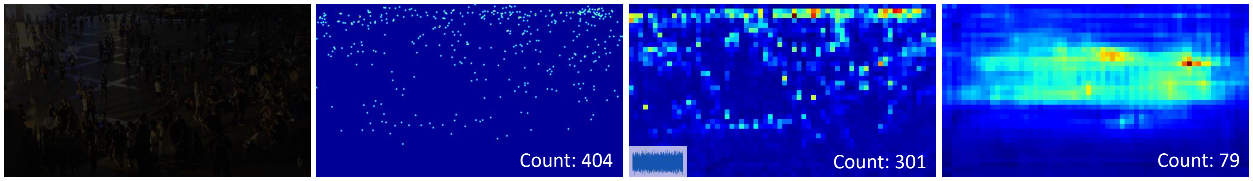
Does Ambient Sound Help? - Audiovisual Crowd Counting

Di Hu[1,*] Lichao Mou[2,*] Qingzhong Wang[3,*] Junyu Gao [4] Yuansheng Hua [2,5]
Dejing Dou [1] Xiao Xiang Zhu [2,5]
[1]Baidu Research [2]German Aerospace Center [3]City University of Hong Kong
[4]Northwestern Polytechnical University [5]Technical University of Munich
{hudi04,doudejing}@baidu.com qingzwang2-c@my.cityu.edu.hk
{lichao.mou,yuansheng.hua,xiaoxiang.zhu}@dlr.de gjy3035@gmail.com



(a) Input Image (b) Ground-truth (c) Audiovisual Prediction (d) Vision-only Prediction

Figure 1. Crowd counting on low-quality images. From left to right: input image with low illumination and strong noise, ground truth density map, predicted density map using both auditory and visual information and predicted density map only using visual information.

1. Introduction

Crowd counting has recently been a hot research topic [9, 13, 12], as it can benefit a wide range of applications, to name a few, safety monitoring, public space design, and disaster management. Consequently, crowd counting techniques, particularly computer vision-based approaches, have received increased interest. The success of current state-of-the-art visual crowd counting models can be attributed to the development of convolutional neural network (CNN) architectures that aim at learning better visual representations from images for this task [7, 8]. Albeit successful, vision-based crowd counting approaches could fail to capture informative features in extreme conditions[1] (c.f., Figure 1).

Investigations in the field of neurobiology show that human perception usually benefits from the integration of both visual and auditory information [15], e.g., lip reading, where correlations between lip movements and speech provide a strong cue for linguistic understanding [1]. This gives us an incentive that ambient sound could be an important cue for identifying the number of people in a scene. This hypothesis is in line with our daily experiences: *the louder we perceive the ambient sound to be, the more people there are.* However, incorporating the ambient sound into a visual crowd counting model and its contributions to this task still remain underexplored in the community. On the other hand, with the now widespread availability of smartphones, digital cameras, and video surveillance equipments, audiovisual data have been accessible at a reasonable cost. This enables us to explore the topic in this paper.

In this paper, we are interested in a novel task, audiovisual crowd counting. We pose and seek to answer the following questions:

- Is combining features coming from visual and auditory modalities better than only using visual features for crowd counting in extreme conditions?
- How do audiovisual crowd counting results vary under different illumination, noise, and occlusion conditions?
- How do we impose the audio information for effectively assisting the visual perception, i.e., how to fuse both modalities?

2. Dataset

To jointly utilize ambient sounds and visual contexts for crowd counting, an auDIoviSual CrOwd dataset ("DISCO" for short) is constructed.

Data Collection. To simultaneously capture the visual image sequences and record the audio signals, we use four video cameras, HDR-CX900E produced by *Sony Corporation*. In the collection process, we simulate the view of a surveillance camera and record crowds in some typical scenes at different time. As a result, we collect 248 video clips, around 20 hours and 385 GB data in total. Specifically, the resolution of each video is $1,920 \times 1,080$, and the frame rate is 25. For the audio information, the DV record 2-channel stereo with the sample rate of $48,000$.

From these raw data, 1,935 images and audios from various typical scenes are selected to construct our proposed

[*] The first three authors contribute equally to this work.

[1] In this paper, the extreme condition refers to a) low resolution, b) noise, c) occlusion, and d) low illumination.

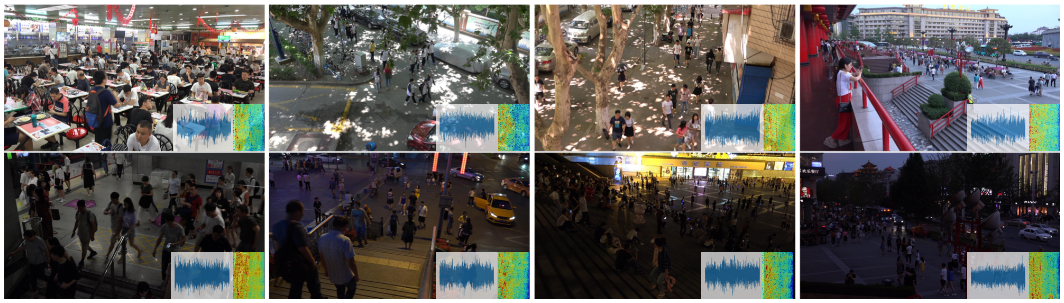


Figure 2. Examples of the DISCO dataset. We collect images and ambient sounds from a wide range of scenes, including indoor, outdoor, day- and night-time. For each image, an audio signal of one second is clipped from raw ambient sounds as auxiliary information. Raw ambient audio signals and their corresponding spectrograms are shown at the right bottom of each example.

dataset. For an image at t in a video, we extract its corresponding audio signals from $t - 0.5s$ to $t + 0.5s$. Some visual examples and their corresponding audio waveforms are shown in Figure 2.

Data Characteristics. DISCO consists of $1,935$ crowd images, a total of $170,270$ instances annotated with the head locations. The average, minimum and maximum number of people for each image are 87.99, 1 and 709, respectively. Compared with some traditional crowd counting datasets [17, 5], the proposed DISCO dataset are the first to record ambient sounds as auxiliary information of crowd scenes to reduce defects of single-vision sensors. In addition, we capture images at different times in one day to ensure their various illuminations (see Figure 2).

In a summary, DISCO dataset has three advantages comparing with others: 1) both audio and visual signals are provided; 2) cover different illuminations; and 3) a large variety of scenes are considered.

3. Our Approach

3.1. Overview

In order to benefit crowd counting with ambient sounds, a novel AudioVisual Counting (AVC) network is designed and consists of three modules (see Figure 3): (1) visual feature extraction, (2) audio feature extraction, (3) audiovisual feature fusion. Notably, comparing with traditional methods [2, 5, 17], where only visual information is employed, our network is characterized by the second and third module. To imitate such human capacity that we can perceive the scene by hearing, we introduce an audio module into the traditional counting framework, resulting in *AudioVisual Counting (AVC)* model.

3.2. Visual Feature Extraction

Similar to CSRNet [7], we employ the first ten layers of VGG16 [14] as the front-end CNN $\mathcal{V}_{CNN}$ to extract visual features. Given an RBG image I with a spatial size of $W \times H$, visual features v_{feat} can be extracted with the following equation:

$$v_{feat} = \mathcal{V}_{CNN}(I), \tag{1}$$

where $v_{feat} \in \mathbb{R}^{C \times \frac{W}{8} \times \frac{H}{8}}$, and C denotes the number of channels, i.e., 512.

3.3. Audio Feature Extraction

In this work, we use Log Mel-Spectrogram (LMS) for representing audio and CNN arch for modeling due to following considerations: 1) The audio feature of LMS has been widely used in CNN-like neural model for sound event detection and shown noticeable performance[3], and 2) Stoter et al.[16] demonstrate that using spectrogram-like feature can achieve comparable performance to the conventional MFCC in the counting task and much simpler. Even so, we still provide some discussions about different audio features and modeling settings in the experiments.

Given a raw audio signal $A_{raw} = \{a_1, a_2, \cdots, a_T\}$, we first sub-sample A_{raw} at 16kHz, and then employ short-time Fourier transform (STFT) using Hann window with the window size of 400 and a hop length of 160, to generate a 98×257 time-frequency map. Afterwards, Mel filter bank is applied, and a 96×64 representation A_{spec} can be then obtained for each raw audio signal. Finally, we utilize a VGG-like deep convolutional neural network [3] to extract audio features a_{feat} as follows:

$$a_{feat} = \mathcal{A}_{CNN}(A_{spec}), \tag{2}$$

where $a_{feat} \in \mathbb{R}^{C \times W_a \times H_a}$, and $C = 512$.

3.4. Feature-wise Audiovisual Fusion

To effectively fuse both audio and visual information in crowd counting, we introduce a feature-wise fusion module which aims at adaptively adjusting visual feature responses with transformed audio embeddings. Concretely, based on the extracted audio features, two feature-wise parameters γ and β are learned to model such cross-modal influence in terms of multiplicative and additive aspects, respectively. The formula is shown here:

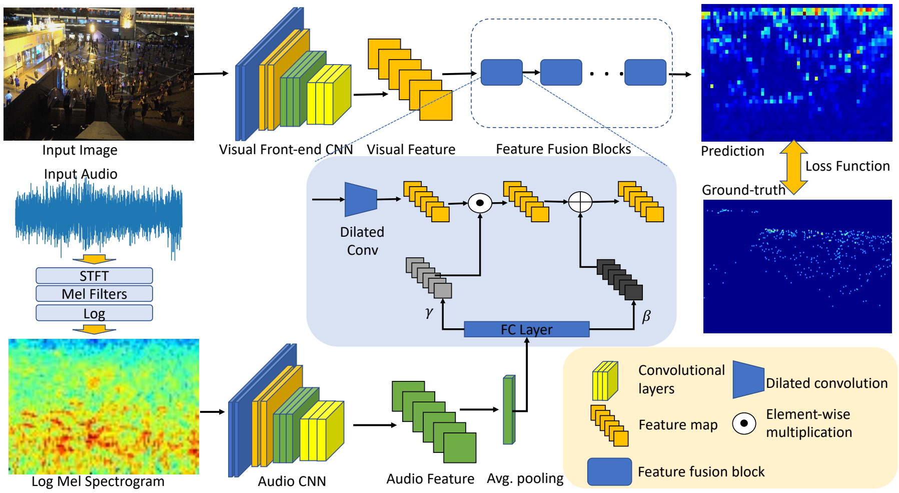


Figure 3. Overview of the proposed AudioVisual Counting model (AVC). The proposed AVC model is composed of three modules (1) visual feature extraction, (2) audio feature extraction, and (3) feature fusion. Note that our AudioVisual Counting framework can be applied to any vision-based counting model.

$$v_{l+1} = \mathcal{F}_l\left(\gamma_l \odot \mathcal{D}^l_{CNN}(v_l) + \beta_l\right), \tag{3}$$

where $v_l \in \mathbb{R}^{C_l \times W_l \times H_l}$ indicates outputs of the lth feature fusion block, $\mathcal{D}^l_{CNN}$ denotes the lth dilated convolution layer, $\mathcal{F}_l$ and $\odot$ suggest the activation function and element-wise multiplication, respectively. Notably, l ranges from 0 to 6, and $v_0 = v_{feat}$. Normally, γ and β can be learned via different affine transformations, such as single or multiple neural networks. In this work, we simply use fully-connected layers to learn γ_l and β_l with the following two equations:

$$\gamma_l = FC^{\gamma}_l(AvgP(a_{feat})), \tag{4}$$

$$\beta_l = FC^{\beta}_l(AvgP(a_{feat})). \tag{5}$$

In these two equations, $AvgP$ represents average pooling, and $\gamma, \beta \in \mathbb{R}^{C_{l+1}}$. To implement Eq.3, γ_l and β_l are tiled to match the size of visual features before fusion, see Figure 3. L_2 norm is selected as the loss function in our experiments.

4. Experiments

4.1. Experimental Settings

First, we split our DISCO dataset into three sets: 200 images for validation, 300 images for testing, and the remaining 1,435 images for training. To obtain the ground-truth density maps, we convolve each binary annotations (centers of human heads are one, and the others are zero) with a 15×15 Gaussian kernel $\mathcal{K} \sim \mathcal{N}(0, 4.0)$.

In the training phase, we select Adam as the optimizer and set its parameters as recommended. The learning rate is initialized as $1e-5$ and decays by 0.99 every epoch. To alleviate overfitting, weight decay is employed with a λ of 1e-4. It is noteworthy that except for those with a low resolution of 128×72, we resize images into 1024×576 to reduce computational resources and time. In our experiments, the batch size is set as 4, and the maximum training epoch is 500. To fairly compare all models, we report their performances on the test split.

Regarding the audio CNN, we use VGGish [3] pre-trained on audioSet [6] and discard its last three fully-connected layers, resulting in a 6-layer CNN. For the visual front-end CNN, we use the first ten layers of VGG16 [14] pre-trained on ImageNet. In our feature fusion blocks, we employ dilated convolutions with the kernel size of 3 and the dilation rate of 2 to enlarge the receptive field. Similar to CSRNet [7], we stack 6 fusion blocks and up-sample outputs with a factor of 8 to yield density maps of full resolution.

4.2. Baselines and Evaluation Metrics

To investigate the task of audiovisual crowd counting, we compare our audiovisual counting model with several vision-based models, such as MCNN [17], CSRNet [7], SANet [2], and CANNet [8]. Notably, we use one of the state-of-the-art models, CSRNet [7], as the backbone of our audiovisual counting model, leading to the proposed AudioCSRNet. The architecture of our proposed network is shown in Figure 3. To assess the performance of each model, we employ **Mean Absolute Error (MAE)** and **Mean Square Error (MSE)** scores.

4.3. Experimental results

To evaluate the performance of our AudioCSRNet in crowd counting under extreme conditions, we conduct experiments on two extreme scenarios: 1) the quality of images is very low, and 2) occlusion exists in images. Table 1

Table 1. Performance on low-quality images. For Gaussian noise, the standard deviation denoted by σ is a fixed value. While in low illumination&Gaussian noise, the illumination decay rate r and standard deviation σ of Gaussian noise are random values and R and B represent the hyper-parameters to compute r and σ, respectively. The bold numbers denote the best performance and the blue numbers represent the second best performance.

Model & Image Quality	Low resolution		Gaussian noise				Low illumination&Gaussian noise				**Avg. Score**	
	128×72		$\sigma = 25/255$		$\sigma = 50/255$		$R = 0.2, B = 25$		$R = 0.2, B = 50$			
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
MCNN [17]	60.17	89.35	53.47	84.04	53.92	84.04	70.72	96.11	70.58	96.11	61.77	89.93
CANNet [8]	22.16	39.60	13.31	**27.23**	14.20	**28.04**	26.03	**49.11**	33.14	58.27	21.77	40.45
CSRNet [7]	17.14	**30.64**	13.79	28.01	14.55	29.15	35.78	62.76	45.88	75.40	25.43	45.19
AudioCSRNet	**16.88**	31.46	**13.07**	27.45	**13.70**	28.67	**25.06**	51.58	**27.33**	**45.16**	**19.21**	**36.86**
PSNR [4] ↑	22.27		30.05		24.13		9.94		10.43		—	
BRISQUE [11] ↑	29.75		82.19		69.06		56.08		66.39		—	

report results on the first scenario, where models are compared on three low-quality conditions: low illumination, low resolution, and strong noise. Specifically, we mimic images taken in the dark environment with the method proposed by [10]. To quantitatively measure the quality of input images, here we calculate PSNR [4] and BRISQUE [11]. Notably, images with high PSNR and BRISQUE scores are regarded as high-quality ones.

Comparisons between AudioCSRNet and its counterpart, CSRNet, directly demonstrate that introducing audio information can benefit crowd counting, in particular on lower-quality images. Besides, on images with low resolution (PSNR is 22.27) and Gaussian noise (PSNR is 30.05 and 24.13), AudioCSRNet surpasses all competitors and achieves the lowest MSE in comparison with visual models as well. Another advantage of introducing audio into crowd counting is that audiovisual models show strong robustness on variant scenarios, e.g., AudioCSRNet obtains the lowest average MAE and MSE score (19.21 and 36.86). Also, we show the density map predicted by CSRNet (d) and AudioCSRNet (c) in Figure 1.

Another scenario that we study is occlusion, where an input image is randomly occluded by a black rectangle. Figure 4 shows results of CSRNet and AudioCSRNet, and we can see that performances of them on occluded images dramatically decrease. Moreover, trends of curves in Figure 4 demonstrate that AudioCSRNet can often achieve lower MAE and MSE scores.

5. Conclusion

In this paper, we investigated a novel audiovisual task, that imposes audio information for assisting visual crowd counting in extreme conditions. We developed an audiovisual crowd counting dataset to facilitate progress in this field, which covers different scenes in different illuminations. Meanwhile, a feature-wise fusion model was developed to achieve audiovisual perception for crowd counting. Extensive experiments were conducted to explore audio effects in different visual conditions. We found that introducing audio is able to benefit crowd counting, in particular in the extreme conditions, such as low illumination, strong noise, low resolution and occlusion.

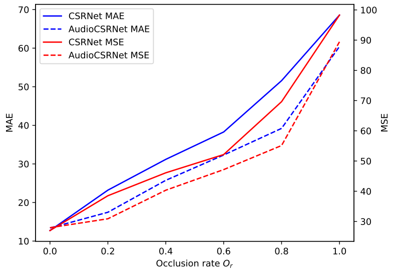


Figure 4. Performance of CSRNet and AudioCSRNet on occluded images. Occlusion rate $O_r = 0$ represent original images and $O_r = 1.0$ means there is no visual information.

References

[1] G. Calvert, E. Bullmore, M. Brammer, R. Campbell, S. Williams, P. McGuire, P. Woodruff, S. Iversen, and A. David. Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596, 1997.

[2] X. Cao, Z. Wang, Y. Zhao, and F. Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018.

[3] S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, R. Moore, M. Plakal, D. Platt, R. Saurous, et al. CNN architectures for large-scale audio classification. In *ICASSP*, 2017.

[4] A. Hore and D. Ziou. Image quality metrics: PSNR vs. SSIM. In *ICPR*, 2010.

[5] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. *arXiv:1808.01050*, 2018.

[6] Q. Kong, Y. Xu, W. Wang, and M. Plumbley. Audio set classification with attention model: a probabilistic perspective. In *ICASSP*, 2018.

[7] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, 2018.

[8] W. Liu, M. Salzmann, and P. Fua. Context-aware crowd counting. In *CVPR*, 2019.

[9] X. Liu, J. van de Weijer, and A. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. *arXiv:1803.03095*, 2018.

[10] K. Lore, A. Akintayo, and S. Sarkar. LLNet: A deep autoencoder approach to natural lowlight image enhancment. *Pattern Recognition*, 61:650–662, 2017.

[11] A. Mittal, A. Moorthy, and A. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[12] M. Shi, Z. Yang, C. Xu, and Q. Chen. Revisiting perspective information for efficient crowd counting. In *CVPR*, 2019.

[13] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M. Cheng, and G. Zheng. Crowd counting with deep negative correlation learning. In *CVPR*, 2018.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[15] B. Stein and M. Meredith. *The merging of the senses*. 1993.

[16] F. Stöter, S. Chakrabarty, B. Edler, and E. Habets. Classification vs. regression in supervised learning for single channel speaker count estimation. In *ICASSP*, 2018.

[17] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.