An Improved Method of Crowd Counting Based on Regression

Mei Jiang[1] and Yanyun Zhao

Abstract An improved method of crowd counting based on regression is proposed to support intelligent management over crowd in video surveillance systems. According to the fact that human body has an articulate structure and complicated contours of shape, we propose a new low-level feature, the number of corner points, to highlight the describable capability of the feature set. We then introduce relevance vector regression (RVR) to model the correspondence between features and the pedestrian number, and propose a fusion scheme of RVR and Gaussian process regression (GPR) to further advance the performance of the proposed algorithm. Experimental results on two crowd datasets (one is UCSDpeds) demonstrate that the proposed work outperforms state-of-the-art methods and can fulfill the real-time requirement.

Keywords: Crowd counting • Feature extraction • Relevance vector regression (RVR) • Gaussian process regression (GPR)

1 Introduction

In public places, high density of crowd may be regarded as an indicator of congestions, threats or other abnormalities. Techniques based on analysis of video data can predict and estimate the size of crowd, which is of great significance to the security and surveillance community. Although intelligent crowd counting approaches have been studied in recent years, the precision of counting is yet not satisfactory. In this paper, we propose an improved method of crowd counting based on regression.

1.1 Related Work

Recently intelligent crowd counting has attracted researchers' attention in computer vision and related fields [1-9]. The existing predominant techniques for crowd counting fall into two categories: 1) object detection and tracking based crowd counting; 2) crowd density estimation based on features and regression analysis.

In the first category work always involves pedestrian detection and tracking. Since the counting results depend heavily upon detection response, applying a state-of-the-art detector has becoming an essential key to improve system performance. In [1-4] generic detectors including HOG based head-shoulder and JRoG based part detectors have been validated fairly effective in detecting and counting people. This kind of methods shows excellent performance in scenes with sparse crowd due to its accurate localization. However, as the crowd becomes large and heavy occlusion arises, individual detection and tracking both become almost impossible. Methods in the second category, such as [5-7], estimate the crowd density by extracting a set of holistic or local features in regions of interest (ROI) and then modeling the number of people based on features. This kind of methods does not consider individual localization and has been proved effective in dense crowd counting, especially in public environments where heavy occlusions happen frequently. In [8], the feature-regression based system is improved by applying human template matching. Antoni B. and Vasconcelos have shown in [9] that regression based crowd estimates are substantially more accurate than those produced by state-of-the-art pedestrian detectors. Thus they are more generally used in crowd counting systems.

Although methods of the second category are superior to the former one in heavily occluded scenes,

[1] Mei Jiang (✉)
Beijing University of Posts & Telecommunications, Beijing, China
e-mail: muyanlee@gmail.com

they still suffer from some obvious flaws, such as relatively low prediction accuracy and high computation time. Based on the framework in [6], we propose an improved crowd counting method in this paper.

There are 3 main contributions of this paper. 1) We propose a new low-level feature, the number of corner points, to highlight the relation between pedestrian number and the features. The unique and discriminative properties of corners make it effective in describing the density level of the crowd. 2) We introduce relevance vector regression (RVR) to model the correspondence between features and the number of people since it uses few relevance vectors but shows good generalization capability. 3) A fusion scheme of RVR and Gaussian process regression (GPR) is proposed to further improve the algorithm performance.

The paper is organized as follows. We briefly discuss the pre-processing steps of our method in Section 2. In Section 3, we present the details of feature extraction. We synopsize the principles of RVR and GPR and elaborate the training and predicting methods in Section 4. In Section 5 we validate our algorithm on two datasets and present the experimental results. We conclude this paper in Section 6.

1.2 Overview of Our Method

The flowchart of our proposed crowd counting method is illustrated in Figure 1. It consists of two main parts including training stage and testing stage. In both stages, the video is firstly segmented and perspective normalized, which are called pre-processing steps. Then some appearance features are extracted in ROI which is masked by the foreground. In the training, features and annotated pedestrian number are used to learn the parameters of regression model. During the testing, the features are fed into the learned model to estimate the pedestrian number.

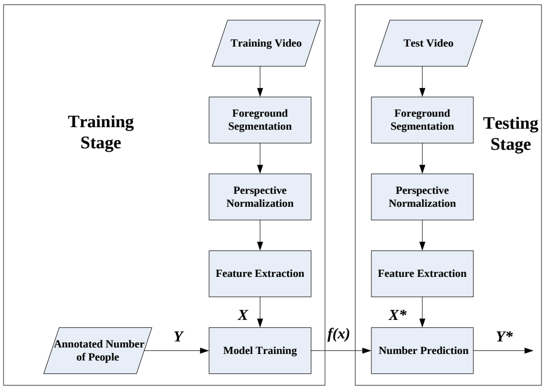

Fig. 1 Flowchart of the proposed method

2 Pre-processing

2.1 Foreground segmentation

We adopt the MID (Mosaic Image Difference) method proposed in [1] to segment the foreground. This method assumes that motion approximately satisfies temporal and spatial uniform distributions in a considerably long period of time since when and where it happens are completely random. We improve this algorithm by replacing the fixed threshold with an adaptive one calculated by Eq.(1). This change makes the method more robust than its original version and applicable to videos in different environments.

$$\begin{aligned} f_1 &= a + (I_t^1(x,y) + I_{t-1}^1(x,y))/b \\ f_2 &= a + (I_t^2(x,y) + I_{t-1}^2(x,y))/b \\ f_3 &= a + (I_t^3(x,y) + I_{t-1}^3(x,y))/b \\ TH_t(x,y) &= \max\{f_1, f_2, f_3\} \end{aligned} \tag{1}$$

where $I_t^i(x,y), i=1,2,3$, are intensity values of B, G and R channels at position (x,y) and time t respectively, and a and b are constants.

2.2 Perspective normalization

To take the perspective effect brought by static camera into account, a perspective map is calculated in the same way as [6]. Each pixel value in the perspective map is the weight of its corresponding pixel in original image. The feature is then rectified by the weight of pixel to accomplish perspective normalization.

As shown in Figure 2, a base line ab is plotted parallel to the bottom of the road in the frame and h_1 is the height of a pedestrian whose center is on the line. At the same time, another base line cd and a height h_2 are plotted and measured similarly. Given the weight of pixels on line ab by 1, the pixels on cd would have a weight of:

$$w(i,j)=\frac{|ab|h_1}{|cd|h_2} \tag{2}$$

The weights of pixels between the two lines can then be obtained by linear interpolation. For area-based features the weights are applied to each pixel, while for edge-based features the root square of the weight is used.

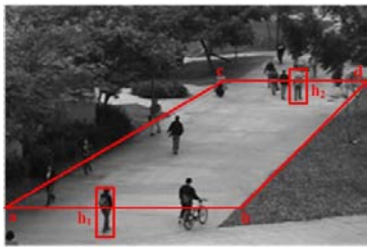

Fig. 2 Overview of perspective normalization

3 Feature Extraction

In regression based crowd counting methods, the features extracted must reflect the size of crowd to a great extent. So choosing good representative features is one of the keys to advance the performance of an algorithm. In this paper a carefully selected feature set is employed. We did not use the area feature which is widely used in many crowd counting methods, because it depends heavily on the results of foreground segmentation and performs unsteadily. More importantly, we propose a new feature, the number of corner points, to increase the representative power of the feature set. All features we extracted are as follows.

Edge. The total number of pixels on the edge, extracted by applying a Canny detector to the gray-level image.

EOH. The orientation histogram of pixels on the edge. A Sobel operator of aperture size 3 is applied to the image, and the amplitude and orientation of each pixel's gradient are given by:

$$G=\sqrt{{G_x}^2+{G_y}^2} \tag{3}$$

$$\theta=\arctan(G_y/G_x) \tag{4}$$

The orientation is in the range of $0^\circ \sim 180^\circ$ and uniformly quantized into 9 directions. The EOH is counted by the gradient intensity of each pixel on the edge.

Perimeter-Area Ratio. The ratio of foreground perimeter to area. This ratio is an effective measure of shape complexity and can indicate the number of people to some extent.

Texture. Marana *et al.* have validated that texture feature shows superior description when reflecting the density degree of crowd in image [10]. In this paper, we adopt a similar manner to measure texture features. The image is uniformly quantized into 8 levels, and then 4 grey level co-occurrence matrixes (GLCM) with distance $d = 1$ and angle $\theta = \{0°, 45°, 90°, 135°\}$ are computed respectively. Based on these matrixes, 3 parameters including energy, homogeneity and entropy are derived to describe the texture property in a parameterized way, resulting in a total 12-d feature vector.

Corners. The total count of corner points in ROI. A corner is deemed to be a unique and distinct point. There are numbers of corners shown in a human object since the human body has an articulate structure and complicated contours of shape. Due to its high representative capability, we propose to estimate the number of people by the corner feature. Figure 3 presents comparisons among corner maps of crowds with different densities. From left to right, as the density of crowd increases, the number of corners becomes correspondingly more plentiful and informative.

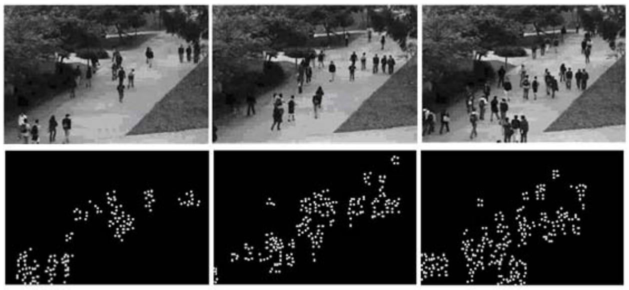

Fig. 3 Corner maps of crowds with different densities: the first row is original frames, the second row is corresponding corner maps

4 Regression Analysis

Regression analysis is a classical learning method for modelling the relation between obervation and continuous world state. In this paper, training samples consisting of features and annotated people numbers are used to build up the regression model and to estimate the model parameters in training, and then the learned model is used for people number prediction in testing. It can be described in a mathematical language as follows. Let x denote the feature, and w be the number of people. We need to estimate a function $f(x)$ so that $w = f(x)$. We focus on RVR and GPR models in this paper.

4.1 Relevance Vector Regression (RVR)

Relevance vector regression[11] is a sparse Bayesian learning model. It has several advantages over support vector regression (SVR): 1) Except for the estimated value of world state, RVR also gives the probability density distribution of the state; 2) The kernel function of RVR is not required to satisfy Mercers theorem so that it can be chosen from a wide range; 3) RVR uses few relevance vectors but shows superb generalization ability.

Let $\mathbf{x}$ denote the feature vector and w is the corresponding people number. As w is resulted from co-effect of multiple independent factors, it is supposed to satisfy the univariate normal distribution in Eq. (5) based on central limit theorem.

$$P_r(w) = \mathrm{Norm}_w[\mu_w, \sigma_w^2] \tag{5}$$

where $\mu_w = \boldsymbol{\psi}^T \mathbf{x}^T \mathbf{x}$ is a function of μ_w, variance σ_w^2 is a constant.

To encourage sparsity over training samples we choose the dual perimeter $\boldsymbol{\psi}$ as the form of a product of one dimension t-distributions as Eq. (6):

$$P_r(\boldsymbol{\psi}) = \prod_{i=1}^{I} \mathrm{Stud}_{\psi_i}[0, 1, \nu] \tag{6}$$

Let $\mathbf{X}$ denotes the whole training feature matrix and $\mathbf{w}$ is the corresponding people number vector. Given a new feature vector $\mathbf{x}^*$, the predicted value w^* follows the distribution as:

$$P_r(w^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{w}) = \mathrm{Norm}_{w^*}[\mu_{w^*|x^*}, \sigma^2_{w^*|x^*}] \tag{7}$$

$$\mu_{w^*|x^*} = \mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{A}^{-1}\mathbf{K}[\mathbf{X}, \mathbf{X}]\mathbf{w} / \sigma^2 \tag{8}$$

$$\sigma^2_{w^*|x^*} = \mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{A}^{-1}\mathbf{K}[\mathbf{X}, \mathbf{x}^*] + \sigma^2 \tag{9}$$

$$\mathbf{A} = \mathbf{K}[\mathbf{X}, \mathbf{X}]\mathbf{K}[\mathbf{X}, \mathbf{X}] / \sigma^2 + \mathbf{H} \tag{10}$$

$\mathbf{H}$ is a diagonal matrix whose elements are hidden variables. $\mathbf{K}[\mathbf{X}, \mathbf{X}]$ represents a matrix of dot products where element (i, j) is given by $k[x_i, x_j]$ .This is a kernel function that efficiently computes the inner product of each pair of data embedded in some high-dimension feature space. Commonly used kernel functions include:

•Linear

$$k[x_i, x_j] = x_i^T x_j \tag{11}$$

•Degree p polynomial

$$k[x_i, x_j] = (x_i^T x_j + 1)^p \tag{12}$$

•Radial basis function (RBF)

$$k[x_i, x_j] = \exp[-0.5(\frac{(x_i - x_j)^T (x_i - x_j)}{\lambda^2})] \tag{13}$$

Since the precisions of fitting and prediction of the regression model depend much on the type of kernel used, it is vital to choose a proper kernel function. In this paper, we have tried three kinds of kernel functions: linear, RBF and a mixed kernel similar to [6]. The mixed kernel combines linear and RBF kernels as shown in Eq. (14), aiming to capture both linear and non-linear relations between data and states.

$$k[x_i, x_j] = \alpha_1 + \alpha_2 x_i^T x_j + \alpha_3 \exp[-\alpha_4 (\frac{(x_i - x_j)^T (x_i - x_j)}{\lambda^2})] \tag{14}$$

where $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \lambda\}$ are hyperparameters.

Figure 4 shows the correspondences between estimates of RVR models with different kernels and the normalized number of corners. It is obvious that their relation follows a linear trend roughly while the data fluctuate nonlinearly due to occlusion, low resolution and observation noise etc. The model with linear kernel captures the main trend well and is capable of extrapolation, but its prediction could be sensitive to outlier data which may affect the gradient of the line. The RBF kernel allows the slight fluctuation of the data points, but it shows low generalization ability since the estimate is far from ideal when there comes a new data point that hasn't appeared in training. The drawbacks of the two types of kernels are overcome by applying a combination of them. Our later experimental results also reveal the superiority of the mixed kernel.

4.2 Gaussian Process Regression (GPR)

Gaussian process regression[11] is a nonlinear regression using Bayesian method. Differing from RVR, it doesn't consider the sparsity over training samples. The simple training process and wide kinds of available kernels make it of high value in machine learning.

Similarly, suppose the prior over w is univariate normal distribution and its mean $\mu_w = \boldsymbol{\varphi}^T\mathbf{x}$ is a function of $\mathbf{x}$. Since GPR doesn't introduce sparsity on training samples, the gradient vector $\boldsymbol{\varphi}$ has a distribution of normal with zero mean and spherical covariance matrix as shown in Eq. (15):

$$P_r(\boldsymbol{\varphi}) = \text{Norm}_{\boldsymbol{\varphi}}[\mathbf{0}, \sigma_p^2 \mathbf{I}] \tag{15}$$

Let $\mathbf{X}$ denotes the whole training feature matrix and $\mathbf{w}$ is the corresponding people number vector. Given a new feature vector $\mathbf{x}^*$, the predicted value w^* follows the distribution as:

$$P_r(w^* | \mathbf{x}^*, \mathbf{X}, \mathbf{w}) = \text{Norm}_{w^*}[\mu_{w^*|x^*}, \sigma^2_{w^*|x^*}] \tag{16}$$

$$\mu_{w^*|x^*} = (\sigma_p^2 / \sigma^2)\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{w} - (\sigma^2 / \sigma_p^2)\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{A}^{-1}\mathbf{K}[\mathbf{X}, \mathbf{X}]\mathbf{w} \tag{17}$$

$$\sigma^2_{w^*|x^*} = \sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{x}^*] - \sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{A}^{-1}\mathbf{K}[\mathbf{X}, \mathbf{x}^*] + \sigma^2 \tag{18}$$

$$\mathbf{A} = \mathbf{K}[\mathbf{X}, \mathbf{X}] + (\sigma^2 / \sigma_p^2)\mathbf{I} \tag{19}$$

The three types of kernels mentioned in 4.1 can also be used in GPR to give the estimate of pedestrian number.

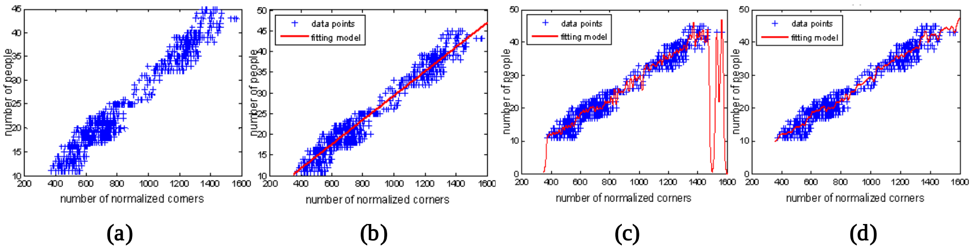

Fig. 4 RVR model for corners: a) training data points; b) RVR model with linear kernel; c) RVR model with RBF model; d) RVR model with mixed kernel

4.3 Fusion of RVR and GPR

For the probability distribution based regression method, we refine the mean value by the variance in the following manner:

$$w = \begin{cases} \mu_{w^*|x^*} & \text{if } \sigma^2_{w^*|x^*} < thr \\ \mu_{w^*|x^*} - d & \text{otherwise} \end{cases} \tag{20}$$

where thr and d are threshold and constant respectively. This treatment is based on the hypothesis that the estimate with higher probability variance would be less convincible, and the refinement of Eq.(20) indeed improves the performance of our algorithm effectively.

In our study, we have tried Bayesian linear regression (BLR), RVR, GPR and SVR for mapping the features into the number of people. As RVR and GPR show better performances among them in aspects of accuracy and generalization in regression and prediction, we use a weighted sum to fuse the estimates of the two models to give the final counting result:

$$w = \alpha_1 w_r + \alpha_2 w_g \tag{21}$$

where w_r and w_g are the people numbers estimated by RVR and GPR respectively, α_1 and α_2 are coefficients that determine the contribution ratio of the two models.

5 Experiments and Discussions

5.1 Crowd Counting Datasets

We use two crowd datasets in experiments. Dataset 1 is UCSDpeds video [6] which was collected from a stationary digital camcorder overlooking a pedestrian walkway at UCSD as illustrated in Figure 5a. The video has a length of 2000 frames. To make comparisons with the state-of-the-art methods, we use 800 frames (601-1400) for training and the remaining for testing as most previous works did [6,12,13].

Dataset 2 is collected by us from an outdoor digital camera in the campus of BUPT as showed in Figure 5b. This dataset contains two clips of videos. The one of 1600 frames is used for training and another of 4200 frames is for testing.

In order to avoid the result with fast fluctuations, we apply an averaging filter of length $2n+1$ to smooth the raw estimates. The filter returns the average values of estimates in a fixed period. This is implemented by using a data buffer with a fixed size.

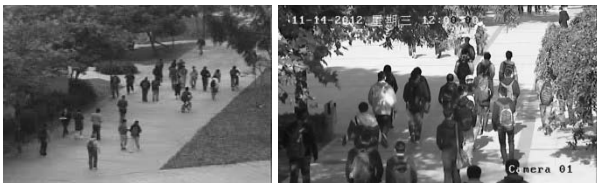

Fig. 5 Screenshots of crowd videos: a) Dataset 1; b) Dataset 2

5.2 Crowd Counting Results

We have tried different kinds of regression methods in training and the crowd estimates and ground-truth were recorded. Evaluation metrics include average absolute error between estimates and ground-truth (Err) and mean squared error (MSE). We also compare our results with the state-of-the-art works on Dataset 1. Table 1 shows the comparison results. It is shown that:

1) For both RVR and GPR models, the Err and MSE of results estimated by mixed kernel are substantially lower than that generated by a single linear kernel or RBF kernel, which demonstrates that the regression with mixed kernel can make a more precise fitting over training data and produce better predictions in test.

2) Our best result given by (RVR+GPR) method has the lowermost Err and MSE and exceeds the state-of-the-art works, validating the high accuracy of our proposed crowd counting algorithm.

The Err and MSE of the result given by (RVR+GPR) method on Dataset 2 is 1.084 and 2.431 respectively, demonstrating the high generalization of out method.

Figure 6 gives the comparisons between ground-truth and estimated results using (RVR+GPR) method on the two datasets.

5.3 Processing Speeds

Our experiments are conducted on an Intel (R) Core (TM) computer with CPU 3.30GHz, 3.16 GB memory, Windows XP OS, Microsoft Visual Studio 2005 and OpenCV 2.0 library. The average processing speeds on Dataset 1 using different regression methods are presented in Table 2.

The algorithm using (RVR+GPR) model runs at a speed of 33.17 ms per frame, demonstrating that our method can fulfill the real-time requirement and is of great practicality.

6 Conclusions

This study proposes an improved method of crowd counting based on feature and regression. By introducing the number of corners as a feature and combining RVR with GPR models, we validate that our algorithm outperforms most of state-of-the-art works at a real-time processing speed. Thus the proposed method is of great practical value for various public environments.

Table 1 Crowd counting results: Err and MSE on Dataset 1 using different regression methods

Work	Regression Method	Err	MSE
Proposed	BLR	1.692	4.163
Proposed	RVR(linear kernel)	1.509	3.846
Proposed	RVR(RBF kernel)	1.447	3.349
Proposed	RVR(mixed kernel)	1.393	3.128
Proposed	GPR(linear kernel)	1.811	5.342
Proposed	GPR(RBF kernel)	1.566	3.772
Proposed	GPR(mixed kernel)	1.561	3.694
Proposed	SVM(linear kernel)	1.646	4.294
Proposed	SVM(RBF kernel)	2.032	6.083
Proposed	RVR+GPR(mixed kernel)	**1.343**	**3.008**
[6]	GPR(mixed kernel)	.>1.621	>4.181
[12]	KDR(linear kernel)	-	4.817
[13]	Linear	1.353	3.065

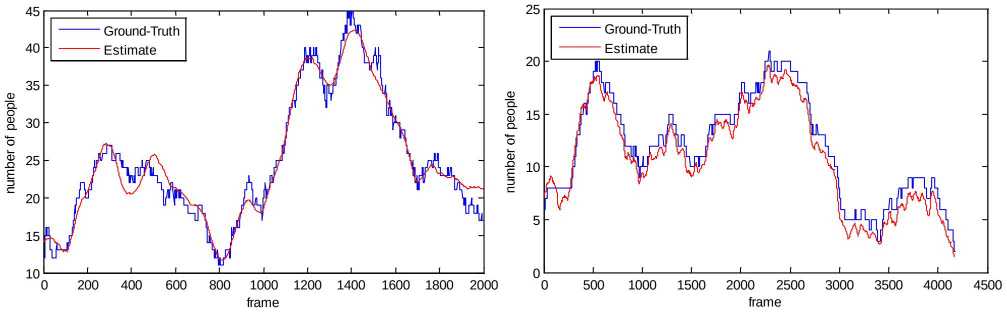

Fig. 6 Comparisons between ground-truth and estimated results on two datasets: a) Dataset1; b) Dataset 2

Table 2 Average processing speeds using different regression methods

Regression Method	Average Processing Speed(ms/frame)
RVR	27.92
GPR	23.31
RVR+GPR	33.17

Acknowledgements This work is funded by Chinese National Natural Science Foundation (90920001, 61101212), National High and Key Technology Research and Development Program (2012AA012505, 2012BAH63F00), National S&T Major Project of the Ministry of S&T (2012ZX03005008), the Special Funds of Beijing Municipal Co-construction Project, and the Fundamental Research Funds for the Central Universities.

References

1. Li, M., Z. Zhang, K.,Huang,et al. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection, ICPR 2008. Tampa: IEEE Press, 2008: 1-4.
2. Subburaman, V.B., Descamps, A.,Carincotte, C. Counting people in the crowd using a generic head detector, AVSS 2012. Beijing: IEEE Press, 2012: 470-475.
3. Nguyen, D., L. Huynh, T.B., Dinh, et al. Video monitoring system: counting people by tracking, RIVF 2012. Vietnam: IEEE Press, 2012: 1-4.
4. Rodriguez, M., I. Laptev, J. Sivic, et al. Density aware person detection and tracking in crowds, ICCV 2011. Barcelona: IEEE Press, 2011: 2423-2430.
5. Kong, D., D. Gray, H. Tao. A viewpoint invariant approach for crowd counting, ICPR 2006. Hong Kong: IEEE Press, 2006: 1187-1190.
6. Chan, A.B., Z.S.J. Liang, N. Vasconcelos. Privacy preserving crowd monitoring: counting people without people models or tracking, CVPR 2008. Anchorage: IEEE Press, 2008: 1-7.
7. Shimosaka, M., S. Masuda, R. Fukui,et al. Counting pedestrian in crowded scenes with efficient sparse learning, ACPR 2011. Beijing: IEEE Press, 2011: 27-31.
8. Li, J., L. Huang, C. Liu. Robust people counting in video surveillance: dataset and system, AVSS 2011.Klagenfurt: IEEE Press, 2011: 54-59.
9. Chan, A.B., N. Vasconcelos. Counting people with low-level features and Bayesian regression. IEEE Trans on Image Processing, 2012, 21: 2160-2177.
10. Marana, AN, SA Velastin, LF Costa, et al. Estimation of crowd density using image processing. IEE Colloquium on Image Processing for Security Applications, 1997, 74: 11/1-11/8.
11. Simon J.D. Prince. Computer vision: models, learning and inference. London: Cambridge University, 2012: 150-159.
12. Zhang, J., B. Tan, F. Sha,et al. Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. IEEE Trans on Intelligent Transportation Systems, 2011, 12: 1037-1046.
13. Ryan, D., S. Denman, C. Fookes, et al. Crowd counting using multiple local features, DICTA 2009. Melbourne:IEEE Press, 2009: 81-88.