Indoor Crowd Counting by Mixture of Gaussians Label Distribution Learning

Miaogen Ling and Xin Geng, *Member, IEEE*

***Abstract*— In this paper, we tackle the problem of crowd counting in indoor videos, where people often stay almost static for a long time. The label distribution, which covers a certain number of crowd counting labels, representing the degree to which each label describes the video frame, is previously adopted to model the label ambiguity of the crowd number. However, since the label ambiguity is significantly affected by the crowd number of the scene, we initialize the label distribution of each frame by the discretized Gaussian distribution with adaptive variance instead of the original single static Gaussian distribution. Moreover, considering the gradual change of crowd numbers in the adjacent frames, a mixture of Gaussian models is proposed to generate the final label distribution representation for each frame. The weights of the Gaussian models rely on the frame and feature distances between the current frame and the adjacent frames. The mixed $\ell_{2,1}$-norm is adopted to restrict the weights of predicting the adjacent crowd numbers to be locally correlated. We collect three new indoor video datasets with frame number annotation for further research. The proposed approach achieves state-of-the-art performance on seven challenging indoor videos and cross-scene experiments.**

***Index Terms*— Label ambiguity, label distribution learning, mixture of Gaussians model.**

I. Introduction

CROWD counting in videos draws a lot of attention for its intense demand in video surveillance and urban security system. Most existing approaches pay more attention to outdoor scenes where people are moving for most of the time. However, in some indoor scenes, people often stay almost static for a long time (Fig. 1(a)). Sometimes people are severely occluded with each other and the background could hardly be seen (Fig. 1(b)). We compare the difficulties and easinesses of the typical indoor and outdoor crowd counting scenes in Table I. In an indoor scene, people usually gather together to perform a certain task in a constrained space. The position of the camera is usually restricted within the scene. Therefore, the indoor scenes are often close to the camera, crowded with people of little movement and have heavy occlusion from both people and obstacles, leading to slow crowd number variation. In contrast, people in outdoor scenes usually move in an open space. The outdoor scenes are mostly far from the camera and the crowd is sparse with large movement and has light occlusion, leading to fast crowd number variation. As shown in Fig. 1(d), the crowd number varies all the time as people go in and out of the scene and thus it causes fast crowd number variation. In fact, indoor and outdoor scenes are a relative concept. According to the characteristics of the indoor and outdoor scenes in Table I, an outdoor shop with people standing in lines could be regarded as an indoor scene, while an indoor corridor or passageway with people moving fast would actually be regarded as an outdoor scene, as shown in Fig. 1(c). Indoor scenes are usually more crowded than outdoor scenes. However, some outdoor scenes might also be crowded and have heavy occlusion (ShanghaiTech [1]). In this paper, we focus on the scenes which satisfy characteristics of indoor scenes in Table I.

(a) (b) (c) (d)

Fig. 1. Indoor scenes: students in classroom3_1 (a) and passengers on a bus (b). Outdoor scenes: people in the mall [2] (c) and crowds in the walkway [3] (d).

In the above indoor environments, the performance of most traditional outdoor foreground extraction methods would drop significantly since the stationary people are often missed due to moving foreground segmentation. Therefore, the foreground extraction method specialized for indoor scenes should be adopted. After the foreground extraction, the head [4], [5] or head-shoulder detection models [6], [7] are usually proposed to detect people for indoor crowd counting. However, these methods suffer when heads or shoulders of people are too crowded to be seen and the crowd counting accuracy relies much on foreground extraction. In contrast, constructing regression or classification models for the extracted foreground features would be more robust as they do not need to detect

Manuscript received August 22, 2018; revised May 2, 2019 and June 10, 2019; accepted June 10, 2019. Date of publication June 19, 2019; date of current version August 28, 2019. This work was supported in part by the National Key Research & Development Plan of China under Grant 2017YFB1002801, in part by the National Science Foundation of China under Grant 61622203, in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, and in part by the Collaborative Innovation Center of Wireless Communications Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mireille Boutin. *(Corresponding author: Xin Geng.)*

M. Ling is with the MOE Key Laboratory of Computer Network and Information Integration, School of Computer Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: mgling@seu.edu.cn).

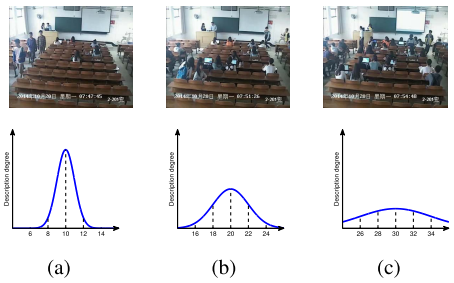
X. Geng is with the College of Information Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China, and also with the School of Computer Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: xgeng@seu.edu.cn).

Digital Object Identifier 10.1109/TIP.2019.2922818

TABLE I

THE DIFFICULTIES AND EASINESSES OF THE TYPICAL INDOOR AND OUTDOOR CROWD COUNTING SCENES

Scene	Difficulties	Easinesses
Indoor	• crowded • heavy occusion • little movement	• close to the camera • slow crowd number variation
Outdoor	• far from the camera • fast crowd number variation	• sparse • light occusion • large movement



(a) (b) (c)

Fig. 2. Single Gaussian label distributions for different crowd numbers in classroom1: 10 people (a), 20 people (b) and 30 people (c). Horizontal axis stands for crowd numbers and vertical axis stands for description degrees.

complete heads or shoulders and they remain valid for very crowded situations, such as in Fig. 1(b).

In the traditional regression or classification based crowd counting methods, only one crowd number label is adopted for each frame in the video. Considering label ambiguity of the crowd number, the label distribution [8] is proposed to cover a certain number of crowd number labels, representing the degree to which each label describes the frame. The description degrees of all labels sum up to 1. Some examples of label distribution representations for video frames are shown in Fig. 2. In this way, a frame instance will contribute to not only the learning of its real crowd number, but also the learning of its neighboring crowd numbers. Hence, the training data is increased significantly and the classes with insufficient training samples are supplied with more training data, which belongs to their adjacent crowd numbers. Label distribution learning (LDL) is confirmed to achieve good performance in outdoor crowd counting [8]. It has also been successfully applied to many other fields, such as prediction of crowd opinion on movies [9], video parsing [10], head pose estimation [11], age estimation [12], [13], sense facial beauty [14] and so on.

However, the label distribution for each frame is not available in the original training data and must be generated with proper assumption. For indoor scenes, we notice that the label ambiguity is significantly affected by crowd number of the frame. When there are a few people as shown in Fig. 2(a), the frame is easy to be labeled accurately and the label ambiguity is low. However, when people are occluded with each other in a crowded scene as shown in Fig. 2(c), the label ambiguity would be rather high. Accordingly, the label distribution generation should accord with such crowd number variances. If the discretized Gaussian distribution centered at the ground truth crowd number is adopted to generate the label distribution, the adaptive standard deviations should be adopted for different crowd numbers. Moreover, the label ambiguity for each frame is not only affected by its crowd number but also crowd number variances of the adjacent frames. For example, we compare two circumstances where crowd numbers of several continuous frames are $\{19, 20, 21, 22, 23\}$ and $\{21, 21, 21, 21, 21\}$. The crowd numbers of the third frames are both 21 but label ambiguity degrees are different. In the first case, the crowd number keeps changing in the sequence, which is probably caused by people entering the spot constantly. The uncertainty of crowd number label of the third frame significantly increases as crowd number varies in its adjacent frames. On the contrary, in the second case the crowd number keeps unchanged around the third frame. Thus, the crowd number label would be more accurate and label ambiguity degree is low. Considering the above reasons, we adopt a mixture of Gaussians model to generate the label distribution for each frame, where each Gaussian model is a discretized Gaussian distribution with adaptive variance. The weights of Gaussian models rely on the frame and feature distances between the current frame and the adjacent frames.

The main contributions of this paper include:

1) It is one of the first times the classification-based method is adopted for indoor crowd counting.
2) Instead of the single static Gaussian label distribution, the adaptive mixture of Gaussians label distribution is generated for each frame.
3) The mixed $\ell_{2,1}$ norm restricts the weights of our label distribution learning model to be locally correlated, which accords with the graduality of the crowd number in indoor scenes.
4) Three new indoor crowd counting datasets are collected and annotated frame by frame. They would be publicly available to the research community in the future.

The rest of this paper is organized as follows. Section II reviews some related works. After that, a mixture of Gaussians label distribution learning model is proposed in Section III. Then, experimental results on seven challenging indoor videos and the cross-scene analysis are reported in Section IV. Finally, conclusions and future works are given in Section V.

II. RELATED WORK

A. Indoor Crowd Counting

At an earlier time, a graph cut method [15] is proposed to cope with indoor crowd counting. Each image of the video is preprocessed by image intensity normalization and the frame differences of the adjacent normalized images are put into the autoregressive model to extract the foreground blobs. The graph constructed by the foreground blobs is partitioned into clusters based on graph cut. The method would fail if people are overlapped with each other in crowded situations. Most of the recent methods on indoor crowd counting are based on head [4], [5] or head-shoulder detectors [6], [7]. For example, a cascaded head detector [5] is trained by manually annotated positive and negative samples. Sub-windows around the interest points of the image are classified as head or no-head region using the trained classifier. Finally a multi-instance pruning is applied to obtain final crowd numbers estimations. However, it is prone to produce false detections for complex background. Therefore, head-shoulder detectors [6], [7] are

proposed to generate more reliable results. For example, a multi-view head-shoulder detection method [6] normalizes all annotated head-shoulder samples to a fixed grid-model, where the blocks in the grid-model with more edge points are selected as key points and assigned different weights. The head-shoulder candidates are extracted by matching the trained grid-model with foreground gradient map by sliding. Finally the crowd number is obtained by K-means clustering for all head and head-shoulder candidates. However, it suffers when heads or shoulders of people are severely occluded and crowd counting accuracy relies on foreground extraction results.

B. Outdoor Crowd Counting

Outdoor crowd counting methods could be broadly divided into two categories: traditional regression-based methods and deep learning-based methods. For most regression-based methods [2], [16], [17], handcrafted features are extracted from the foreground blobs and different regression models are adopted for crowd number estimation. Gaussian process regression (GPR) [18] is one of the most popular nonlinear methods for crowd counting, since it allows any number of basis functions for different data complexity. Based on the conventional GPR, various extended approaches have been proposed. For example, Poisson regression [16] models the counting output as a Poisson random variable, where the mean parameter is a function of the input feature. The Poisson regression model is analyzed in a Bayesian setting by adding a Gaussian prior on the weights of the linear log-mean function and solved by a closed-form Bayesian inference. Moreover, a generalized Gaussian process model [19] based on the single-parameter exponential family distribution is proposed for crowd counting. New Gaussian process models can be created for different computer vision tasks by simply applying different likelihood function through its parameterization. Ridge regression (RR) [20] is another useful tool for crowd counting for its superior robustness in coping with multi-collinearity problem [21]. The cumulative attribute based ridge regression (CA-RR) [2] first projects the low level foreground features onto a cumulative attribute space as a intermediate representation, which captures how the crowd number changes continuously and cumulatively. After that, the ridge regression is adopted to learn the mapping from cumulative attribute space to crowd counting. All the regression-based methods can be adopted in indoor scenes once the features of the video frames are obtained after indoor foreground extraction.

The deep learning-based methods [22]–[24] are mostly based on density map estimation, where each person in the training images should be annotated and various deep neural networks are adopted. For example, multi-column convolutional neural network (MCNN) [22] uses three parallel CNNs with different filter sizes to model the crowd density at different scales. For perspective distortion, a geometry-adaptive kernel is adopted to generate the density map instead of the traditional Gaussian kernel. The output feature maps of the three CNNs are averagely fused by 1×1 convolution. However, congested scene recognition network (CSRNet) [24] notices that the branch structure of MCNN [22] dose not perform better than a deeper network with smaller convolutional filters. They deploy first 10 layers of VGG-16 network [25] as the front-end and dilated convolutional layers as the back-end to deliver larger receptive fields. Thus, CSRNet is easy-trained for its pure convolutional structure. Furthermore, body part map [23] is proposed to generate the structured density map. A pre-trained deep neural network is first employed to parse each pedestrian image into several body parts. After that, a multi-task crowd counting framework including density map, body part map and crowd count estimation is adopted to learn the model. However, since all indoor videos in our experiments are annotated by crowd number instead of dotted head annotations, the density map based methods cannot be directly applied and their loss function should be adapted to fit our datasets.

Moreover, we notice that some of the deep learning-based methods directly treat the crowd number as the output. In [26], the pretrained AlexNet is adapted by replacing the output of the last fully-connected layer with 1 item that means the predicted crowd number. Then pre-trained parameters of five convolutional layers and two fully-connected layers are adopted and weights for the last layer are fine-tuned by adding some collected samples with no person. Moreover, an end-to-end crowd counting method [27] uses the pre-trained deep network for features generation and features are decoded to the local counts by the long-short time memory (LSTM) [4]. Then two fully connected layers are adopted to generate the global count of the image. However, the local count corresponding to part of the image is not available in our indoor videos.

III. Indoor Crowd Counting by Mixture of Gaussians LDL

A. Foreground Extraction for Indoor Conditions

In indoor scenes, people often stay almost static for a long period of time. Therefore, most of the traditional outdoor foreground extraction methods would fail since the stationary people would mostly be merged into the background. However, body part movements of short duration still occur constantly in the video. Accordingly, the motion accumulation method [15] is proposed to deal with indoor foreground extraction. The motion accumulator is able to integrate, both spatially and temporally, the motion cues that are found by frame to frame differencing. Each frame in the video is firstly converted to the gray image and a Gaussian filter is applied to remove noise. Then each pixel in the image is divided by the maximum value in the $R\times R$ neighborhoods to make the image more robust to illumination changes [28]. After that, an autoregressive (AR) model is adopted to accumulate the frame difference pixels to form the foreground. Let $g(x, y, t)$ denotes the AR filtered output corresponding to the pixel at location (x, y) in the t-th frame, which is calculated as

$$g(x,y,t) = C_{000}\cdot d(x,y,t) + \sum_{u=1}^{T}\sum_{(i,j)\in R} C_{ijk}\cdot g(x+i, y+j, t-u), \quad (1)$$

where C_{ijk} are AR coefficients, $d(x, y, t)$ denotes the frame difference of the normalized image at location (x, y) in the

(a) (b) (c) (d)

Fig. 3. Perspective normalization on 4 indoor scenes: classroom3_2 (a), canteen2 (b) and bus (c) and a scene without the vanishing lines (d).

t-th frame, T is the temporal order of the AR process and R represents the spatial order in terms of the local neighborhood considered. Finally, the pixels with $g(x, y, t)$ larger than a threshold are judged as foreground. However, as AR always combines the current and historical movements of people to extract foreground, foreground pixels of the current frame would sometimes be lost. Therefore, three-frame difference method [28] is adopted to preserve the foreground area for the current frame. In our experiments, we combine the result of AR and three-frame difference method to be the foreground.

B. Perspective Normalization

We extract 29 features from the foreground of each frame the same as in [8], including the foreground segment features, the edge features and the texture features. The effects of perspective cannot be neglected for the features. Objects closer to the camera usually appear larger in frames. Thus, the features should be normalized. First, the ground plane in each scene is first marked by two vanishing lines $\overline{ac}$ and $\overline{bd}$ as shown in Fig. 3. Lengths of $\overline{ab}$ and $\overline{cd}$ are manually measured. The length of any line $\overline{ef}$ parallel to $\overline{ab}$ can be computed by the linear interpolation of $|\overline{ab}|$ and $|\overline{cd}|$, where $|\cdot|$ means length of the line. For classroom scenes in Fig. 3(a), the weights of pixels above the line $\overline{cd}$ are all assigned as 1 since the area of the front wall and the blackboard are vertical to the ground plane and no perspective normalization should be considered. Meanwhile, any pixel on the line $\overline{ef}$ below the line $\overline{cd}$ is given the weight of $|\overline{cd}|/|\overline{ef}|$. For canteen scenes in Fig. 3(b), the weights of pixels on the first row of the image $\overline{cd}$ are set to 1 and the weights of the rest pixels on the line $\overline{ef}$ are determined by $|\overline{cd}|/|\overline{ef}|$. For the bus scene in Fig. 3(c), the weights of pixels on the line $\overline{ab}$ are set to 1, the weights of pixels above the line $\overline{ab}$ are determined by $|\overline{ab}|/|\overline{ef}|$ and the weights of pixels below the line $\overline{ab}$ are determined by $|\overline{ij}|/|\overline{gh}|$, where lengths of $\overline{ab}$, $\overline{cd}$, $\overline{gh}$ and $\overline{kl}$ are manually measured.

Sometimes, the vanishing lines may not exist in the videos, we could still perform the perspective normalization as follows if the perspective map is linearly fitted along the vertical axis (as is often the case). For example in Fig. 3(d), we find one person close to the bottom line of the image and the other one far from it. The rectangular bounding boxes of the two persons could be obtained by annotations of widths and heights of them. The size of the two persons at the center points of the bounding boxes are estimated as square roots of bounding box areas. The size of person at any point in the image can be calculated as linear interpolation of the two persons' sizes along the vertical axis and the weight at that point would be reciprocal of the person's size.

In consideration of the perspective normalization, the features generated by pixels in the frames are multiplied by their corresponding weights. The area feature of the foreground segment is weighted by the square of the original pixel weight as it changes quadratically whilst the other features are weighted by the original pixel weights as they change linearly.

C. Label Distribution Learning With Mixed $\ell_{2,1}$ Norm

Let $\mathcal{X}$ denote the video frames, $\mathcal{Y} = \{y_1, y_2, \cdots, y_C\}$ denote the C crowd number labels. Given a training set $S = \{(\boldsymbol{x}_1, \hat{\boldsymbol{d}}_1), (\boldsymbol{x}_2, \hat{\boldsymbol{d}}_2), \cdots, (\boldsymbol{x}_n, \hat{\boldsymbol{d}}_n)\}$, where $\boldsymbol{x}_i \in \mathcal{X}$, $\hat{\boldsymbol{d}}_i = [\hat{d}_{\boldsymbol{x}_i}^{y_1}, \hat{d}_{\boldsymbol{x}_i}^{y_2}, \cdots, \hat{d}_{\boldsymbol{x}_i}^{y_C}]$. As mentioned before, the label distribution for each training frame is not available in the original training data and must be generated with proper assumption. Here, the original label distribution of the training frame is first initialized by a single discretized Gaussian distribution centered at the ground truth crowd number. In this paper, we denote the single Gaussian label distribution as $d_{y_j,\boldsymbol{x}_i}^k$ and denote the Mixture of Gaussians label distribution as $\hat{d}_{y_j,\boldsymbol{x}_i}^k$ for frame $\boldsymbol{x}_i$ in the k-th iteration. After initialization, each training frame $\boldsymbol{x}_i$ with crowd number α_i is assigned an initial label distribution $\hat{\boldsymbol{d}}_i^0 = [\hat{d}_{y_1,\boldsymbol{x}_i}^0, \hat{d}_{y_2,\boldsymbol{x}_i}^0, \cdots, \hat{d}_{y_C,\boldsymbol{x}_i}^0]$ which is calculated as

$$\hat{d}_{y_j,\boldsymbol{x}_i}^0 = d_{y_j,\boldsymbol{x}_i}^0 = \frac{1}{\sigma_{\alpha_i}^0 \sqrt{2\pi} Z} \exp\left(-\frac{(y-\alpha_i)^2}{2(\sigma_{\alpha_i}^0)^2}\right), \quad j = 1, 2, \cdots, C, \tag{2}$$

where $\sigma_{\alpha_i}^0$ is the initial standard deviation for crowd number α_i, and Z is a normalization factor that ensures $\sum_y d_{y,\boldsymbol{x}_i}^0 = 1$.

The goal of label distribution learning is to find the parameter vector $\boldsymbol{\Theta}$ in a conditional mass function $p(y|\boldsymbol{x}_i; \boldsymbol{\Theta})$ that can generate a label distribution similar to the ground truth output $\hat{\boldsymbol{d}}_i^{k-1}$, which is updated in the last iteration. One reasonable choice of $p(y|\boldsymbol{x}; \boldsymbol{\Theta})$ is the maximum entropy model [29], which has the exponential form $p(y|\boldsymbol{x}; \boldsymbol{\Theta}) = \frac{1}{Z} \exp\left(\boldsymbol{\theta}_y^T \boldsymbol{\phi}(\boldsymbol{x})\right)$, where $Z = \sum_y \exp\left(\boldsymbol{\theta}_y^T \boldsymbol{\phi}(\boldsymbol{x})\right)$ is the normalization factor, $\boldsymbol{\theta}_y$ is the y-th column in $\boldsymbol{\Theta}$, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_C]$ and $\boldsymbol{\phi}(\boldsymbol{x})$ is the feature of $\boldsymbol{x}$. Here, we note that $\boldsymbol{\theta}_y$ is the weight for predicting the y-th crowd number label. Observing that the crowd number variation of the video is a smoothing process, we consider that the weights for the adjacent crowd number should be locally correlated. Therefore, the mixed $\ell_{2,1}$ norm is adopted to guarantee the structured sparsity among the adjacent weight matrices [30]. If the Kullback-Leibler divergence is used to measure the similarity between distributions, then the best parameter vector for the k-th iteration is determined by

$$\begin{aligned}\boldsymbol{\Theta}^k &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{i,j} \left(\hat{d}_{y_j,\boldsymbol{x}_i}^{k-1} \ln \frac{\hat{d}_{y_j,\boldsymbol{x}_i}^{k-1}}{p(y_j|\boldsymbol{x}_i; \boldsymbol{\Theta})} \right) + \gamma \sum_{c=1}^{C'} \left\| \widehat{\boldsymbol{\Theta}}_c \right\|_{2,1} \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} - \sum_{i,j} \hat{d}_{y_j,\boldsymbol{x}_i}^{k-1} \ln p(y_j|\boldsymbol{x}_i; \boldsymbol{\Theta}) + \gamma \sum_{c=1}^{C'} \left\| \widehat{\boldsymbol{\Theta}}_c \right\|_{2,1}, \end{aligned} \tag{3}$$

where γ is the trade-off factor, $\widehat{\boldsymbol{\Theta}}_c = [\boldsymbol{\theta}_c, \boldsymbol{\theta}_{c+1}, \cdots, \boldsymbol{\theta}_{c+\epsilon-1}]$ (represents ϵ adjacent columns in $\boldsymbol{\Theta}$), ϵ is a factor to control

the width of temporal window of $\widehat{\boldsymbol{\Theta}}_c$ (set as 6), $C' = C - \epsilon + 1$. We would show the necessity of adding the mixed $\ell_{2,1}$ norm in our experiments.

Substituting the maximum entropy model into Eq. (3) yields the target function of $\boldsymbol{\Theta}$:

$$T(\boldsymbol{\Theta}) = \sum_i \ln \sum_j \exp\left(\boldsymbol{\theta}_{y_j}^T \boldsymbol{\phi}(\boldsymbol{x}_i)\right) - \sum_{i,j} \hat{d}_{y_j,\boldsymbol{x}_i}^{k-1} \cdot \boldsymbol{\theta}_{y_j}^T \boldsymbol{\phi}(\boldsymbol{x}_i) + \gamma \sum_{c=1}^{T'} \left\| \widehat{\boldsymbol{\Theta}}_c \right\|_{2,1}. \quad (4)$$

The optimization of Eq. (4) can be efficiently solved by the limited-memory quasi-Newton method (L-BFGS) [31].

D. Updating Label Distributions by Mixture of Gaussians

After the above LDL step, the conditional p.m.f. $p(y|\boldsymbol{x}; \boldsymbol{\Theta}^k)$ is obtained. Thus, the label distribution of each training frame $\boldsymbol{x}_i$ can be predicted as $p(y|\boldsymbol{x}_i; \boldsymbol{\Theta}^k)$. According to the predicted label distributions, the crowd number of $\boldsymbol{x}_i$ can be estimated as $\hat{\alpha}_i = \arg\max_y p(y|\boldsymbol{x}_i; \boldsymbol{\Theta}^k)$. The absolute error of each crowd number estimation is then calculated by $e_i = |\alpha_i - \hat{\alpha}_i|$, and those predicted label distributions with an absolute crowd number estimation error lower than the MAE (mean absolute error) of the whole training set, i.e., MAE $= \frac{1}{n}\sum_i e_i$, are selected as the training set for fitting the standard deviation σ to each crowd number. The selected training set are divided into c subsets by their crowd number, i.e., the frames in each subset have the same crowd number.

For each frame, the variance of its adjacent frames also affect the label ambiguity. Therefore, we propose a mixture of Gaussians model to generate the label distribution for each frame in the video. The nearest neighbor frames and their corresponding weights for generating the mixture of Gaussians model should be determined. For the i-th frame $\boldsymbol{x}_i$ with crowd number α_i, we consider the p nearest neighbor frames with different crowd number label from $\boldsymbol{x}_i$. The feature similarity and frame distance are included for choosing the nearest neighbor frames. Here, the distance metric between frame $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as: $D(i,j) = \left\|\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_j)\right\|_2^2 + P_{\alpha_i}^k \cdot |i-j|$, where $P_{\alpha_i}^k$ is the parameter for balancing the importance of the feature similarity and frame distance for crowd number α_i in the k-th iteration. We denote the p nearest neighbor frames of $\boldsymbol{x}_i$ as $N_{i,p}$. Thus, the weight of the j-th nearest neighbor frame is assigned as $W_{ij} = \frac{1}{Z'}\exp(-D(i,j))$, $j \in N_{i,p}$, where $Z' = 1 + \sum_{j \in N_{i,p}} \exp(-D(i,j))$ is the normalization factor and the weight of the original label distribution for the current frame is $W_{ii} = \frac{1}{Z'}\exp(-D(i,i)) = \frac{1}{Z'}$. Suppose the frame index set corresponding to the subset of the crowd number α_i is denoted by I_{α_i}, the label distribution for frame $m \in I_{\alpha_i}$ would be generated by the mixture of Gaussians model as

$$\hat{d}_{y_j,\boldsymbol{x}_m}^k = \sum_{s \in N_{m,p} \cup m} W_{ms}^k d_{y_j,\boldsymbol{x}_s}^{k-1}, \quad j = 1, 2, \cdots, C, \quad (5)$$

where $d_{y_j,\boldsymbol{x}_s}^{k-1} = \frac{1}{\sigma_{\alpha_s}^{k-1}\sqrt{2\pi}Z}\exp(-\frac{(y_j-\alpha_s)^2}{2(\sigma_{\alpha_s}^{k-1})^2})$, $W_{ms}^k = \frac{1}{Z'}\exp(-\|\boldsymbol{\phi}(\boldsymbol{x}_m) - \boldsymbol{\phi}(\boldsymbol{x}_s)\|_2^2 - P_{\alpha_i}^k \cdot |m-s|)$.

The generated mixture of Gaussians label distributions of the selected training set are assumed to be close to the label distributions generated by the maximum entropy model $p(y|\boldsymbol{x}; \boldsymbol{\Theta}^k)$. The balancing parameters P_α^k ($\forall \alpha \in \mathcal{Y}$) are initialized as all ones. The parameters σ_α^k and P_α^k are optimized by alternating the following two steps iteratively.

1) Given P_α^k, updating σ_α^k as

$$\sigma_\alpha^k = \arg\min_{\sigma_\alpha > 0} \sum_{m \in I_\alpha} \sum_j \hat{d}_{y_j,\boldsymbol{x}_m}^k \ln \frac{\hat{d}_{y_j,\boldsymbol{x}_m}^k}{p(y_j | \boldsymbol{x}_m; \boldsymbol{\Theta}^k)}, \forall \alpha \in \mathcal{Y}, \quad (6)$$

where $\hat{d}_{y_j,\boldsymbol{x}_m}^k$ is initialized as the mixture of Gaussians label distribution calculated by Eq. (5). Meanwhile, the single Gaussian label distribution $d_{y_j,\boldsymbol{x}_s}^k$ ($j = 1, 2, \cdots, C$) for each frame $\boldsymbol{x}_s$ is updated with σ_α^k.

2) Given σ_α^k, updating P_α^k as

$$P_\alpha^k = \arg\min_{P_\alpha > 0} \sum_{m \in I_\alpha} \sum_j \tilde{d}_{y_j,\boldsymbol{x}_m}^k \ln \frac{\tilde{d}_{y_j,\boldsymbol{x}_m}^k}{p(y_j | \boldsymbol{x}_m; \boldsymbol{\Theta}^k)}, \ \forall \alpha \in \mathcal{Y}, \quad (7)$$

where $\tilde{d}_{y_j,\boldsymbol{x}_m}^k = \sum_{s \in N_{m,p} \cup m} W_{ms}^k d_{y_j,\boldsymbol{x}_s}^k$. Meanwhile, W_{ms}^k is updated with P_α^k and $\hat{d}_{y_j,\boldsymbol{x}_m}^k = \tilde{d}_{y_j,\boldsymbol{x}_m}^k$.

Both Eq. (6) and (7) can be effectively solved by the log barrier interior-point method [32]. The stopping condition in the above optimization is the sum of differences of all the P_α^k and σ_α^k ($\forall \alpha \in \mathcal{Y}$) for two adjacent iterations are both under predefined thresholds, which are set as 20 and 0.1, respectively.

After σ_α^k is determined for every crowd number in $\mathcal{Y}$, the updated label distribution of each frame $\boldsymbol{x}_s$, i.e., $\hat{d}_{y_j,\boldsymbol{x}_s}^k$ ($j = 1, 2, \cdots, C$), is obtained and again sent into the LDL step in Eq. (3) to start the next iteration $k+1$. The whole process repeats until the MAE difference between two adjacent iterations is less than a predefined threshold, which is set as 0.01 in our experiments. Finally, the predicted crowd number for each test frame x' is determined by $y^* = \arg\max_y p(y|x'; \boldsymbol{\Theta}^*)$, where $\boldsymbol{\Theta}^*$ is the optimal weight parameter after convergence.

IV. Experiments

A. Experimental Setup

1) Datasets: We conduct experiments on seven indoor datasets: four videos in three different classrooms, two videos in different canteens and one video on a bus. Four classroom videos are collected by [6]. We annotate each frame of the four videos (24 fps) by ourselves. In the experiments, the regions of interest (ROI) are all set as the whole image except the red rectangle area where the digits of second changes as shown in Fig. 4(a). The crowd number varies a lot in the classroom videos. The four classroom scenes have much difference in illumination conditions and facilities in the background. Classroom 1/3_1/3_2 are all recorded in the daytime, while classroom2 are shot at night. For videos in the daytime, the curtain in the left top of image is always open in classroom1, half closed in classroom3_1 and first half closed and then completely closed in classroom3_2. In most time of

(a) (b) (c) (d)

Fig. 4. ROI settings in classroom2 (a), canteen1 (b), canteen2 (c) and bus (d).

TABLE II

VIDEO STATISTICS IN THIS PAPER. #FRAMES IS THE NUMBER OF TOTAL FRAMES; MIN, MAX, AVG AND TOTAL DENOTE THE MINIMUM, MAXIMUM, AVERAGE AND TOTAL CROWD NUMBER IN THE VIDEO

Video	Resolution	#Frames	Min	Max	Avg	Total
Classroom1	704*576	116602	1	53	37.2	4336952
Classroom2	704*576	43860	24	37	31.8	1395397
Classroom3_1	704*576	115572	0	55	36.1	4176666
Classroom3_2	704*576	119751	9	59	43.7	5236213
Canteen1	352*288	59990	0	16	5.2	314225
Canteen2	352*288	60000	0	19	11.1	668850
Bus	704*576	28253	19	40	30.1	851695

the classroom 1/3_1/3_2, a teacher is giving a class to the students and most of the students are staying almost static for a long time. In contrast, the students are self-studying for most of the time in classroom2. To evaluate the robustness of our algorithms to the change of scene and illumination condition, the cross-scene analysis is also carried out in the experiments. Some statistics of all the datasets are shown in Table II.

Two canteen videos (25 fps) are collected and annotated for each frame by ourselves in the canteens of a university campus. For canteen scenes, each frame is partitioned into two parts by the red line as shown in Fig. 4(b) and (c). The ROI are all set as the left part since we concentrate on the number of consumers in the canteens. In canteen scenes, people are standing in different lines and waiting for the meals. People standing in front of the lines usually leave as soon as they get their meals while other people move forward slowly in their lines. Moreover, some people would walk around to choose which lines they would join. Thus, the percentages of moving people are larger than the classroom scenes and far less than most outdoor scenes (such as in Mall [2] and UCSD [3]).

The bus video (25 fps) comes from real city public transport system. People are severely occluded with each other for most part of the video and the scene is affected by the outdoor lighting now and then. Each frame is annotated by ourselves. For the bus scene, each frame is partitioned into three parts by the red lines as shown in Fig. 4(d) and ROI is set as center part of the image for precluding the effects of outdoor vehicles and lights. In the video, passengers get on or off the bus and stand almost static on the bus when the bus is traveling.

The total video length of all the seven datasets is more than six hours. The datasets will be publicly available to the research community by request. Since people often stay almost static for a long time, we extract the frames with a gap of 10 frames for the video to construct each dataset. As the three-frame different method is adopted for foreground extraction, the first and last frames of each video are not used in the experiments. Except first and last frames of each video, all other frames are equally divided into two parts. Former half frames are adopted for training and latter half for testing.

2) Baselines: We compare our methods with GPR [19], CA-RR [2], AlexNet [26], CSRNet [24], LDL [8], LDL with mixed $\ell_{2,1}$ norm (LDLN) and head-shoulder detection method (HSD) [6]. In the foreground extraction step, spatial order R and temporal order T in Eq. (1) are set as 11 and 4, respectively. For AR coefficients, C_{000} is set as 0.5 and all other C_{ijk}s are chosen to be the same satisfying all coefficients sum up to one, which is the same as in [15]. The threshold for judging the foreground of the AR process is set as 0.05. AlexNet [26] replaces the original network by replacing the output of the last fully-connected layer with one crowd number output. The original input frames are all resized to $227 \times 227 \times 3$ without the foreground extraction. The batch size, momentum and learning rate are set to 10, 0.9 and 0.0005, respectively. The loss function of the CSRNet [24] is adapted to be the Euclidean distance between the ground truth and the estimated crowd number to fit our datasets. The batch size, momentum and learning rate are set to 10, 0.95 and 10^{-7}.

TABLE III

THE COMPARATION OF MAE IN LDL RELATED METHODS WITH THE WEIGHTED AVERAGE OR THE MAXIMUM OF THE LABEL DISTRIBUTION AS THE FINAL PREDICTION ON SEVEN DATASETS

Method	Class1	Class2	Class3_1	Class3_2	Canteen1	Canteen2	Bus	Avg
LDLW	3.76	**0.98**	1.71	1.02	1.18	**1.89**	**3.34**	1.98
LDL	**3.35**	1.07	**1.42**	**1.00**	**0.85**	2.17	3.48	1.91
LDLNW	1.97	1.07	3.00	**0.98**	**3.06**	1.73	**2.04**	1.98
LDLN	**1.50**	**0.94**	**1.25**	1.04	3.48	**0.95**	2.28	1.63
MoG-LDLW	1.02	0.66	0.44	**0.86**	3.76	1.30	**1.96**	1.43
MoG-LDL	**0.81**	**0.65**	**0.37**	1.10	**3.71**	**0.79**	2.11	1.36
MoG-LDLNW	0.89	0.72	0.38	0.87	1.22	2.00	**2.99**	1.30
MoG-LDLN	**0.79**	**0.67**	**0.37**	**0.83**	**0.78**	**1.78**	3.14	1.19

Both LDL [8] and LDLN adopt the single static Gaussian distribution as the label distribution for each frame. The HSD method [6] clusters all the head and head-shoulder candidates by the adapted K-means. Our methods, the mixture of Gaussians label distribution learning with/without the mixed $\ell_{2,1}$ norm, are denoted as MoG-LDLN/MoG-LDL in the experiments. In all LDL related methods, the maximum crowd number C is set as 71 (from 0 to 70) and the initial standard deviations for all crowd numbers are set as 2.

Moreover, the original LDL [8] uses the weighted average of the label distribution as the final prediction of crowd number. However, we find that crowd number with the maximum description degree achieves better performances (highlighted by boldface) in most cases in MAE of all LDL related methods by 19 to 9, as shown in Table III, where the suffix 'W' means the weighted average of the label distribution is used and no suffix 'W' means the maximum description crowd number is adopted. We notice that the maximum prediction only needs the maximum description crowd number to be right. In contrast, the weighted average method requires that all the outputs of the label distribution should be reasonable and any small fluctuation of the outputs would influence the final result. Therefore, the maximum prediction is probably more robust than the weighted average prediction and it is adopted as the final prediction for all LDL related methods in the remainder of the experiments.

3) Evaluation Metric: By following the convention of existing works [1], [22] for crowd counting, the mean absolute error (MAE) and root mean squared error (RMSE) are adopted as the metrics to evaluate the compared methods. Moreover,

TABLE IV
MAE, RMSE, FA OF THE COMPARED METHODS ON FOUR CLASSROOM VIDEOS

Metric	MAE					RMSE					FA				
Scene	Class1	Class2	Class3_1	Class3_2	Avg	Class1	Class2	Class3_1	Class3_2	Avg	Class1	Class2	Class3_1	Class3_2	Avg
GPR [19]	1.84 (5)	0.98 (5)	1.26 (4)	2.13 (5)	1.55 (4)	3.00 (6)	1.28 (5)	2.42 (4)	3.64 (6)	2.59 (5)	0.28 (4)	0.33 (5)	0.41 (5)	0.27 (5)	0.32 (5)
CA-RR [2]	3.10 (7)	1.30 (7)	1.76 (6)	3.35 (7)	2.38 (7)	3.59 (7)	1.67 (6)	2.38 (3)	4.31 (7)	2.99 (6)	0.07 (8)	0.25 (7)	0.20 (6)	0.10 (7)	0.16 (7)
Alexnet [26]	2.41 (6)	2.59 (8)	2.19 (8)	4.37 (8)	2.89 (8)	2.99 (5)	2.96 (8)	2.85 (6)	5.02 (8)	3.46 (7)	0.14 (6)	0.09 (8)	0.17 (7)	0.05 (8)	0.11 (8)
CSRNet [24]	1.69 (4)	0.88 (3)	2.03 (7)	2.64 (6)	1.81 (6)	1.92 (3)	1.09 (3)	2.51 (5)	3.26 (5)	2.20 (3)	0.11 (7)	0.35 (4)	0.12 (8)	0.11 (6)	0.17 (6)
LDL [8]	3.35 (9)	1.07 (6)	1.42 (5)	1.00 (2)	1.71 (5)	7.46 (9)	1.74 (7)	4.16 (9)	2.03 (4)	3.85 (8)	0.30 (3)	0.41 (3)	0.55 (3)	0.53 (2)	0.45 (3)
LDLN	1.50 (3)	0.94 (4)	1.25 (3)	1.04 (3)	1.18 (3)	2.76 (4)	1.22 (4)	2.99 (7)	1.82 (2)	2.20 (3)	0.28 (4)	0.30 (6)	0.49 (4)	0.45 (4)	0.38 (4)
HSD [6]	3.19 (8)	3.12 (9)	2.50 (9)	4.39 (9)	3.30 (9)	4.06 (8)	3.81 (9)	3.17 (8)	5.18 (9)	4.06 (9)	0.01 (9)	0.01 (9)	0.01 (9)	0.01 (9)	0.01 (9)
MoG-LDL	0.81 (2)	**0.65 (1)**	**0.37 (1)**	1.10 (4)	0.73 (2)	1.61 (2)	**0.95 (1)**	**0.97 (1)**	2.00 (3)	1.38 (2)	0.67 (2)	0.46 (2)	0.73 (2)	0.48 (3)	0.59 (2)
MoG-LDLN	**0.79 (1)**	0.67 (2)	**0.37 (1)**	**0.83 (1)**	**0.67 (1)**	**1.56 (1)**	0.99 (2)	1.03 (2)	**1.64 (1)**	**1.31 (1)**	**0.68 (1)**	**0.48 (1)**	**0.76 (1)**	**0.83 (1)**	**0.69 (1)**

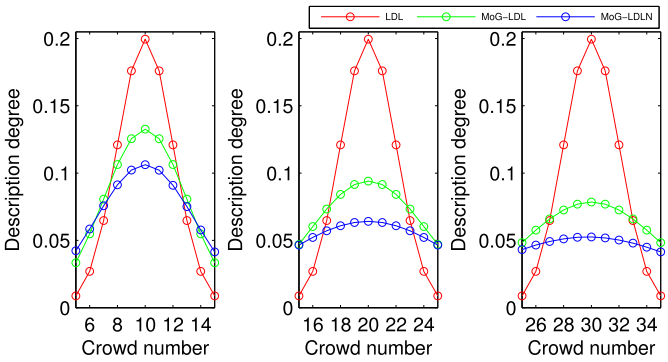


Fig. 5. The generated label distributions of the LDL, MoG-LDL and MoG-LDLN methods in classroom1 for frames with 10, 20 and 30 people.

we think frame accuracy (FA), which represents the rate of correctly predicted counting frame number over the total frame number, can also show effectiveness of algorithms. FA is defined as $\sum_i I(\alpha_i, \hat{\alpha}_i)/N$, where N is the number of test frames, α_i and $\hat{\alpha}_i$ are the ground truth and predicted crowd number of the i-th frame and I is the indicator function. Generally speaking, MAE and FA both indicate accuracy of the estimates and RMSE indicates robustness of the estimates.

B. Classroom Scenes

Instead of the single static Gaussian distribution adopted in LDL [8], our methods use the mixture of Gaussians model to generate the label distribution for each frame through the alternative optimization process in training. The comparisons of generated label distributions among LDL, MoG-LDL and MoG-LDLN in classroom1 are shown in Fig. 5. As can be seen, when the crowd number of the frame is low (10 people), the generated label distributions of MoG-LDL and MoG-LDLN is sharp and similar to that of LDL with static standard deviation (initialized as 2). As the crowd number of the frame increases, label ambiguity degree rises and the generated label distributions become flatter. For the frame with 30 people, the adjacent crowd number labels around 30 has quite close description degree since people are occluded with each other. We notice that MoG-LDLN is usually flatter than MoG-LDL. It means MoG-LDLN with mixed $\ell_{2,1}$ norm tends to generate label distributions with larger ambiguity degrees, which better fits the crowded situation in most indoor scenes.

Comparing with LDL, as shown in Table IV, our MoG-LDL method decreases MAE by 76%/39%/74%, RMSE by 78%/45%/77% and increases FA by 123%/12%/33% in classroom 1/2/3_1. In classroom3_2, MoG-LDL is a bit less accurate than LDL in MAE and FA. However, with the mixed $\ell_{2,1}$ norm, our MoG-LDLN decreases MAE by 76%/37%/74%/17%, RMSE by 79%/43%/75%/19% and increases FA by 127%/17%/38%/57% in classroom 1/2/3_1/3_2. The above results illustrate that through the training process, we have obtained various label distribution representations of the training frames in different crowdedness, which helps to learn a better mapping from features to the label distributions.

Comparing with MoG-LDL, our MoG-LDLN decreases MAE/RMSE by 25%/18% and increases FA by 73% in classroom3_2 and the average performance on four videos is decreased by 8%/5% in MAE/RMSE and increased by 17% in FA. It demonstrates that the mixed $\ell_{2,1}$ norm adopted in MoG-LDLN is much helpful in restricting the weights of predicting adjacent crowd numbers to be locally correlated. Moreover, comparing with LDL, LDLN decreases MAE by 55%/12%/12%/−4%, RMSE by 63%/30%/28%/10% and increases FA by 7%/27%/11%/15% in classroom 1/2/3_1/3_2. It confirms the effectiveness of mixed $\ell_{2,1}$ norm in LDLN.

As can be seen, our MoG-LDLN outperforms all the compared methods in MAE and RMSE on the four classroom datasets. The rankings of all the compared methods are shown in the brackets after the results. The best (rank *1st*) performance among the nine compared algorithms is highlighted by boldface. HSD performs worse than all four regression-based methods, GPR, CA-RR, AlexNet and CSRNet in all the metrics. We think it is mainly caused by the heavy occlusion of people in classroom scenes that not all heads or shoulders can be detected. AlexNet also performs poorly in the four classroom videos. We think the single crowd number output is less effective than the density map output in training the complicated network and the network structure is simple and not very deep. By using the VGG-16 network and the dilated convolution, CSRNet performs better/worse than AlexNet on 10/2 out of 12 measures and the average performance of CSRNet is better than AlexNet in MAE/RMSE/FA by 37%/36%/55%. However, it is still worse than the proposed MoG-LDL and MoG-LDLN in all metrics. We believe CSRNet's attempt not to loose resolution of the predicted density maps is not relevant when only a number is required. The complex backgrounds and varying lighting conditions might influence the feature extraction effect of the deep network. Moreover, both AlexNet and CSRNet do not consider the crowd number variation among the adjacent frames in videos and it might be the one of the reasons why they perform worse than the proposed methods.

The results of the former 2000 frames of MoG-LDLN and two best compared methods in 4 classrooms are shown

TABLE V
MAE, RMSE AND FA OF THE COMPARED METHODS ON TWO CANTEEN VIDEOS

Metric	MAE			RMSE			FA		
Scene	Canteen1	Canteen2	Avg	Canteen1	Canteen2	Avg	Canteen1	Canteen2	Avg
GPR [19]	0.91 (4)	2.14 (7)	1.53 (6)	**1.18 (1)**	2.63 (7)	1.91 (3)	0.35 (4)	0.13 (5)	0.24 (5)
CA-RR [2]	1.05 (6)	1.94 (3)	1.50 (4)	1.42 (6)	2.41 (3)	1.92 (4)	0.34 (5)	0.16 (3)	0.25 (3)
Alexnet [26]	2.30 (9)	1.94 (3)	2.12 (9)	2.79 (9)	2.45 (4)	2.62 (9)	0.14 (8)	0.16 (3)	0.15 (8)
CSRNet [24]	1.21 (7)	**1.67 (1)**	1.44 (2)	1.49 (7)	**2.03 (1)**	1.76 (2)	0.23 (7)	0.18 (2)	0.21 (6)
LDL [8]	0.85 (3)	2.17 (8)	1.51 (5)	1.30 (4)	2.69 (8)	2.00 (6)	0.38 (3)	0.12 (7)	0.25 (3)
LDLN	0.95 (5)	2.28 (9)	1.62 (7)	1.34 (5)	2.84 (9)	2.09 (8)	0.30 (6)	0.12 (7)	0.21 (6)
HSD [6]	1.30 (8)	1.94 (3)	1.62 (7)	1.54 (8)	2.46 (5)	2.00 (6)	0.02 (9)	0.02 (9)	0.02 (9)
MoG-LDL	0.79 (2)	2.11 (6)	1.45 (3)	1.22 (3)	2.62 (6)	1.92 (5)	**0.42 (1)**	0.13 (5)	0.28 (2)
MoG-LDLN	**0.78 (1)**	1.78 (2)	**1.28 (1)**	1.20 (2)	2.29 (2)	**1.75 (1)**	**0.42 (1)**	**0.19 (1)**	**0.31 (1)**

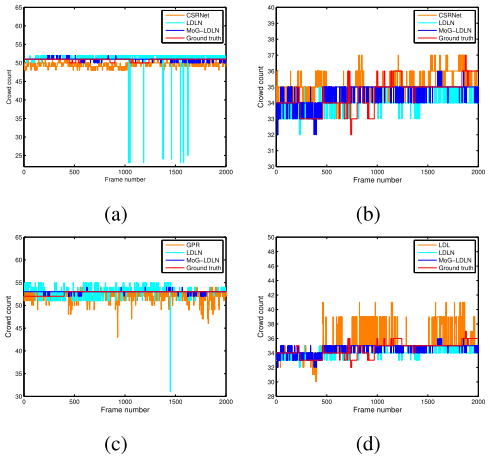


Fig. 6. The crowd counting results of the former 2000 frames of our MoG-LDLN method and the two best compared methods in classroom1 (a), classroom2 (b), classroom3_1 (c) and classroom3_2 (d), respectively.

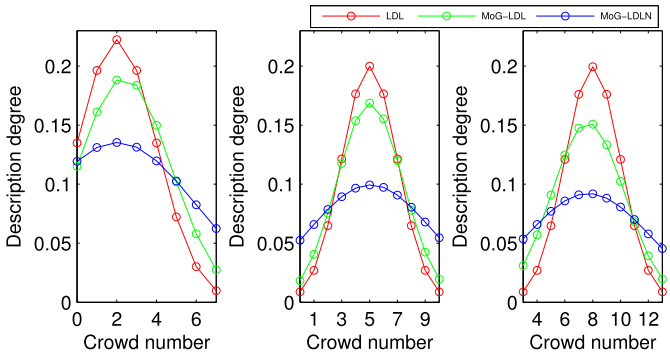


Fig. 7. The generated label distributions of the LDL, MoG-LDL and MoG-LDLN methods in canteen2 with 2, 5 and 8 people, respectively.

in Fig. 6. As can be seen, the results of MoG-LDLN are closer to ground truth annotations in most frames and more stable than the compared methods. We notice that LDLN would sometimes give out crowd number estimates far from the ground truths (big estimated errors) as shown in Fig. 6(a) and (c). Actually, LDLN with the mixed $\ell_{2,1}$ norm largely reduces the big estimated errors, comparing with LDL. Furthermore, with the mixture of Gaussians model, MoG-LDLN shows no big estimated errors on all classroom videos.

C. Canteen Scenes

The generated label distributions of LDL, MoG-LDL and MoG-LDLN for frames with different ambiguity degrees are shown in Fig. 7. As can be seen, the label distributions generated by our methods both become flatter as the crowd number increases. We notice that MoG-LDLN is flatter than MoG-LDL, which means it tends to generate label distributions with larger ambiguity degrees. As shown in Table V, both MoG-LDL and MoG-LDLN outperform LDL in all three metrics on two canteen scenes. The results indicate that label distributions generated by MoG-LDL and MoG-LDLN can better describe the ambiguity degrees of crowd number labels.

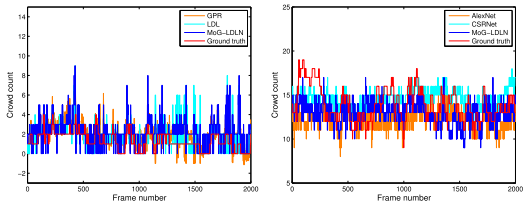
As can be seen, LDLN performs worse than LDL in all metrics. We think a single Gaussian is insufficient to produce a suitable label distribution for the frame sometimes. With the mixture of the Gaussians model, our MoG-LDLN decreases MAE by 8%/18%, RMSE by 8%/15% and increases FA by 11%/58% in canteen 1/2, comparing with LDL.

In canteen1, both our MoG-LDL and MoG-LDLN methods perform better than all the compared methods in all the metrics only except that our methods are a bit less stable than the GPR method in RMSE. In canteen2, the MoG-LDL performs worse than CA-RR, AlexNet, CSRNet and head-shoulder detection method in all three metrics only except it is better than the head-shoulder detection method in FA. With the mixed $\ell_{2,1}$ norm, the MoG-LDLN largely improves the performance. Our MoG-LDLN performs worse than the CSRNet in MAE and RMSE, but outperforms all the other compared methods in all the metrics. CSRNet deploys the well-known VGG-16 network [25] as the front-end and uses dilated convolution to enlarge receptive fields. CSRNet performs better than AlexNet on all 6 measures in canteen scenes and the average performance of CSRNet is improved by 32%/33%/40% in MAE/RMSE/FA. However, our MoG-LDLN can still achieve comparable results and outperforms CSRNet in all three metrics in average. They all confirm the mixed $\ell_{2,1}$ norm is helpful in preventing our MoG-LDLN from overfitting and guaranteeing the structured sparsity among the adjacent weight matrices. The HSD method achieves comparable results to the regression-based methods in both canteen scenes, mainly because people are fewer and less occlusion happens in the two videos.

The results of the former 2000 frames of MoG-LDLN and the two best compared methods on two canteen scenes are shown in Fig. 8. As can be seen, the results of GPR and MoG-LDLN are both close to the ground truth annotations in canteen1, while CSRNet and MoG-LDLN are both accurate and stable in canteen2.

D. Bus Scene

The generated label distributions of LDL, MoG-LDL and MoG-LDLN for the frames with different ambiguity levels are shown in Fig. 9. As can be seen, ambiguity degrees of



(a) (b)

Fig. 8. The crowd counting results of the former 2000 frames of our MoG-LDLN method and the two best compared methods in canteen1 (a) and canteen2 (b), respectively.

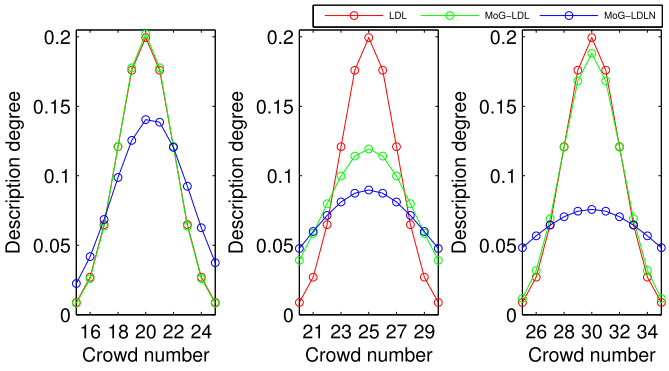


Fig. 9. The generated label distributions of the LDL, MoG-LDL and MoG-LDLN methods on the bus video with 20, 25 and 30 people, respectively.

TABLE VI

MAE, RMSE AND FA OF COMPARED METHODS ON BUS VIDEO

Metric	MAE	RMSE	FA
GPR [19]	3.98 (7)	5.07 (8)	0.10 (5)
CA-RR [2]	3.60 (4)	4.44 (2)	0.08 (6)
AlexNet [26]	4.25 (8)	4.98 (7)	0.03 (8)
CSRNet [24]	3.84 (6)	4.60 (5)	0.07 (7)
LDL [8]	3.48 (2)	4.53 (3)	0.18 (3)
LDLN	3.48 (2)	4.54 (4)	0.16 (4)
HSD [6]	5.96 (9)	7.13 (9)	0.01 (9)
MoG-LDL	3.71 (5)	4.85 (6)	0.20 (2)
MoG-LDLN	**3.14 (1)**	**4.30 (1)**	**0.22 (1)**

the generated label distributions of MoG-LDLN are more in accord with the scene's crowdedness than MoG-LDL on the bus scene. Comparing with LDL, as shown in Table VI, our MoG-LDLN decreases MAE/RMSE by 10%/5% and increases FA by 22%.

The bus video is quite challenging since the passengers are severely occluded with each other for a long period of time and only part of the heads and shoulders of passengers could be seen. Therefore, the head-shoulder detection method (HSD) [6] performs worst on this video. Four regression-based methods, GPR, CA-RR, AlexNet and CSRNet, all perform better than HSD. CSRNet performs better than AlexNet on all 3 measures on the bus scene and the performance of CSRNet is improved by 10%/8%/133% in MAE/RMSE/FA. Our MoG-LDL performs better than all compared methods in FA, but only better than GPR, AlexNet and HSD methods both in MAE and RMSE. With the mixed $\ell_{2,1}$ norm, our classification-based method MoG-LDLN outperforms all compared methods in all the metrics on the bus scene. It again confirms the mixed $\ell_{2,1}$ norm helps MoG-LDLN to be more accurate and robust.

The results of all test frames of MoG-LDLN and two best compared methods on the bus video are shown in Fig. 10. We notice that the curve of our MoG-LDLN is not very

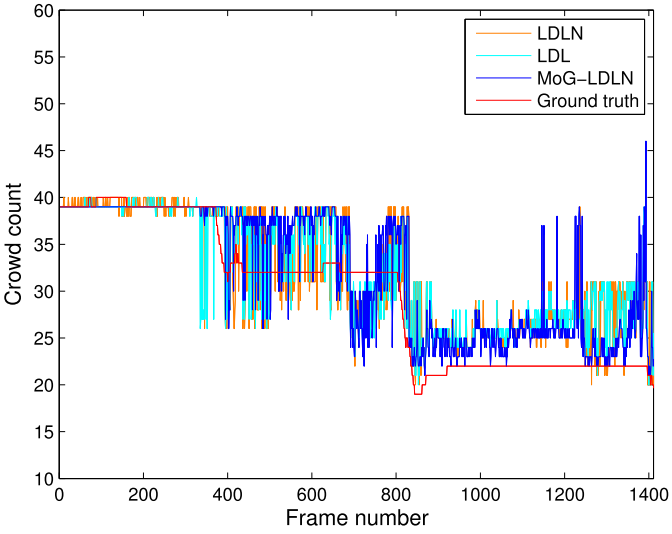
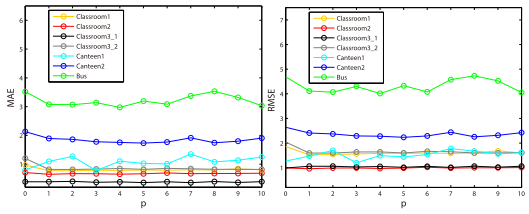


Fig. 10. The crowd counting results of all the test frames of our MoG-LDLN method and the two best compared methods on the bus video.



(a) (b)

Fig. 11. Horizontal axis corresponds to nearest neighbors p, and vertical axis corresponds to MAE (a) and RMSE (b) on all seven indoor videos.

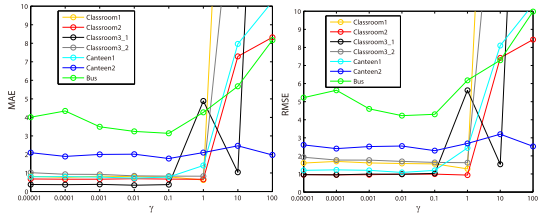
smooth. The mixed $\ell_{2,1}$ norm is adopted in MoG-LDLN to restrict the weights of predicting the adjacent crowd numbers to be locally correlated in training process. However, the changing features of frames, affected by the ambient light and the crowds' movement, could probably be the main reason why our crowd number curve is not very smooth. As can be seen, our MoG-LDLN is more accurate than the compared methods in most frames of the crowded bus scene.

E. Cross-Scene Analysis

To evaluate the robustness of our algorithms to the change of scene and illumination, we design the following cross-scene experiments. To cover all the videos of the classrooms and canteens and guarantee that the crowd numbers of the training video covers the crowd numbers of the test video, we set the three cross-scene experiments as follows: 1) train in classroom3_1 and test in classroom1; 2) train in classroom3_2 and test in classroom2; 3) train in canteen2 and test in canteen1. The illumination conditions and facilities in the background are both different in all three cross-scenes. In cross-scene 1, the curtain in the left top corner is half closed in classroom3_1, while it keeps open in classroom1. In cross-scene 2, classroom3_2 is recorded in the daytime, while classroom3_2 is shot at night. The camera view is different in 3 classroom scenes such that the layout of desks and chairs are distinct. In cross-scene 3, canteen2 is sparser and the illumination is brighter than canteen1.

TABLE VII
MAE, RMSE AND FA OF COMPARED METHODS ON CROSS-SCENES

Metric	MAE				RMSE				FA			
Scene	Class3_1→1	Class3_2→2	Canteen2→1	Avg	Class3_1→1	Class3_2→2	Canteen2→1	Avg	Class3_1→1	Class3_2→2	Canteen2→1	Avg
GPR [19]	3.15 (4)	17.82 (9)	2.87 (2)	7.95 (9)	4.81 (4)	18.68 (9)	4.08 (2)	9.19 (9)	**0.14 (1)**	0.00 (6)	0.21 (3)	0.12 (3)
CA-RR [2]	2.77 (3)	13.97 (8)	5.68 (8)	7.47 (8)	3.34 (3)	14.30 (8)	6.31 (8)	7.98 (6)	0.08 (3)	0.00 (6)	0.01 (8)	0.03 (6)
AlexNet [26]	6.12 (9)	3.29 (2)	3.99 (7)	4.47 (4)	7.47 (7)	**3.87 (1)**	4.40 (5)	5.25 (4)	0.02 (6)	0.06 (3)	0.03 (7)	0.04 (5)
CSRNet [24]	4.53 (5)	4.70 (3)	3.77 (6)	4.33 (3)	5.21 (5)	4.88 (3)	4.35 (4)	4.81 (2)	0.04 (5)	0.00 (6)	0.10 (4)	0.05 (4)
LDL [8]	5.14 (8)	10.06 (6)	3.47 (4)	6.22 (6)	8.43 (9)	11.75 (6)	4.50 (6)	8.23 (8)	0.01 (7)	0.01 (5)	0.06 (5)	0.03 (8)
LDLN	5.12 (7)	10.27 (7)	3.44 (3)	6.28 (7)	8.19 (8)	12.03 (7)	4.22 (3)	8.15 (7)	0.01 (7)	0.03 (4)	0.05 (6)	0.03 (6)
HSD [6]	4.83 (6)	6.04 (5)	6.57 (9)	5.81 (5)	6.69 (6)	6.78 (4)	7.28 (9)	6.92 (5)	0.01 (7)	0.00 (6)	0.00 (9)	0.00 (9)
MoG-LDL	2.54 (2)	5.60 (4)	3.73 (5)	3.96 (2)	**3.24 (1)**	6.83 (5)	5.36 (7)	5.14 (3)	0.08 (3)	0.10 (2)	0.22 (2)	0.13 (2)
MoG-LDLN	**2.52 (1)**	**3.27 (1)**	**2.63 (1)**	**2.81 (1)**	3.28 (2)	4.36 (2)	**3.77 (1)**	**3.80 (1)**	0.09 (2)	**0.16 (1)**	**0.24 (1)**	**0.16 (1)**



(a) (b)

Fig. 12. Horizontal axis corresponds to the trade-off factor γ, and vertical axis corresponds to MAE (a) and RMSE (b) on all seven indoor videos.

As shown in Table VII, with the mixture of Gaussians model, MoG-LDL decreases MAE by 51%/44%/−7% and RMSE by 62%/42%/−19% and increases FA by 700%/900%/267% than LDL in cross-scene 1/2/3. However, MoG-LDL performs worse than AlexNet and CSRNet in cross-scene 2 and 3 on MAE and RMSE. Furthermore, by using the mixed $\ell_{2,1}$ norm, our MoG-LDLN outperforms all the compared methods in MAE on all the three cross-scenes, and is only not as robust as AlexNet in cross-scene 2 and the frame accuracy is lower than GPR in cross-scene 3. CSRNet performs better than AlexNet on 6/3 out of 9 measures on the cross-scenes and the average performance of CSRNet is improved by 3%/8%/25% in MAE/RMSE/FA. Based on the experimental results, MoG-LDLN ranks *1st* in 66.7% cases and ranks *2nd* in the other cases across all the evaluation measures. Thus, MoG-LDLN achieves superior performance over the compared algorithms in cross-scene experiments. It confirms that MoG-LDLN is more robust to the change of scene and illumination than all compared methods.

F. Parameter Analysis

Fig. 11 shows the MAE and RMSE of MoG-LDLN on seven indoor videos with the variation of parameter p, which represents the number of most nearest neighbor frames. As can be seen, MoG-LDLN achieves best MAE and RMSE on 3 out of 7 videos when p is chosen between 3 and 5 and it is not sensitive to the chosen of parameter p. In average, our MoG-LDLN achieves the best MAE and RMSE when $p = 3$. The curve indicates that using the nearest adjacent frames to generate the mixture of Gaussians label distributions for crowd numbers is helpful in improving the prediction performance.

Moreover, γ is trade-off factor of Eq. (3). As shown in Fig. 12, MoG-LDLN achieves the best MAE and RMSE when γ is chosen among $\{1, 0.1, 0.01\}$. Variations of MAE and RMSE are both slow when γ is between 0.01 and 1, which shows MoG-LDLN is stable in this interval. Here, MoG-LDLN achieves best average MAE and RMSE when $\gamma = 0.1$.

V. CONCLUSION

In this paper, we propose a mixture of Gaussians label distribution learning method for indoor crowd counting. Since the label ambiguity is significantly affected by the crowd number, the standard deviation is adopted to generate the initial Gaussian label distribution for each frame. As crowd number variances of the adjacent frames are also considered, the mixture of Gaussians label distribution is used to represent each frame. An alternative optimization method is adopted to obtain the best standard deviations and the weights for each Gaussian label distribution. The mixed $\ell_{2,1}$ norm is adopted to guarantee the structured sparsity among the adjacent weight matrices. Three new indoor video datasets are collected with crowd number annotations. In the experiments, we achieve promising results on seven indoor datasets.

Although the proposed methods have achieved very promising results, they might be further improved in future work. For indoor scenes with hundreds of persons, such as concert and sport scenes, our indoor feature extraction method still works. After that, crowd numbers could be divided into a series of consecutive groups (no overlapping) with the same gap of crowd numbers and assume that the crowd numbers in the same group have the same ambiguity degree. Therefore, the variances of the Gaussian models for crowd numbers in the same group would be simultaneously optimized. However, these kinds of scenes with hundreds of people are much more complicated and the deep learned features might be adopted to achieve better performance in the future work, since the deep learned features have already shown its superiority in many other computer vision tasks, such as in [33]–[36].

REFERENCES

[1] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 833–841.
[2] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.
[3] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.
[4] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2325–2333.
[5] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 470–475.
[6] J. Luo, J. Wang, and H. Xu, "Real-time people counting for indoor scenes," *Signal Process.*, vol. 124, pp. 27–35, Jul. 2016.

[7] R. Yang, H. Xu, and J. Wang, "Robust crowd segmentation and counting in indoor scenes," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 505–514.
[8] Z. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing*, vol. 166, pp. 151–163, Oct. 2015.
[9] X. Geng and P. Hou, "Pre-release prediction of crowd opinion on movies by label distribution learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 3511–3517.
[10] X. Geng and M. Ling, "Soft video parsing by label distribution learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1331–1337.
[11] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1837–1842.
[12] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[13] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.
[14] Y. Ren and X. Geng, "Sense beauty by label distribution learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2648–2654.
[15] A. Khemlani, K. Duncan, and S. Sarkar, "People counter: Counting of mostly static people in indoor conditions," in *Video Analytics for Business Intelligence*. Berlin, Germany: Springer, 2012, pp. 133–159.
[16] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 545–551.
[17] N. Tang, Y.-Y. Lin, M.-F. Weng, and H.-Y. M. Liao, "Cross-camera knowledge transfer for multiview people counting," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 80–93, Jan. 2015.
[18] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.
[19] A. B. Chan and D. Dong, "Generalized Gaussian process models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2681–2688.
[20] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, vol. 1, no. 2, p. 3.
[21] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998, pp. 515–521.
[22] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.
[23] S. Huang *et al.*, "Body structure aware deep crowd counting," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1049–1059, Mar. 2018.
[24] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[26] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1299–1302.
[27] C. Shang, H. Ai, and B. Bai, "End-to-end crowd counting via joint learning local and global count," in *Proc. Int. Conf. Image Process.*, Sep. 2016, pp. 1215–1219.
[28] Y. Zhang, X. Wang, and B. Qu, "Three-frame difference algorithm research based on mathematical morphology," in *Proc. Int. Workshop Inf. Electron. Eng.*, 2012, pp. 2705–2709.
[29] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[30] Z. He *et al.*, "Data-dependent label distribution learning for age estimation," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3846–3858, Aug. 2017.
[31] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, 1989.
[32] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006.
[33] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.
[34] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[35] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.
[36] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

Miaogen Ling received the B.S. degree in mathematical science from Soochow University, China, in 2010, and the M.S. degree in computer science from Southeast University, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include machine learning and its application to computer vision and multimedia analysis.

Xin Geng received the B.Sc. and M.Sc. degrees in computer science from Nanjing University, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Deakin University, Australia, in 2008. He is currently a Professor and the Vice Dean of the School of Computer Science and Engineering, Southeast University, China. His research interests include machine learning, pattern recognition, and computer vision. He has published over 60 refereed papers in these areas. He has been an Associate Editor of the IEEE T-MM, FCS, and MFC, a Steering Committee Member of PRICAI, and the Program Committee Chair for conferences, such as VALSE'13 and PRICAI'18.