised and scaled into $[0, 1]$; the back-propagated information is then computed and represented by the difference between the normalised errors and the upper boundary of the range (i.e. 1 in this paper).

D. Weighted Regression Learning

With the training pair $\{(\mathbf{r}, y)\}_i, i = 1, 2, \cdots, N$ and $v_i, i = 1, 2, \cdots, N$ denoting the back-propagated information, the weighted ridge regression [26], [27] is considered here to "update" the model integrating the importance of each sample with the formulation presented as follows:

$$\min \quad \frac{1}{2}||\hat{\mathbf{w}}||_2^2 + C\sum_{i=1}^{N} v_i \text{loss}(y_i, f(\mathbf{r}_i)). \tag{4}$$

Similarly, with quadratic loss function, Equation (4) can thus be written as:

$$\min \quad \frac{1}{2}||\hat{\mathbf{w}}||_2^2 + C\sum_{i=1}^{N} v_i||y_i - (\hat{\mathbf{w}}^T\mathbf{r}_i + \hat{b})||_2^2. \tag{5}$$

Because of v_i being non-negative, we can obtain the closed-form solution $[\hat{\mathbf{w}}, \hat{b}]^T$of Equation (5) by updating Q and $\mathbf{p}$ in Equation (3) with the following $\hat{Q}$ and $\hat{\mathbf{p}}$:

$$\hat{Q} = \begin{bmatrix} 2C\sum_{i=1}^{N} v_i\mathbf{x}_i\mathbf{x}_i^T + I & 2C\sum_{i=1}^{N} v_i\mathbf{x}_i \\ 2C\sum_{i=1}^{N} v_i\mathbf{x}_i^T & 2C\sum_{i=1}^{N} v_i \end{bmatrix},$$

$$\hat{\mathbf{p}} = \begin{bmatrix} -2C\sum_{i=1}^{N} v_i\mathbf{x}_i y_i^T \\ -2C\sum_{i=1}^{N} v_i y_i^T \end{bmatrix}.$$

The weight v of each instance has an important effect in giving the information of sample mining during training. Specifically, each training sample has a weight scaler indicating the importance, which will contribute to construct a better regressor. Intuitively, the proposed algorithm can be viewed as inserting one latent feature/attribute space and one latent label space between original feature/attribute space and label space mapped by using the square root of $v_i, i = 1, 2, \cdots, N$. Compared to the original regression methods, our proposed model can thus get more benefits from the deeper architecture. An alternative to weighted ridge regression is weighted support vector regression [21], [28] and conditional regression forest [20], while weighted ridge regression is adopted here owing to the success achieved by ridge regression [3], [4] for crowd counting problem and its simplicity in implementation.

IV. Experiments

A. Datasets and Settings

(a) the UCSD Dataset (b) the Mall Dataset

Figure 5: The used public benchmark datasets.

Table II: Dataset properties: N_f = number of frames, R = Resolution, FPS = frame per second, and D = Density (minimum and maximum number of people in the region of interest).

Data	N_f	R	FPS	D
UCSD [1]	2000	238 × 158	10	11–46
Mall [4]	2000	320 × 240	<2	13–53

Datasets – In crowd counting, two benchmarking datasets are adopted in the experiments, i.e. the UCSD [1], [2], [3], [4] and the Mall [3], [4], [15] datasets which feature an outdoor and an indoor scene respectively. Illustrative examples and details of both datasets are given in Figure 5 and Table II.

Settings – Following the settings in [3], [4], the training and testing partition are Frames 601-1400 for training and the rest for testing with the UCSD, while the first 800 frames of the Mall datasets are used for training and the remaining 1200 frames are employed for testing.

Comparative Evaluation – We compared the following counting-by-regression models:

- Ordinary counting-by-regression methods with low-level imagery features including Gaussian Process Regression (GPR) [1] with linear + RBF kernel and Ridge Regression (RR) [3], [4];
- Ridge Regression incorporating Cumulative Attribute Representation (CA-RR) [3];
- Ridge Regression utilising learning with privileged attribute information (LPI-RR) [18];
- The proposed Weighted Ridge Regression (WRR) with the back-propagated information described in Sec. III-D.

The free parameters of the models were tuned using 4-fold cross-validation.

Evaluation Metrics – In addition to three evaluation metrics used in [3], [4], we introduce another metric, namely

cumulative score originally designed for the evaluation in age estimation [29], to evaluate the percentage of testing data within the tolerance of absolute error level. Specially, four evaluation metrics: *mean absolute error* (mae), δ_{abs}; *mean squared error* (mse), δ_{sqr}; *mean deviation error* (mde), δ_{dev}; and *cumulative score* (cs), δ_{cs} are given as:

$$\delta_{\text{abs}} = \frac{1}{N}\sum_{i=1}^{N}|e_i - \hat{e}_i|, \quad \delta_{\text{sqr}} = \frac{1}{N}\sum_{i=1}^{N}(e_i - \hat{e}_i)^2,$$

$$\delta_{\text{dev}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|e_i - \hat{e}_i|}{e_i}, \quad \text{and} \quad \delta_{\text{cs}} = \frac{M_{\epsilon_{abs}\leqslant L}}{M} \times 100\%,$$

where N is the total number of test frames, e_i is the actual count in the whole image, $\hat{e}_i$ is the estimated count of ith frame, $M_{\epsilon_{abs}<L}$ is the total number of testing frames whose absolute error ϵ_{abs} is less than the tolerated error level L, and M is the total number of testing frames. In our experiments, the error level L is set to be 2 according to mean absolute error reported in [3], [4].

B. Results

Table III: Comparative evaluation using low-level features.

Method	UCSD [1]				Mall [4]			
	mae	mse	mde	cs	mae	mse	mde	cs
GPR [2]	2.30	8.21	0.114	60%	3.72	20.1	0.115	37%
RR [3]	2.25	7.82	0.110	72%	3.59	19.0	0.110	46%
WRR	2.11	7.11	0.105	72%	3.58	19.0	0.110	46%

Comparative Evaluation with Feature Representation – Table III compares the results on person counting with three counting-by-regression methods using two benchmarking datasets. The results illustrate that our proposed weighted ridge regression (WRR) model with the back-propagated information achieves superior performance to the-state-of-art methods for both datasets. More specifically, the results generated by our method is slightly superior to other counting-by-regression methods with the Mall dataset, whereas our method can significantly outperform the other evaluated methods with the UCSD benchmark. The reason about leading to the difference of performance on two datasets is the variation of feature representation caused by the different characteristics of the scenes. As Figure 6 shows, more than 80% of training samples in UCSD can be fitted well within the range of error level 2 when using standard ridge regression, while the error level of the Mall dataset is increased to 4, which indicates that the features of the Mall dataset have larger variation and noises. As a result, the back-propagated information with the Mall dataset is less discriminative than that of the UCSD dataset, which leads to performance gap between two datasets.

Table IV: Comparative evaluation using attributes.

Method	UCSD [1]				Mall [4]			
	mae	mse	mde	cs	mae	mse	mde	cs
CA-RR [3]	2.07	6.86	0.102	74%	3.43	17.7	0.105	48%
WRR	2.05	6.75	0.102	74%	3.44	18.0	0.105	48%

Comparative Evaluation with Attribute Representation – In addition to low-level feature representation, recently-proposed cumulative attribute [3] is also adopted to evaluate the proposed learning paradigm wit back-propagated information for sample mining. As shown in Table IV, using intermediate attribute representation, the counting performance of the proposed WRR model is comparable to that of the-state-of-art CA-RR [3]. Evidently, the similar results generated by CA-RR and WRR can be caused due to the usage of powerful cumulative attribute. In details, with more discriminative attribute representation capturing the cumulative dependency of regression labels, the back-propagated information used in our WRR is consistent with the information discovered by cumulative attribute.

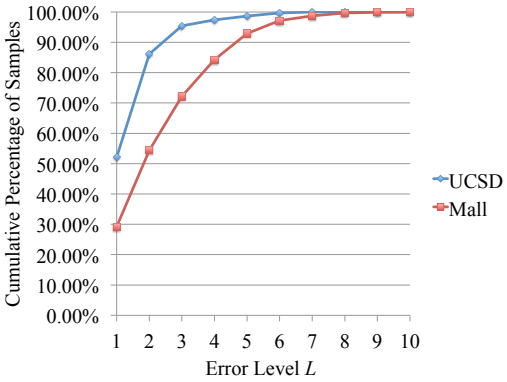


Figure 6: The absolute error distributions in the terms of cumulative percentage of training samples and count error using low-level features. The similar phenomenon can also be observed using the attribute representation.

Table V: Back-Propagated vs. Privileged Information.

Method	UCSD [1]				Mall [4]			
	mae	mse	mde	cs	mae	mse	mde	cs
LPI-RR [18]	2.10	7.05	0.104	72%	3.58	19.0	0.110	45%
WRR	2.05	6.75	0.102	74%	3.44	18.0	0.105	48%

Back-Propagated vs. Privileged Information – In [18], attributes can be employed as privileged information rather than representation, which inspires us to propose the learning framework with back-propagated information. Table V illustrates the comparative results by using back-propagated and privileged information respectively. Compared to learning with privileged information (LPI-RR) model [18], [19], [20], our WRR model can perform better in all four evaluation metrics for both datasets. Since both models use the same cumulative attribute and the same regression model, the performance gain can only be explained by the fact that our proposed learning framework with attribute representation is superior to the learning framework with privileged attribute information in pedestrian counting problem. Moreover, our WRR model with feature representation shown in Table III is still comparable or even better than attribute-based LPI-RR [18], which further demonstrates the superiority of our proposed model to learning with privileged information.

Absolute vs. Square Error – The results with feature and attribute representation in Table VI are to compare two types

Table VI: Our method using absolute (WRR-abs) vs. square (WRR-sqr) errors with low-level features (the top two rows) and attributes (the bottom two rows).

Method	UCSD [1]				Mall [4]			
	mae	mse	mde	cs	mae	mse	mde	cs
WRR-abs	2.11	7.13	0.105	72%	3.58	19.0	0.110	46%
WRR-sqr	2.11	7.11	0.105	72%	3.58	19.0	0.110	46%
WRR-abs	2.05	6.76	0.102	73%	3.44	18.0	0.105	48%
WRR-sqr	2.05	6.75	0.102	74%	3.45	17.9	0.105	48%

of error measurement presented in Section III-C. The performance with two types of error measurement for both datasets are almost identical, which shows that the type of the back-propagated error adopted in this paper has less effect on the final performance.

V. Conclusion

We have proposed a counting-by-regression method based on a new learning paradigm – learning with back-propagated information. The importance of samples is discovered and mined in the manner of weights to improve the regression performance with low-level features and intermediary attributes. In addition, compared to other supervised information learning frameworks, no additional efforts in terms of labels or annotation are needed but are automatically estimated efficiently with the existing training data. The experimental results on pedestrian counting substantiate that our proposed method can outperform the-state-of-arts. Moreover, the results also illustrate that the "*quality*" (i.e. discriminative) of the back-propagated information plays an important role for contributing to a better regression model. In the light of this, our future work to construct more robust back-propagated information could be a promising direction.

Acknowledgement

This work is funded by the Academy of Finland Grant No. 267581 "Learning 10,000 Visual Classes and Some Stuff".

References

[1] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: counting people without people models or tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2160–2177, 2012.

[3] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[4] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proceedings of British Machine Vision Conference*, 2012.

[5] W. Ma, L. Huang, and C. Liu, "Crowd density analysis using co-occurrence texture features," in *Proceedings of International Conference on Computer Sciences and Convergence Information Technology*, 2010.

[6] W. Ge and R. Collins, "Marked point processes for crowd counting," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[7] 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 *Proceedings of International Conference on Pattern Recognition*, 2008.

[8] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1198–1211, 2008.

[9] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[10] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[11] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *Proceedings of Digital Image Computing: Techniques and Applications*, 2009.

[12] X. Wu, G. Liang, K. Lee, and Y. Xu, "Crowd density estimation using texture analysis and learning," in *Proceedings of IEEE International Conference on Robotics and Biomimetics*, 2006.

[13] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proceedings of Advanced in Neural Information Processing Systems*, 2010.

[14] L. Fiaschi, R. Nair, U. Koethe, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proceedings of International Conference on Pattern Recognition*, 2012.

[15] C. C. Loy, S. Gong, and T. Xiang, "From semi-supervised to transfer counting of crowds," in *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, 1986.

[17] Y. Zhang, Z. Li, K. Chen, and B. Cai, "Common nature of learning exemplified by bp and hopfield neural networks for solving online a system of linear equations," in *Proceedings of IEEE International Conference on Networking, Sensing and Control*, 2008.

[18] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," in *Proceedings of IEEE International Conference on Computer Vision*, 2013.

[19] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 56, pp. 544 – 557, 2009.

[20] H. Yang and I. Patras, "Privileged information-based conditional regression forest for facial feature detection," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2012.

[21] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[22] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358 – 3378, 2007.

[23] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proceedings of IEEE International Conference on Data Mining*, 2003.

[24] A. Marana, L. da Fontoura Costa, R. Lotufo, and S. Velastin, "Estimating crowd density with minkowski fractal dimension," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[25] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[26] S. M. Zahari, M. S. Zainol, M. I. Al-Banna, and B. Ismail, "Weighted ridge mm-estimator in robust ridge regression with multicollinearity," *Proceedings of Mathematical Models and Methods in Modern Science*, 2012.

[27] P. W. Holland, *Weighted ridge regression: combining ridge and robust regression methods*. National Bureau of Economic Research Cambridge, Mass., USA, 1973.

[28] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Proceedings of International Conference on Pattern Recognition*, 2010.

[29] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234–2240, 2007.