Air Quality Index Estimation Based on Image Analysis

Mohd Ashraf[1] | Niket[1] | Devender[1] | Dr. Vinod Kumar[2]

[1]UG scholar, Department of Computer Engineering, Delhi Technological University, Delhi, India.
[2]Associate Professor, Department of Computer Engineering, Delhi Technological University, Delhi, India

To Cite this Article
Mohd Ashraf., Niket., Devender & Dr. Vinod Kumar. Air Quality Index EstimationBased on Image Analysis. *International Journal for Modern Trends in Science and Technology* 7, 172-177 (2021).

ABSTRACT

Air pollution is an issue that is out of the control of an average citizen. Controlling air pollution requires preventive and control measures on a large scale implemented by the government. However, what an individual can do is protect him/her from the harmful effects of pollution by taking precautions such as not going out in times of severe pollution or wearing an air mask when travelling out. It will be very helpful if a person is able to find out the pollution level around him.

Government provides measures of pollution in terms of AIR QUALITY INDEX (AQI). However this is provided only at certain centre places. AQI may change drastically between these centres.

In this report, an effort was made to solve this problem by enabling an individual to find an estimate of the Air Quality Index near them with their smartphone, even without an Internet connection, by simply clicking an image of their surroundings. Using this information a person can take preventive measures to take care of his health. This will not only spread awareness about air pollution but also protect people from the harmful effects of air pollution.

We have used Machine Learning to achieve this goal. We prepared a dataset of images of sky and trained a model using several algorithms and compared them. We then used this model to recognise almost accurate AQI of the surroundings.

KEYWORDS: *AQI, Machine Learning , Convolution Neural Network, classification Algorithm, Image Recognition.*

I. INTRODUCTION

Air contamination is a blend of particles and gases that can achieve unsafe fixations both outside and inside. Its belongings can go from higher malady dangers to rising temperatures. Sediment, smoke, shape, dust, methane, and carbon dioxide are only a couple of instances of basic poisons.

The AQI estimates five criteria air contaminants (particulate issue, sulfur dioxide, carbon monoxide, nitrogen dioxide, and ozone), and changes over the deliberate poison fixations in a network's air to a number on a size of 0 to 500.

The most significant number on this scale is 100, since this number compares to the

National Ambient Air Quality Standard built up under the Clean Air Act.

An AQI level more than 100 implies that a toxin is in the unhealthful range on a given day; an AQI level at or underneath 100 implies that a poison perusing is in the acceptable range.

As of now aqi in India is discovered precisely by the Ministry of Environment, Forest and Climate Change and a couple of other private associations. There are a few focal points of having the option to identify aqi of the region by utilizing a cell phone.

Knowing the air nature of spot an individual can without much of a stretch choose whether he should design an excursion. Or then again whether he ought to have any air veil with him.

Having an application that can discover the air quality list of an area precisely can be of extraordinary significance since government and different offices measures AQI just at spots at a critical separation separated and since air quality record may change steeply it is exceptionally useful to have an application that can quantify the air quality list at a specific spot.

II. RELATED WORKS

There are so many research already done in the same field already which are as follows

1. There was a research paper named *image based pm 2.5 estimation and its applications on depth estimation* has been published by *Jian Ma, Kun Li, Yahong Han, Pufeng Du, Jingyu Yang* in which they proposed an image-based method for PM2.5 estimation by capturing a single image. They extracted high-level features based on convolutional neural networks (CNN) and learned the mapping between the features and PM2.5 by support vector regression (SVR). Given a captured image, they can estimate the PM2.5 value in real time.

2. There was another research paper named *utilizing user generated online photos to estimate and screen air pollution in major cities* published by *Yuncheng Li, Jifei Huang, Jiebo Luo* in which they proposed to employ the state of the art in computer vision techniques to analyze photos that can be easily acquired from online social media. Next, they established the correlation between the haze level computed directly from photos with the official PM 2.5 record of the taken city at the taken time. Their experiments based on both synthetic and real photos had shown the promise of their image-based approach to estimating and monitoring air pollution.

3. There was research paper published named *particle pollution estimation based on image analysis* proposed by *Chenbin Liu,Francis Tsow,Yi Zou,Nongjian Tao* to estimate PM air pollution based on analysis of a large number of outdoor images available for Beijing, Shanghai (China) and Phoenix (US).

III. LIMITATIONS OF EXISTING WORK

1. The resultant precision can be improved in the known execution techniques. More improvement at all means is required.

2. Datasets have typically been explicit to either a spot or specific contamination level. It's augmentation is fundamental for both preparing and examining more places.

3. There are various purposes behind wrong acknowledgment of contamination levels, for example, each spot has various poisons and have diverse contamination levels, a need to enhance include extraction procedures, preparing dataset ought to have tests out of every other place on earth and stretched out to recognize contamination levels all the more precisely.

IV. METHODOLOGY

1. DATA COLLECTION

Dataset consists of images of the sky taken at different angles with respect to the sun and at various times of the day to get a more general application.

Dataset is prepared by clicking images everyday of the sky at our nearest air quality measuring location as given on the site https://aqicn.org/ city/delhi/ which has data from various government organizations.

Clicking images near the air quality measuring locations gives us more accuracy in the dataset which in turn would give more accuracy in the final recognition of pollution levels of the air surrounding us. All the

images used by us in the dataset have been clicked using a HD camera of a regular smartphone and all the air quality index measures taken correspondingly with the images have been taken from authentic government sources.

2. FEATURES EXTRACTED FROM DATA

We extracted a total of 4 features which we considered important in understanding the relationship between the images and air quality index.

1) Average of Red, Green, Blue colour

As our image is made up of pixels and to represent the pixels values RGB colour code is considered better and the air quality index also changes the colour of the atmosphere as it increases. As our atmosphere is getting polluted every day and so it is changing the colour of the sky which clearly shows the relation between air quality and colour of an image.

2) Image Contrast

Contrast is that the distinction in physical property or color that creates an object (or its illustration in a picture or display) is distinguishable. The human sensory system is a lot more sensitive to distinction than absolute luminance. The image contrast of the images is calculated using Root mean square contrast.

$$RMS = \sqrt{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(I_{ij} - avg(I))^2}$$

where I_{ij} is intensity at (i,j) pixel of the image with size M by N, and $avg(I)$ is the average intensity of all pixels in the image.

Root mean sq. (RMS) distinction doesn't rely upon the angular frequency content or the spatial distribution of distinction within the image.

3) Entropy of the image

Image entropy quantifies the information contained in an image and is related to image texture. Image entropy used in here is Shannon entropy which visits every pixel and then compares the every next surrounding pixels with the checking pixel, when entropy is calculated it is stored according to the respective images.

$$entropy = -\sum_{i=1}^{M} p_i \log_2 p_i$$

where pi is the probability that the pixel intensity is equal to i, and M is the maximum intensity of the image

4) Average Brightness

Brightness may be a relative term. It depends on your seeing. Since brightness may be a relative term, thus brightness is outlined because the quantity of energy output by a supply of sunshine relative to the supply we tend to area unit examination it to. In some cases we are able to simply say that the image is bright, and in some cases, it's tough to understand.

V. MODELS USED

Figure 1: Basic Steps involved in Classification Models

IV. LOGISTIC REGRESSION

Logistic Regression is employed once the dependent variable (target) is categorical. In simple logistic regression output is 0 or 1.Its hypothesis is z=wx+b.

hΘ(x) = sigmoid (Z)

If 'z' goes to infinity, y which is to be predicted will become 1 and if 'z' goes to negative infinity y predicted will become 0.

error function used to minimize the error while performing classification

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Examples: To predict whether or not associate email is spam (1) or (0).

V. DECISION TREE

It is the most powerful and most used model for the classification and prediction and in this project it turned out to be the most accurate model with much greater accuracy than all the models. It is similar or just the same as a flow chart in which each node is an event which occurs on the feature of the data extracted or the parameter on which the prediction is to be made. It follows or applies the event of the following features to make a result or reach a conclusion. By applying the events on features basically it selects or splits the tree into two half it branches and then goes on to the next nodes and finally when leaves arrives then it gives the answer or one can say that the leaves contain the Answers.
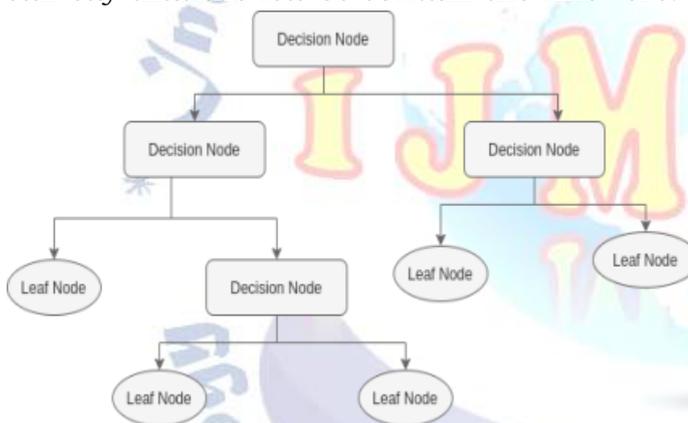
Decision tree uses information gain. Information gains calculate difference between entropy before split and entropy after split.

$$Info(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

Info(D) = Average amount of information needed to identify the class label of a tuple in D and here pi is the probability that an attribute tuple in D.

VI. RANDOM FOREST

It is a model which is used as a classifier and uses a decision tree as its most basic module in our project and this has the algorithm which forms a forest using numerous trees.

In this we create numerous of decision trees from the dataset and random forest can be used for both classification and regression.

It can easily handle the values which are missing in the data and thus does not cause a problem with our prediction or model. We can have any number of trees in our model and as we increase our number of trees it will not over fit our model but will fit it to a certain level.

VII. NAÏVE BAYES

It is not just a single algorithm but a family or bunch of algorithms where all of them have the same process. This is a classifier which has an algorithm which is based on the bayes theorem.

p(y/x) = (p(x/y)*p(y)) / p(x).

The simple idea that Naive Bayes follows is we have the input and we have to predict the output class probability.

In this every feature which is being classified is totally independent of each other.

In our project we used the Gaussian nb classifier in which continuous values which are related to points are associated with each feature is distributed according to the Gaussian distribution

VIII. KNN

It is a simple algorithm that takes all the cases and then classifies new cases which are based on the similarity or the relationship of the measures. This has been used in statistical estimations and pattern recognition.

In this project we use knn to classify the images into different categories of the health of air.

VI. RESULTS

Model	Accuracy (%)
Decision Tree	82.35
Random Forest	73.52
Logistic Regression	64.70
Naïve Bayes	53.84
k-nearest neighbors	52.94

In all models

Test Size = 40% of total Data. Train size = 60% of total Data.

VII. CONCLUSION

The air quality estimation based on the image analysis designed and implemented can effectively classify the image into categories

on the basics of health. It can also tell air quality index that is PM value using Logistic Regression with 64.7% accuracy up till now and for classification decision tree upholds highest accuracy of 82.35% and can classify the images into different categories according to the quality of the PM based on the features extracted. This process of feature extraction and possible plotting charts represent that air quality of an image can be linked to various factors and could also be linked to many more unknown factors which are not represented or known in this research.

This air quality estimation motive was to help people to know their surrounding better and take care of their health as one could use this feature of air quality estimation based on image and can find out the health of the air on the basis of which people could take extra care of their own health.

It started with collection of data set which was done by clicking pictures with phone hdr mode and respective pm values was noted from the government site which provides overall pm values for certain region, after collecting the data set features were extracted and final data set of features respective to the images was made.

This final data set of air quality features and respective pm values was used on the various classification models like decision tree, random forest, logistic regression, k-nearest neighbors, naive bayes was used to find out the value of the air quality pm out of the data.

LIMITATION

Some limitation identified after evaluating the current framework, Firstly, we are classifying the images into various categories of air quality health but our data set does not have all the categories of data as it is very rare for the air quality to go to the worse or very unhealthy area and neither it is going to the very good side it usually fluctuates with in 3 or 4 categories, so a real world framework needs to classify in all different categories and should be more flexible.

Second , our decision tree got the best accuracy whereas new models were introduced accuracy began to reduce this might be due to less data, as data will increase the accuracy might go up.

Third, air is composed of much more small particles which our camera cannot detect which might reduce our accuracy because new factors could be introduced to overcome such problems.

FUTURE SCOPE

The future scope of this project is to introduce many various factors which might provide better relationship between air quality pm value and images and given the expectation of more than 5 Billion smartphones in the world by 2020 of which almost a Billion in even small location we do believe that the approach represents a viable additional method for people to know about their surrounding air health and it will provide a better way to keep their health in check.

Images data from a phone may be supplemented with location, respective air quality index values and time information for additional improvements accuracy.

REFERENCES

[1] Yuncheng Li, Jifei Huang, Jiebo Luo, "Using User Generated Online Photos to Estimate and Monitor Air Pollution in Major Cities." University of Rochester, 20 August 2015.

[2] Chenbin Liu, Francis Tsow, Yi Zou, Nongjian Tao, "Particle Pollution Estimation Based on Image Analysis". School of Chemistry & Chemical Engineering, Nanjing University, Nanjing, JiangSu, China, Center for Bioelectronics and Biosensors, Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, Beijing Kinto Investment Management Co., Ltd, Beijing, China, February 2016.

[3] Jian Ma, Kun Li, Yahong Han, Pufeng Du, Jingyu Yang, Image-Based PM2.5 Estimation and its Application on Depth Estimation". Tianjin University, Tianjin 300350, China 2018 .

[4] Xiaoyang Liu, Zheng Song, Edith Ngai, Jian Ma, Wendong Wang, "PM2.5 Monitoring using Images from Smartphones in Participatory Sensing State Key Laboratory of Networking and Switching Technology", Beijing University of Posts and Telecommunications, Beijing 100876, China, 20 August 2015.

[5] A Baklanov, PG Mestayer, Alain Clappier, S Zilitinkevich, S Joffre, A Mahura, and NW Nielsen. Towards improving the simulation of meteorological fields in urban areas through updated/advanced surface fluxes description. Atmospheric Chemistry and Physics, 8(3):523–543, 2008.

[6] Can Li, N Christina Hsu, and Si-Chee Tsay. A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmospheric environment, 45(22):3663–3675, 2011.

[7] Xiang Li, Ling Peng, Yuan Hu, Jing Shao, and Tianhe Chi. Deep learning architecture for air quality predictions. Environmental Science and Pollution Research, 23(22):22408–22417, 2016.

[8] C. Liu, F Tsow, Y. Zou, and N. Tao. Particle pollution estimation based on image analysis. Plos One, 11(2):e0145955, 2016.

[9] Pablo E Saide, Gregory R Carmichael, Scott N Spak, Laura Gallardo, Axel E Osses, Marcelo A Mena-Carrasco, and Mariusz Pagowski. Forecasting urban pm10 and pm2.5 pollution episodes in very stable nocturnal conditions and complex terrain using wrf–chem co tracer model. Atmospheric Environment, 45(16):2769–2780, 2011.

[10] Ana Suarez Sanchez, Paulino Jose Garcıa Nieto, Francisco Javier Iglesias-Rodr´ıguez, and Jose Antonio Vilan Vilan. Nonlinear air quality modeling using support vector machines in gijon urban area (northern spain) at local scale. ´International Journal of Nonlinear Sciences and Numerical Simulation, 14(5):291–305, 2013.

[11] Litian Tao, Lu Yuan, and Jian Sun. Skyfinder: attribute-based sky image search. In ACM Transactions on Graphics (TOG), volume 28, page 68. ACM, 2009.

[12] Haoqian Wang, Xin Yuan, Xingzheng Wang, Yongbing Zhang, and Qionghai Dai. Realtime air quality estimation based on color image processing. In Visual Communications and Image Processing Conference, 2014 IEEE, pages 326–329. IEEE, 2014.

[13] Yibing Zhan, Zhang Rong, Wu Qian, and Wu You. A new haze image database with detailed air quality information and a novel no-reference image quality assessment method for haze images. In IEEE International Conference on Acoustics, 2016.

[14] Zheng Zhang, Huadong Ma, Huiyuan Fu, and Xinpeng Wang. Outdoor air quality inference from single image. In International Conference on Multimedia Modeling, pages 13–25. Springer, 2015.

[15] Zheng Zhang, Huadong Ma, Huiyuan Fu, Liang Liu, and Cheng Zhang. Outdoor air quality level inference via surveillance cameras. Mobile Information Systems, 2016, 2016.

[16] Bin Zou, J Gaines Wilson, F Benjamin Zhan, and Yongnian Zeng. Air pollution exposure assessment methods utilized in epidemiological studies. Journal of Environmental Monitoring, 11(3):475–490, 2009.

[17] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1436–1444. ACM, 2013.

[18] Image-Based Air Quality Estimation? Qin Li1[0086−0147−8992−1850] and Bin Xie2[0086−0138−7312−4038] Central South University, Changsha, China, Central South University, Changsha, China.

[19] Cabaneros, S.M.S. Calautit, J.K.S. Hughes, B.R. A review of artificial neural network models for ambient air pollution prediction. Environ. Model. Softw. 2019, 119, 285–304. [CrossRef]

[20] Huang W, Tan J, Kan H, Zhao N, Song W, Song G, et al. Visibility, air quality and daily mortality in Shanghai, China. 2009; 407(10): 3295–3300. pmid:19275954.

[21] Wang H, Yuan X, Wang X, Zhang Y, Dai Q. Real-time air quality estimation based on color image processing. IEEE Conference on Visual Communications and Image Processing; 2014 Dec 7–10; Valletta, Malta; 2014. p. 326–329. https://doi. org/10.1109/VCIP.2014.7051572