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

Crowd counting in public places has a wide spectrum of applications especially in crowd control, public space design, and pedestrian behaviour profiling. Existing counting by regression methods, which aim to learn a direct mapping between low-level features and people count without segregation or tracking of individuals, can be categorised into either global approaches or local approaches. Global approaches [1, 3, 4] learn a single regression function between image features extracted globally from the entire image space and the total people count in that image space. 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. To overcome these limitations of a global approach, local models [5, 7] 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. 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.

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 [6], which takes inter-dependent local features from local spatial regions as input and people count from individual regions as multi-dimensional structured output. 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.

Figure 1 gives an overview of our framework: (Step-1) We first infer a perspective normalisation map using the method described in [2]. (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.

For a training video frame i, where $i = 1, 2 \dots N$ and N denotes the total number of training frames, we first partition the frame into K cell regions (see Step-3 in Figure 1). 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, $\mathbf{y}_i \in \mathbb{R}^m, i = 1, 2 \dots N$

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

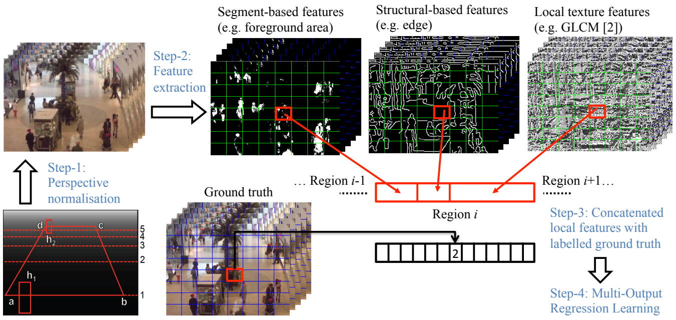

Figure 1: A multi-output regression framework for localised crowd counting by feature mining.

Let ($\mathbf{x}_i$, $\mathbf{y}_i$) be 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. Owing to its inbuilt ability for feature mining according to changing crowd conditions presented in different local spatial cell regions in the scene, our model outperforms multiple localised regressors and also compares favourably against existing single global regressor based crowd counting models on existing UCSD benchmark dataset and a new more challenging shopping mall dataset.

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.