Feature Mining for Localised Crowd Counting

Ke Chen[1]
cory@eecs.qmul.ac.uk

Chen Change Loy[2]
ccloy@visionsemantics.com

Shaogang Gong[1]
sgg@eecs.qmul.ac.uk

Tao Xiang[1]
txiang@eecs.qmul.ac.uk

[1] School of Electronic Engineering and Computer Science
Queen Mary, University of London
London E1 4NS, UK

[2] Vision Semantics
London E1 4NS, UK

Abstract

This paper presents a multi-output regression model for crowd counting in public scenes. Existing counting by regression methods either learn a single model for global counting, or train a large number of separate regressors for localised density estimation. In contrast, our single regression model based approach is able to estimate people count in spatially localised regions and is more scalable without the need for training a large number of regressors proportional to the number of local regions. In particular, the proposed model automatically learns the functional mapping between interdependent low-level features and multi-dimensional structured outputs. The model is able to discover the inherent importance of different features for people counting at different spatial locations. Extensive evaluations on an existing crowd analysis benchmark dataset and a new more challenging dataset demonstrate the effectiveness of our approach.

1 Introduction

Crowd counting in public places has a wide spectrum of applications especially in crowd control, public space design, and pedestrian behaviour profiling. In some applications, e.g. crowd counting on a train platform, estimating a global count for the whole scene is sufficient. For more complex scenarios, it is necessary to estimate the counts at different spatial locations as well. For instance, for crowd counting in a shopping mall, one needs to know not only how many people in total are in the scene, but also where they are distributed, i.e. which shop is more popular.

Existing people counting techniques fall into three categories: counting by detection, counting by clustering, and counting by regression. In counting by detection [8, 16, 24], people count is estimated through detecting instances of people. Typically, the detection process is time-consuming since it involves exhaustive scanning of image space using a pre-trained detector with different scales. In counting by clustering [4, 20], a crowd is assumed to be composed of individual entities, each of which has unique yet coherent motion patterns
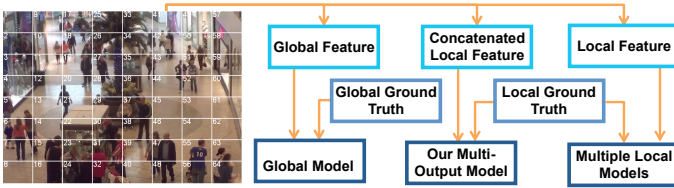
Figure 1: A flow chart illustrating the processing pipeline of global and local counting by regression methods, and our multi-output model.

that can be clustered to approximate the number of people. Such a motion clustering based method only works well with sufficiently high video frame rate so motion information can be extracted reliably. The counting by detection and by clustering approaches either rely on explicit object segmentation or feature point tracking. They are not suitable for crowded scenes with cluttered background and frequent inter-object occlusion. In contrast, a counting by regression model aims to learn a direct mapping between low-level features and people count without segregation or tracking of individuals. This approach is more suitable for crowded environments and is computationally more efficient.

Existing counting by regression methods can be categorised into either global approaches or local approaches (see Figure 1). Global approaches [5, 7, 14, 18, 21] learn a single regression function between image features extracted globally from the entire image space and the total people count in that image. Since spatial information is lost when computing global features, such a model assumes implicitly that a feature should be weighted the same regardless where in the scene it is extracted. However, this assumption is largely invalid in real-world scenarios. In particular, crowd structures[1] can vary spatially due to density, scene layout, and self-organisation of crowd induced by elementary individual interactions, boundary conditions, and regulations [11]. Thus, different features can be more reliable and relevant for crowd counting at different spatial locations. Furthermore, a global regression model is unable to provide information about spatially local crowd count information, which is desired in some applications.

To overcome these limitations of a global approach, local models [17, 23] aim to relax the global assumption to certain extent by dividing the image space into cell regions, each of which modelled by a separate regression function. The regions can be cells having regular size, or different resolutions determined by the scene perspective to compensate for camera geometric distortions [17]. Local counts can be estimated in each region and a global count can then be obtained by summing up the cell-level counts. In an extreme case, Lempitsky et al.[15] go one step further to model the crowd density at each pixel, casting the problem as that of estimating an image density whose integral over any image region gives the count of objects within that region. In general, unlike global approaches [5, 7, 14, 18], local models aim to weigh features differently by local crowd structures in order to facilitate localised crowd counting. However, existing local methods suffer a scalability issue due to the need to learn multiple regression models, the number of which can become very large. In addition, an inherent drawback of existing local models is that no information is shared across spatially localised regions in order to provide a more context-aware feature selection for more accurate crowd counting. In many real-world cases, low-level imagery features can be highly ambiguous due to cluttered background and severe inter-object occlusions. Therefore, har-

---

[1]Systematic granular motion of crowd resembling the flow of gas, fluid, and granular media.

nessing common properties and features among different local spatial regions should benefit the estimation of crowd density.

We consider that *localised feature importance mining* and *information sharing among regions* are two key factors for accurate and robust crowd counting, which are missing in all existing techniques. To this end, we propose a single multi-output model for joint localised crowd counting based on ridge regression [22], which takes inter-dependent local features from local spatial regions as input and people count from individual regions as multi-dimensional structured output (see Figure 1). Unlike global regression methods, our model relaxes the one-to-one mapping assumption by learning spatially localised regression functions jointly in a single model for all the individual cell regions in a scene, as such our model can capture feature importance locally. Unlike existing approaches to building multiple local regression models, our single model is learned by joint optimisation to enforce dependencies among cell regions. Therefore information from all local spatial regions can be shared to achieve more reliable count prediction. We demonstrate the effectiveness of our model on both an existing crowd analysis benchmark dataset and a new more challenging shopping mall dataset. In summary, the main contributions and novelties of this study are three-fold:

- This is the first study that achieves robust crowd counting by mining local feature importance and sharing visual information among spatially localised regions in a scene.
- This is achieved by considering a single multi-output ridge regression model for localised crowd counting which has advantages over both existing global approaches in providing local estimates and existing local approaches being more scalable.
- We introduce a new public scene dataset of over 60,000 pedestrian instances for crowd analysis. To our best knowledge, it is the largest dataset to date with the most realistic and challenging setting of a crowded scene in a public space.

2 Methodology

Figure 2 gives an overview of our framework: (Step-1) We first infer a perspective normalisation map using the method described in [6]. (Step-2) Given a set of training images, we extract low-level imagery features, including local foreground, edges and texture features, from each cell region. (Step-3) Local features from each cell are used to construct a local intermediate feature vector before all local intermediate feature vectors are concatenated into a single ordered (location-aware) feature vector. (Step-4) A multi-output regression model based on multivariant ridge regression is trained using the single concatenated feature vector and the vector, each element being actual count in each region, as a training pair. Given a new test frame, features are extracted and mapped to the learned regression model for generating a structured output that estimates the crowd count in each local region simultaneously.

Note that the training/testing procedure adopted in our framework is similar to that in a global counting framework (see Figure 1), but with a different and new learning strategy to enable spatially localised features weighting and inter-region feature sharing. This variation is important to our approach. As in [15], our method requires dot annotations on each pedestrian so we can generate a training count for each cell region. This may appear a laborious task but dotting/pointing is the natural way of how human numerate objects. In practice dotted annotation is no harder than raw count as in the global counting methods [5]. All the labelled data with ground truth is released publicly (http://www.eecs.qmul.ac.uk/~ccloy/downloads_mall_dataset.html).
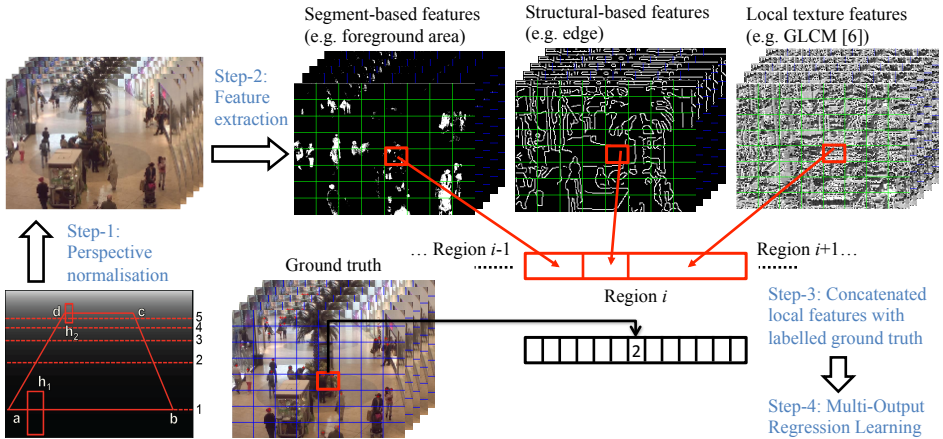
Figure 2: A multi-output regression framework for localised crowd counting by feature mining.

2.1 Feature Representation

Given a training video frame i, where $i = 1, 2 \ldots N$ and N denotes the total number of training frames, we first partition the frame into K cell regions (see Step-3 in Figure 2). We then extract low-level imagery features $\mathbf{z}_i^j$ from each cell region j and combine them into an intermediate feature vector $\mathbf{x}_i \in \mathbb{R}^d$. We also concatenate the localised labelled ground truth u_i^j from each cell region into a multi-dimensional output vector, $\mathbf{y}_i \in \mathbb{R}^m, i = 1, 2 \ldots N$

$$\mathbf{x}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^{K-1}, \mathbf{z}_i^K], \quad \mathbf{y}_i = [u_i^1, u_i^2, \ldots, u_i^{K-1}, u_i^K].$$

We train the proposed model using $\{(\mathbf{x}, \mathbf{y})\}_i, i = 1, 2 \ldots N$. In this study, we adopt three types of features as in [6]:

- Segment-based features: foreground area, total number of pixels of the foreground perimeter, perimeter-area ratio, and histogram of perimeter edge orientation;
- Structural-based features: total number of edge pixels, histogram of the edge orientation, and the Minkowski dimension [19];
- Local texture features: Gray Level Co-occurrence Matrix (GLCM) [10].

Note that all images are transformed to grayscale prior to feature extraction. In addition, features are perspective normalised using the method described in [6] and scaled into $[0, 1]$.

2.2 Multi-Output Regression Model

For learning a multi-output regression model, we exploit the ridge regression function [1, 9]. In its conventionally form, a ridge regression function learns a single output mapping. In our case, we adapt it to cope with a multi-outputs regression learning problem for simultaneous localised crowd counting in different spatial cell regions. The rational for exploiting ridge regression is that the model offers superior robustness in coping with multicollinearity problem[2], due to its regularised least-square error minimisation, as opposed to ordinary

---

[2]Some low-level features may be highly co-linear, unstable estimate of parameters may occurs [8], leading to very large magnitude in the parameters and therefore a clear danger of severe over-fitting.

least-square in classic regression methods such as linear regression. Ridge regression has been exploited elsewhere for face recognition [1]. This is the first attempt to exploit it for crowd analysis.

Formally, given $(\mathbf{x}_i, \mathbf{y}_i)$ as the observation and target vectors, multivariate ridge regression can be presented as follows

$$\min \quad \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^{N}\|\mathbf{y}_i^T - \mathbf{x}_i^T\mathbf{W} - \mathbf{b}\|_F^2, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times K}$ and $\mathbf{b} \in \mathbb{R}^{1 \times K}$ denote a weight matrix and a bias vector respectively. The $\|\cdot\|_F$ denotes the Frobenius-norm, and C is a parameter that controls the trade-off between the penalty and the fit.

The weight matrix $\mathbf{W}$ plays an important role in capturing the local feature importance and facilitating the sharing of features. In particular, for each localised cell, we formulate our model to jointly weigh the features extracted from both the corresponding localised cell and other cell regions in the image. According to the above Equation (1), for jth cell region in the images, jth column of matrix $\mathbf{W}$ is employed to weigh the concatenated feature vector $\mathbf{x}_i$ for the count estimation in corresponding localised cell region, i.e. jth entry of $\mathbf{y}_i$. Considering the residual error of all cell regions being penalized jointly with Frobenius-norm and feature vector $\mathbf{x}_i$ consisting of feature from both jth cell region and other cell regions in the image, such a regression model can benefit from local feature importance mining, and more importantly, feature information sharing within the whole image space for the localised crowd density estimation of a specific region.

Here we provide more details on the error minimisation. Specifically, the above Equation (1) is transformed as follows

$$\min \quad M(\theta) = \operatorname{tr}(\frac{1}{2}\theta^T Q\theta + P^T\theta), \tag{2}$$

where positive semi-definite matrix Q and matrix P are given as

$$Q = \begin{bmatrix} 2C\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T + I & 2C\sum_{i=1}^{N}\mathbf{x}_i \\ 2C\sum_{i=1}^{N}\mathbf{x}_i^T & 2CN \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \quad P = \begin{bmatrix} -2C\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i^T \\ -2C\sum_{i=1}^{N}\mathbf{y}_i^T \end{bmatrix} \in \mathbb{R}^{(d+1)\times K},$$

where $\theta = [\mathbf{W};\mathbf{b}] \in \mathbb{R}^{(d+1)\times K}$ represents the matrix to be optimized, I denotes the identity matrix, and $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. Different from the standard ridge regression with single output, the coefficient P and parameters θ to be optimized in Equation (2) are matrices instead of vectors, which leads to the usage of the trace $\operatorname{tr}(\cdot)$ for minimisation. Similar to ridge regression, Equation (2) is solved using the Quadratic Programming, which has a global optimal solution, if and only if

$$\frac{\partial M(\theta)}{\partial \theta} = Q\theta + P = 0,$$

and thus, the weights and bias of ridge regression are computed by

$$\theta = -(Q^T Q)^{-1}Q^T P.$$

An alternative to the multi-output ridge regression model is the structural Support Vector Machine [13], which has been applied to pose estimation [12] and object detection [3]. Multivariate ridge regression is adopted owing to its simplicity in implementation.

3 Experiments

Data	N_f	R	FPS	D	Tp
UCSD [6]	2000	238×158	10	11–46	49885
Mall	2000	320×240	<2	13–53	62325

Table 1: Dataset properties: N_f = number of frames, R = Resolution, FPS = frame per second, D = Density (minimum and maximum number of people in the ROI), and Tp = total number of pedestrian instances.

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

Figure 3: (a) the UCSD benchmark dataset and (b) the new Mall dataset.

Datasets – The effectiveness of the proposed method was evaluated on two datasets: the UCSD benchmark dataset [5, 6], and a new Mall dataset we introduce here, which is released with labelled ground truth (http://www.eecs.qmul.ac.uk/~ccloy/downloads_mall_dataset.html). The details of the two datasets are given in Table 1, and the example frames are shown in Figure 3. In contrast to the UCSD dataset, of which the video was recorded from a campus scene using hand-held camera (Figure 3(a)), the new Mall dataset was captured using a publicly accessible surveillance camera in a shopping mall with more challenging lighting conditions and glass surface reflections. The Mall dataset also covers more diverse crowd densities from sparse to crowded, as well as different activity patterns (static and moving crowds) under larger range of illumination conditions at different time of the day. In addition, in comparison to the UCSD dataset, the Mall dataset experiences more severe perspective distortion, which causes larger changes in size and appearance of objects at different depths of the scene, and has more frequent occlusion problem caused by the scene objects, e.g. stall, indoor plants along the walking path.

Settings – For the UCSD dataset, we followed the same training and testing partition as in [6], i.e. we employed Frames 601-1400 for training and the rest for testing. For the Mall dataset, we used the first 800 frames for training and kept the remaining 1200 frames for testing. Based on the resolution of the different datasets, we defined 6×4-cells for the UCSD dataset and 8×8-cells for the Mall dataset.

Evaluation Metrics – We employed three evaluation metrics, namely *mean absolute error* (mae), ε_{abs}; *mean squared error* (mse), ε_{sqr}; and *mean deviation error* (mde), ε_{dev}.

$$\varepsilon_{abs} = \frac{1}{N}\sum_{i=1}^{N}|v_i - \hat{v}_i|, \quad \varepsilon_{sqr} = \frac{1}{N}\sum_{i=1}^{N}(v_i - \hat{v}_i)^2, \quad \text{and} \quad \varepsilon_{dev} = \frac{1}{N}\sum_{i=1}^{N}\frac{|v_i - \hat{v}_i|}{v_i},$$

where N is the total number of test frames, v_i is the actual count in each cell region or the whole image, and $\hat{v}_i$ is the estimated count of ith frame.

Method	Features Level		Learning Level		UCSD [6]			Mall		
	Global	Local	Global	Local	mae	mse	mde	mae	mse	mde
RR [22]	✔	–	✔	–	2.25	**7.82**	0.1101	3.59	19.0	0.1109
GPR [6]	✔	–	✔	–	**2.24**	7.97	0.1126	3.72	20.1	0.1159
MLR [23]	–	✔	–	✔	2.60	10.1	0.1249	3.90	23.9	0.1196
MORR	–	✔	✔	–	2.29	8.08	**0.1088**	**3.15**	**15.7**	**0.0986**

Table 2: Performance comparison between different methods and our multi-output ridge regression (MORR) model on global crowd counting.

Comparative Evaluation – We compared the following models

- Single global model with global feature (1) ridge regression (RR) [22], and (2) Gaussian processes regression (GPR) with linear + RBF kernel as in [6]. These models employ global features as their input and the crowd density of the whole image as their output.
- Multiple localised regressors (MLR) [17]. The input and output for the model is the feature within each cell and the people count in the corresponding cell respectively. The ridge regression model is used to eliminate the effect of using different regression models.
- The proposed multi-output ridge regression (MORR) model described in Sec. 2.2.

For all models free parameters were tuned using 4-fold cross-validation.

Comparison With Single Global Regression Models – The results of different models on the two datasets are shown in Table 2. It can be observed the two global regression models, RR and GPR, yielded very similar results on UCSD compared to our MORR model, but much higher error rates on the more challenging Mall dataset, i.e., 16.03%, 24.52%, and 15.01% higher than our model in mae, mse, and mde on average. It is worth pointing out that in contrast to the other two metrics the mde is more indicative as it takes the level of crowdedness of ith frame into account.

Different performances on two datasets were due to the different characteristics of the two scenes. In the shopping mall scene different local regions can have drastically different lighting conditions (see Figure 3(b)). The different fixed structures in the scene (e.g. stalls and plants in the middle) also introduced different characteristics of occlusion. In comparison, in the UCSD campus scene, there was no occlusion caused by static objects and the lighting condition across the scene were fairly even and stable during the entire recording period.

The result thus suggests that mining features at different spatial location is more critical for a complex scene where lighting conditions are not uniform and can change quickly, and occlusion can occur both inter people and between people and static obstacles. It is also worth pointing out that despite its simpler formulation, the single global ridge regression model achieves comparable or better performance compared to the more complex Gaussian Processes Regression model.

Evaluation of Local Feature Mining of Our Model – The advantage of our multi-output framework for localised crowd counting is to mine local feature importance for supporting crowd density estimation. As Figure 4 shows, at different depths in a scene, certain types of features can play a more important role in estimating specific localised crowd density. Specifically, the plots in Figure 4 shows that the edge orientation 60° (mainly corresponds to shoulder edges) in the Away-from-Camera Cell 43 exhibited a higher importance as com-
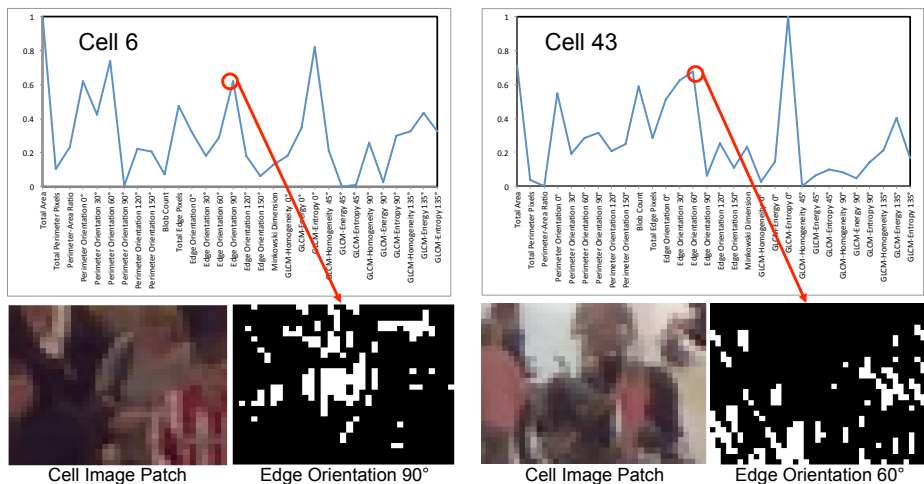
Figure 4: Local feature mining from one Close-to-Camera Cell 6 and one Away-from-Camera Cell 43 selected from the grid image in Figure 1. For each cell, we also show an example of image patch and together with the extracted edge at specific orientation. The horizontal axes of the two plots represent the features described in Section 2.1.
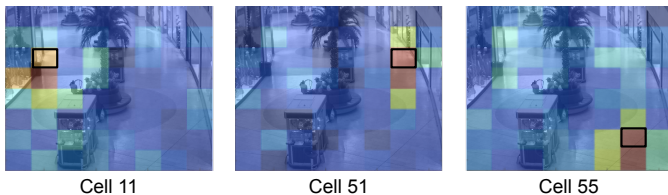

Figure 5: Using the Mall dataset as a study case: the figures depict the weight contributions of neighbouring cells to cells 11, 51, and 55, which are highlighted using black boxes (refer Figure 1 for cell index). Red colour in the heat maps represents a higher weight contribution i.e. more information sharing.

pared to other edge orientations; whilst the edge orientation $90°$ (mainly corresponds to the torso edges) in the Close-to-Camera Cell 6 was automatically assigned a higher weight. To provide a closer look on the different edge orientations at different cells, we depicted two example image patches and the associated edges extracted from cells 6 and 43 at the bottom of Figure 4. The results suggest that the weights learned using our model matched our intuition, i.e. different features, e.g. shoulder and torso edges, would have different importance to estimation at different depths of a scene.

Evaluation of Information Sharing Among Regions – To demonstrate that spatially localised regions in a scene can indeed share information with each other, we used the Mall dataset as a study case and selected three cells to profile how other neighbouring cells contributed to their count estimation. The degree of information sharing (or evidence support) can be quantified by summing the absolute weights of a neighbouring cell that contribute to the count of a cell that we are interested in. We repeated this step for all neighbouring cells

in the image space. The weight contribution/information sharing can be transformed into a heat map as shown in Figure 5. Evidently, the closer the neighbouring cells to the selected cell region, the more information were shared. This observation suggests that our model is capable of seeking evidence support from other cell regions to achieve a more accurate counting estimate.

Comparison With Multiple Localised Regression Model – As shown in Table 2, our multi-output regression model outperformed the multiple local regression model (MLR). It is interesting to note that the performance of the MLR is even worse than the two global regression models, although it was motivated to overcome the limitations of global regression models [23]. Since the MLR model also measures the importance of different features in different local regions as our model, this result highlights the importance of exploiting the correlation between features across regions and sharing information across regions to achieve more robust crowd counting. Without this information sharing, achieved by the multiple structure output regression model formulated in this paper, the local measures are too noisy and brittle to be relied upon in isolation for estimating density. Importantly, our single model based regression approach is more computationally scalable compared to the MLR model, e.g. compared to MLR, MORR is 3-5 times faster in both training and testing. More details about training and testing time for MLR and MORR are given in Table 3.

Analysis of Localised Counting Accuracy – To demonstrate the effectiveness of the proposed MORR in localised counting, we selected two busy regions (right in front of two shops) in the Mall dataset and compared the performance of MORR against MLR. The selected regions are depicted in a figure together with the results on localised crowd counting in Table 3. As compared to MLR, our MORR achieved more accurate localised counting, and yet faster training and testing time. The results again suggest the importance of information sharing among regions.

	Region 1(R1)				Scalability (seconds)	
	mae	mse	mde		Time-tr	Time-te
MLR	0.82	1.45	0.3611	MLR	17.274	0.1028
MORR	**0.76**	**1.22**	**0.3317**	MORR	14.848	0.0196

	Region 2 (R2)		
	mae	mse	mde
MLR	0.71	1.24	0.3317
MORR	**0.67**	**1.12**	**0.3061**

Table 3: Localised counting performance on two busy localised regions in the Mall dataset. Region 1 consists of Cells 11, 12, 19, and 20, while Region 2 includes Cells 43, 44, 51, and 52. Time-tr and Time-te denote the training time and testing time respectively.

4   Conclusion

We presented a single multi-output regression model capable of spatially localised crowd counting. Instead of building multiple localised regressors as adopted by existing techniques, our approach utilises a single joint regressor taking concatenated multiple localised imagery features as input for learning spatially localised crowd counts as multi-outputs. Our model outperforms multiple localised regressors on a challenging shopping mall dataset owing to its inbuilt ability for feature mining according to changing crowd conditions presented in different local spatial cell regions in the scene. On the other hand, it also compares favourably against existing single global regressor based crowd counting models. Extensive and com-

parative experimental results demonstrate the effectiveness of our method. Future work will focus on improving the performance of crowd density estimation by considering dynamic and temporal segmentation of crowd structure. This will facilitate the relaxation of any fixed size local cell region definition and permit learning more accurate crowd sensitive counting models reflecting dynamically the changing crowd structures in feature selection and representation.

References

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

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

[3] M.B. Blaschko and C.H. Lampert. Learning to localize objects with structured output regression. In *European Conference on Computer Vision*, pages 2–15, 2008.

[4] G. J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 594–601, 2006.

[5] A.B. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012.

[6] A.B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[7] S.Y. Cho, T.W.S. Chow, and C.T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(4):535–541, 1999.

[8] W. Ge and R.T. Collins. Marked point processes for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920, 2009.

[9] Y. Haitovsky. On multivariate ridge regression. *Biometrika*, 74(3):563–570, 1987.

[10] R.M. Haralick, K. Shanmugam, and I.H. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[11] D. Helbing, A. Johansson, and Eidgenössische Technische Hochschule. *Pedestrian, crowd and evacuation dynamics*. Encyclopedia of Complexity and Systems Science, 2009.

[12] C. Ionescu, L. Bo, and C. Sminchisescu. Structural SVM for visual localization and continuous state estimation. In *IEEE International Conference on Computer Vision*, pages 1157–1164, 2009.

[13] T. Joachims, T. Finley, and Chun-Nam Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[14] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *British Machine Vision Conference*, 2005.

[15] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, 2010.

[16] 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 *International Conference on Pattern Recognition*, pages 1–4, 2008.

[17] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *International Conference on Computer Sciences and Convergence Information Technology*, pages 170–175, 2010.

[18] A.N. Marana, S.A. Velastin, L.F. Costa, and R.A. Lotufo. Estimation of crowd density using image processing. In *Image Processing for Security Applications*, pages 11/1 – 11/8, 1997.

[19] A.N. Marana, L. da Fontoura Costa, RA Lotufo, and SA Velastin. Estimating crowd density with minkowski fractal dimension. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3521–3524. IEEE, 1999.

[20] V. Rabaud and S. Belongie. Counting crowded moving objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 705–711, 2006.

[21] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications*, 2009.

[22] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning*, pages 515–521, 1998.

[23] X. Wu, G. Liang, K.K. Lee, and Y. Xu. Crowd density estimation using texture analysis and learning. In *IEEE International Conference on Robotics and Biomimetics*, pages 214–219, 2006.

[24] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1198–1211, 2008.