2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Image-based Air Pollution Estimation Using Hybrid Convolutional Neural Network

Jian Ma, Kun Li*, Yahong Han
Tianjin Key Laboratory of Cognitive Computing and Application,
School of Computer Science and Technology,
Tianjin University, Tianjin 300050, China
* Corresponding author: lik@tju.edu.cn

Jingyu Yang
School of Electrical and Information Engineering,
Tianjin University, Tianjin 300072, China

***Abstract*—Air pollution has a serious impact on our daily life, and how to quickly and easily measure the air pollution level without any expensive equipment is a quite challenging task. This paper proposes an air pollution estimation method using deep hybrid convolutional neural network from a single image, e.g., captured by a smartphone. The captured image is input to the main network, a very deep network, which solves the side effects of increased depth (degradation issues) by skip connection. This can improve network performance by simply increasing the depth of the network. Dark channel map is computed and fed into a secondary network to enrich the features with implicit representation. We have collected 1575 images of different scenes with different values of $PM_{2.5}$ to train the network in the end-to-end fusion mode. Experimental results on synthetic dataset and real captured dataset demonstrate that our method achieves excellent performance on classification of air pollution levels from a single captured image.**

I. Introduction

Nowadays, the air quality deteriorates day by day with the development of industry, especially in developing countries. The clouds, fogs, haze, dust from unburned fuels in the industry and in the transport industry are concrete examples of aerosol. The aerosol is dispersed by solid or liquid small particles with a size of 1 to 100 nm and is suspended in a colloidal dispersion formed in a gaseous medium. Among them, the haze, a lot of very fine dry dust particles floating in the air, will have serious damage to the human respiratory system. Therefore, estimating the air pollution level is a meaningful and demanding work.

At present, the classification of air pollution level is mainly based on the value of $PM_{2.5}$. Existing methods measure the $PM_{2.5}$ via special devices. Gravimetric method [1] is the main method that China monitors the $PM_{2.5}$. It has a high measurement accuracy, but the equipment is expensive, cumbersome, and complicated to operate. Moreover, gravimetric method detects the sum of the mass of the particles in the fixed area for a period of time (typically 1 hour), so it is difficult to achieve real-time monitoring or arbitrary regional scanning. Tao *et al.* [2] use the lateral scattering signal of the CCD laser radar to invert the aerosol function and the vertical backscattering coefficient. It is proved that the results of the lateral scattering laser radar and other professional instruments are highly consistent with those of other professional instruments, while the equipment of this method is cheap, simple operation and can achieve real-time measurement.

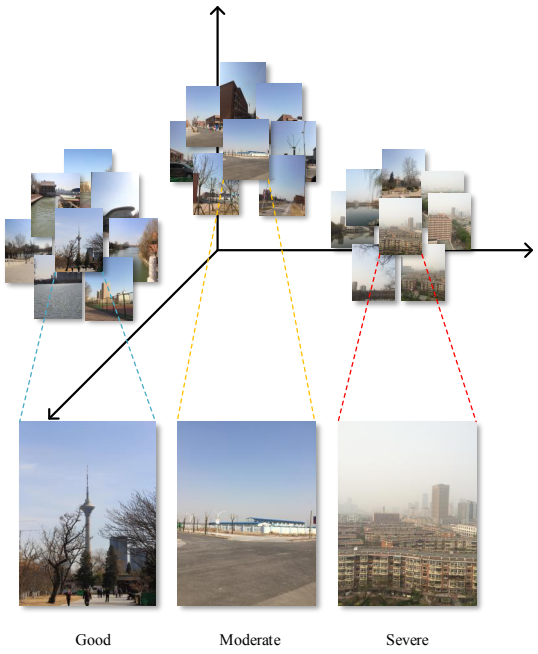


Fig. 1: The classification result of air pollution from a single image

Although the above methods are accurate, it is not convenient for people to carry a special device everywhere and the price of the device is a little high. It is highly demanding to measure air quality with daily-used devices such as smartphones. Zhang *et al.* [3] devise an effective convolutional neural network (CNN) to estimate the air quality based on photos, which achieves good classification results on his own dataset. Chakma *et al.* [4] create a dataset that contains total 591 images taken in Beijing with corresponding $PM_{2.5}$ concentrations, and use a deep convolutional neural network, VGGNet-19 [5], and random forest classifier to classify natural images into different categories based on their $PM_{2.5}$ concentrations. However, these methods have a little low accuracy.

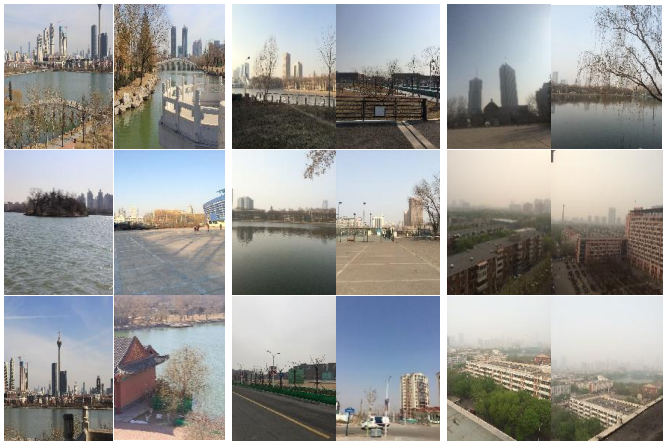
In this paper, we propose a hybrid deep network to rec-

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

ognize air pollution level from a single captured image by daily-used devices such as smartphones. The main network is a very deep network, which solves the side effects of increased depth (degradation issues) by skip connection. Dark channel map is computed and fed into a secondary network to enrich the features with implicit representation. We train the network in the end-to-end fusion mode with our collected 1575 images of different scenes with different $PM_{2.5}$ values. Experimental results on synthetic dataset and real captured images demonstrate that our method achieves more accurate classification of air pollution from a single captured image, compared with state-of-the-art methods. Our code and dataset will be publicly available on the project website.

II. Dataset

Due to the scarcity of outdoor images together with their air qualities, we capture 1575 images of different scenes using an Apple 5s mobile phone, and measure the corresponding $PM_{2.5}$ values with Hanvon M1 that uses laser scattering particle acceleration measurement technology with high precision. The distribution of $PM_{2.5}$ is between 0-300 $\mu g/m^3$, and it is categorized into three classes: Good ($PM_{2.5} < 75\mu g/m^3$), Moderate ($PM_{2.5} \in [75\mu g/m^3, 150\mu g/m^3)$), and Severe ($PM_{2.5} \geq 150\mu g/m^3$). The images are captured during the day (mainly after 10:00 AM) in order to better removing the interference of fog. Because the fog is formed by water vapor which naturally evaporates as the daytime temperature rises, while the haze is formed by solid particles which are not changed due to changes in environmental parameters. Fig. 2 shows some examples of our dataset.



(a) (b) (c)

Fig. 2: Some examples of our dataset: (a) Good ($PM_{2.5} < 75\mu g/m^3$), (b) Moderate ($PM_{2.5} \in [75\mu g/m^3, 150\mu g/m^3)$), (c) Severe ($PM_{2.5} \geq 150\mu g/m^3$).

III. The Proposed Method

Given a single RGB image I, our goal is to identify the level of air pollution $level$ from the image I, which can be formulated as:

$$level = f(I), \tag{1}$$

where I refers to the input RGB image, and f is a function which defines the relationship between the input image and air pollution level. This relationship is typically very complex, because different images with the same pollution level may have different degrees of blurring due to different weather parameters such as light and humidity.

Hence, we propose to learn this relationship. Inspired by the recent success of convolutional neural networks (CNN) in a variety of applications, especially classification, a straightforward way to do so is to directly model the function f with a CNN in which the input is the captured RGB image and the output is the air pollution level. However, the relationship between the RGB image and the pollution level cannot be well modelled by only using photometric features. Therefore, we compute a dark channel map from the RGB image, and feed it to a secondary network to learn an implicit representation. In this way, the system can better estimate the air pollution level with more accuracy and robustness, and the final problem is formulated as

$$level = f(I, J^{dark}), \tag{2}$$

where J^{dark} represents the dark channel map.

A. Dark Channel Map

In most non-sky local regions, some pixels always have at least one color channel with a very low value. In other words, the minimum value of light intensity of the region is a small value [6]:

$$J^{dark}(x) = \min_{c \in \{r,g,b\}} [\min_{y \in \Omega(x)} J^c(y)], \tag{3}$$

where J^c is a color channel of J which is the true radiance of natural scene, and $\Omega(x)$ is a window centered at pixel x. Accordingly, y is a pixel that belongs to $\Omega(x)$. A large number of experiments [6] show that the dark channel value of a hazy free image is almost zero (*i.e.*, $J^{dark} \to 0$).

The reasons for the low intensities in the dark channel map of the haze-free image are three factors [6]: 1) shadows of glass windows in cars, buildings, and cities, or projections of natural landscapes such as leaves, trees and rocks; 2) colorful objects or surfaces, such as green grass / trees / plants, red or yellow flowers / leaves, and blue water; 3) darker objects or surfaces such as dark colored tree trunks and stones. Under foul weather, the captured images will become brighter due to the light absorption by suspended particles, *i.e.*, dark channel maps of hazy images tend to be generally bright instead of dark according to the statistics. The denser the haze is, the higher the intensity in regions of the dark channel map is. Hence, the intensity of dark channel map gives a rough approximation for the thickness of the haze. We learn this implicit representation using a secondary deep network.

B. Our Model

The proposed network consists of convolutional layers, pooling layers, batch normalization, and a fully connection layer. Fig. 3 shows the architecture of our network. The RGB

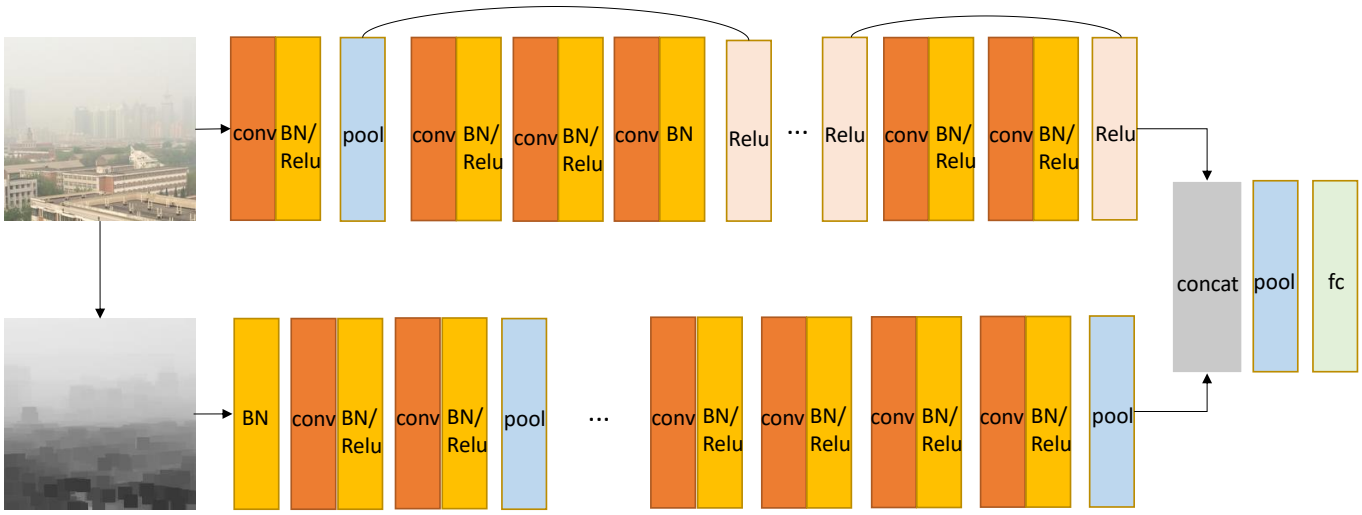


Fig. 3: The architecture of our network.

image is input to a very deep network, the residual network [7], to extract the photometric features, while the dark channel map is fed to the VGG network [5] to extract the implicit features. These two kinds of features are merged together at the end of the network.

The convolutional layer is represented as:

$$\mathcal{F} = \sigma(I \circ W + b), \tag{4}$$

where σ is the ReLU activation function, I represents the captured image, $\circ$ represents the convolution operator, W is a convolution kernel, and b is the offset.

The residual block is represented as:

$$\mathcal{H}(x) = \mathcal{F}(x, \{W_i, b_i\}) + \alpha x, \tag{5}$$

where the number of layers i used to learn the residuals should be more than 1, otherwise the decay is linear. α is a coefficient of the linear mapping to satisfy the dimension matching. This residual learning structure can be implemented by a forward neural network and shortcut connection as shown in the block diagram. The shortcut connection equivalent to a simple implementation of the same mapping, will not generate additional parameters, and hence will not increase the computational complexity. Moreover, the entire network can still be trained using end-to-end backpropagation.

For feature maps, unlike the residual network uses shortcut connections for the corresponding element addition, we transfer the feature maps from two networks into a fully connected layer:

$$F = F_{res} \oplus F_{vgg}, \tag{6}$$

where F_{res} and F_{vgg} represent the features extracted from the residual network and the VGG network, respectively. F refers to the final feature map, and $\oplus$ is the operation of dimension splicing.

C. Training

In order to achieve high-precision classification that can be used in daily life, we train the network by minimizing the loss function between the predicted result and the ground truth:

$$\delta(y, z) = -\log(\frac{e^{z_y}}{\sum_j e^{z_j}}), \tag{7}$$

where $\delta(y, z)$ is *Softmax* loss function, y is the air pollution level of image and z_j is the feature of the j^{th} image which is extracted by our network. At each iteration of the training, we use these gradients to update the weights of both networks in the opposite direction of the gradients. Experimental results show that the training loss tends to stabilize after 40000 iterations.

Since we have only 1575 images in our dataset, data augmentation is necessary for deep learning. Although there are some popular methods such as adding gaussian noise and changing the image brightness / saturation / contrast, they are unsuitable for our problem. Therefore, rotation (90, 180, 270 degrees) and random crop are adopted to extend the dataset. The final training set includes 18574 images transformed from 1375 images in our dataset and the remaining 200 images in our dataset are as the test set. Their labels are levels of air pollution (good, moderate, severe), which are set according to the measured $PM_{2.5}$ values.

Because training on the original images is slow, we resize the images to the ones with size of 224×224 and use mini-batches with size of 8 to have the best trade-off between speed and convergence. When the batch size is set to be 8, it need up to 11G memory while Titan X's memory is only 12G. Therefore, we use pre-trained models [7], [5] as initialized weights which have been trained in ImageNet classification task and train our system using the ADAM solver [8], with a

learning rate of 0.0001 and "step" as learning rate of decline strategy. The momentum is 0.9.

IV. Experiment

In this section, we evaluate the proposed deep hybrid convolutional neural network on synthetic dataset (Sec. IV-A) and our real captured dataset (Sec. IV-B). we compare the results of our method with four state-of-the-art methods which have achieved huge success in image classification tasks. We also compare three variants with different inputs in Sec. IV-B. The proposed method is implemented by Caffe [9] which provides multimedia scientists and practitioners with a clean and modifiable framework for state-of-the-art deep learning algorithms and a collection of reference models. All the test images are resized to 224×224. In test period, the batch size is set to be 16 and the iteration is 50. The experiments are run on a desktop with an Intel Core i7-4800K CPU, a NVIDIA GeForce GTX TITAN X GPU, and a 32GB RAM.

A. Results on Synthetic Dadaset

We first evaluate the methods on a synthetic dataset, which is generated by adding artificial haze to the haze-free color images in the 3D video dataset [10] that includes RGB images and depth maps. By setting different coefficients β, the hazy images with different pollution levels can be obtained by

$$\begin{cases} I(x) = J(x)t(x) + A[1 - t(x)], \\ t(x) = \mathrm{e}^{-\beta d(x)}, \end{cases} \tag{8}$$

where I is the synthesized hazy image, J is the original haze-free image, A is the global atmospheric light composition, and t is the medium transmission which depends on the distance. $\beta(\beta > 0)$ is the atmospheric scattering coefficient and $d(x)$ is the scene depth at pixel x. We keep the global atmospheric light composition A constant which decides the color of the haze we added. From previous work [11], the atmospheric scattering coefficient β has a linear relationship with the values of $\mathrm{PM}_{2.5}$. Therefore, we manually set different $\mathrm{PM}_{2.5}$ values and calculate the β values based on the linear relationship. Then, we can obtain the hazy image according to Eq. (8). In the experiments, we set $A = [0.755, 0.77, 0.77]$ and the values of $\mathrm{PM}_{2.5}$ are set from 130 $\mu g/m^3$ to 180 $\mu g/m^3$. Finally, we get 17442 hazy images which are distributed to different levels of air pollution.

Table. I shows the quantitative evaluation results compared with four famous models which are used for image classification tasks: AlexNet [12], VGGNet-19 [5], ResNet-152 [7], DenseNet [13]. We use the pre-trained models which were trained in the ImageNet dataset to fine-tune the networks. As shown in the table, ResNet-152 and DenseNet have obviously better performance than AlexNet and VGGNet-19 due to the use of batch normalization. In all the methods, our method achieves the most accurate estimation for air pollution level from a single image. Fig. 4 shows some example results of our deep hybrid convolutional neural network. Our method achieves accurate classification with over 75% probability of classification (*Softmax* output) for *Lawn* dataset [10] and over 98% probability of classification (*Softmax* output) for *Road* dataset [10]. The visual results also demonstrate the effectiveness of the proposed method.

(a) Good (b) Moderate (c) Severe

Fig. 4: Example results on synthetic *Lawn* dataset (Top) and *Road* dataset (Bottom).

Methods	Classification Accuracy
AlexNet [12]	51.13%
VGGNet-19 [5]	52.19%
ResNet-152 [7]	60.00%
DenseNet [13]	59.17%
Our Model	**62.38**%

TABLE I: Quantitative evaluation on synthetic dataset.

We also test our method on two virtual-world images which come from a famous game named as PLAYERUNKNOWN'S BATTLEGROUNDS[1] in Fig. 5. It can be seen that our model also has good performance on virtual-world images.

(a) Moderate (b) Severe

Fig. 5: Results on synthetic images from a famous game.

B. Results on Real Captured Dataset

We evaluate our method on real captured images compared with four popular methods: AlexNet [12], VGGNet-19 [5],

[1]http://pubg.com/

(a) Good (b) Moderate (c) Severe

Fig. 6: Example results on real captured dataset.

ResNet-152 [7], and DenseNet [13]. We use the pre-trained models which were trained in the ImageNet dataset to fine-tune the networks. Table. II gives quantitative evaluation for these methods. It can be observed that ResNet-152 and DenseNet also have better performance than AlexNet and VGGNet-19 due to the use of batch normalization. Our method has the most accuracy in all the methods due to the elegant design of the hybrid deep convolutional neural network. Fig. 6 shows our results on 6 randomly selected test images, which also demonstrate that our method achieves reasonable classification results.

Methods	Classification Accuracy
AlexNet [12]	86.75%
VGGNet-19 [5]	85.00%
ResNet-152 [7]	91.13%
DenseNet [13]	90.11%
Our Model	**92.19**%

TABLE II: Quantitative evaluation on real captured dataset.

In order to evaluate the influence of the input, especially the role of the dark channel map, we also compare three variants with different inputs in Table. III. DC represents dark channel map [6] and RGB refers to the original captured image. It can be seen that the dark channel map cannot be used as input independently or as a dominant input because the RGB image contains richer information. However, the dark channel map can be used as a secondary input to help extract implicit features, which improves the classification accuracy of air pollution levels. Therefore, our method has the highest accuracy when RGB + DC are as inputs.

Inputs	Classification Accuracy
DC + DC	35.86%
DC + RGB	36.50%
RGB + RGB	91.75%
RGB + DC (our method)	**92.19**%

TABLE III: Quantitative comparison with different variants.

V. Conclusion

In this paper, we have presented a novel deep learning approach for classification of pollution levels based on images. We train the proposed hybrid convolutional neural network simultaneously using RGB images and dark channel maps, and combine the feature maps in the end. Our method is validated with qualitative and quantitative evaluations against serval state-of-the-art methods. Experimental results on both synthetic and real captured datasets demonstrate that our method has achieved the best classification results. Besides, we provide a public dataset including outdoor images and their corresponding $PM_{2.5}$ values. In the future, we will collect more hazy images with larger range of $PM_{2.5}$ values from different places to improve the classification accuracy, consider more air quality factors, *e.g.*, PM_{10}, and study the relationship between the hazy images and the air quality factors.

VI. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61571322 and Grant 61771339.

References

[1] W. Zhang, S. Gao, X. Song, J. Liu, W. Liu, and Z. Chen, "Concentration measurement and metrical technology of fine particulate matter pm_(2.5)," *China Powder Science & Technology*, 2013.

[2] Z. Tao, D. Liu, Z. Wang, X. Ma, Q. Zhang, C. Xie, G. Bo, S. Hu, and Y. Wang, "Measurements of aerosol phase function and vertical backscattering coefficient using a charge-coupled device side-scatter lidar." *Optics Express*, vol. 22, no. 1, pp. 1127–34, 2014.

[3] C. Zhang, J. Yan, C. Li, X. Rui, L. Liu, and R. Bie, "On estimating air pollution from photos using convolutional neural network," in *ACM on Multimedia Conference*, 2016, pp. 297–301.

[4] A. Chakma, B. Vizena, T. Cao, J. Lin, and J. Zhang, "Image-based air quality analysis deep convolutional neural network," in *IEEE International Conference on Image Processing*, 2017, pp. 3949–3952.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1956–1963.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[10] G. Zhang, J. Jia, T. T. Wong, and H. Bao, "Consistent depth maps recovery from a video sequence," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 31, no. 6, p. 974, 2009.
[11] M. Jia, K. Li, Y. Han, J. Yang, and P. Du, "Image-based pm2.5 estimation and its application on depth estimation," in *submitted to IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
[13] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.