RESEARCH ARTICLE

Particle Pollution Estimation Based on Image Analysis

Chenbin Liu[1], Francis Tsow[2], Yi Zou[3], Nongjian Tao[1,2]*

1 School of Chemistry & Chemical Engineering, Nanjing University, Nanjing, JiangSu, China, **2** Center for Bioelectronics and Biosensors, Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, **3** Beijing Kinto Investment Management Co., Ltd, Beijing, China

* njtao@asu.edu

Abstract

Exposure to fine particles can cause various diseases, and an easily accessible method to monitor the particles can help raise public awareness and reduce harmful exposures. Here we report a method to estimate PM air pollution based on analysis of a large number of outdoor images available for Beijing, Shanghai (China) and Phoenix (US). Six image features were extracted from the images, which were used, together with other relevant data, such as the position of the sun, date, time, geographic information and weather conditions, to predict $PM_{2.5}$ index. The results demonstrate that the image analysis method provides good prediction of $PM_{2.5}$ indexes, and different features have different significance levels in the prediction.

Citation: Liu C, Tsow F, Zou Y, Tao N (2016) Particle Pollution Estimation Based on Image Analysis. PLoS ONE 11(2): e0145955. doi:10.1371/journal.pone.0145955

Editor: Huafeng Liu, Zhejiang Univ, CHINA

Received: September 17, 2015

Accepted: December 10, 2015

Published: February 1, 2016

Data Availability Statement: Data are available from Figshare at http://dx.doi.org/10.6084/m9.figshare.1603556.

Funding: This study was supported by the National Natural Science Foundation of China, grant numbers 21575062, 21327902, http://isisn.nsfc.gov.cn/egrantweb/, Nongjian Tao. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Beijing Kinto Investment Management Co., Ltd provided support in the form of salaries for an author [YZ], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of

Introduction

Among various air pollutants, airborne particulate matter (PM), especially fine particles with diameters less than 2.5 micrometers ($PM_{2.5}$), has a huge adverse effect on human health [1], including increased rates of cardiovascular, respiratory and cerebrovascular diseases [2]. Various techniques have been developed to measure the mass concentrations of PM in air. The most popular methods include filter-based gravimetric methods [3], tapered element oscillating microbalance [4], beta attenuation monitoring [5], optical analysis [6,7] and black smoke measurement [8]. All these methods require sophisticated equipment, which is out of reach for most people. A simple, fast and cheap method to monitor PM in air have the potential to increase public awareness, alert those with respiratory diseases to take proper prevention measures, and provide local air quality data that are not otherwise available.

PM pollution is often characterized by poor visibility, arising from scattering of sunlight by airborne particles. A layperson can tell the difference between clear and hazy sky, but it is much more difficult to distinguish if the hazy sky is caused by PM or fog, and to quantify the degree of PM pollution. Digital cameras are widely available to provide high quality photos, which, together with the ever-increasing computational power for sophisticated image processing with even a mobile device, provide a new opportunity to qualify and analyze airborne particles based on digital photography. Wang et al. [9] examined air quality from light extinction estimated from photographs. However, airborne PM affects a photograph via complex scattering of light, depending on angle and intensity of sunlight, position and angle of the camera,

this author is articulated in the author contributions section.

Competing Interests: Yi Zou is employed by Beijing Kinto Investment Management Co., Ltd. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

distance between the objects and camera, as well as weather conditions, which are reflected in multiple ways: obscuring the images of distant objects, discoloring the sky and reducing the image contrast [10]. Accurate assessment of PM pollution requires us to consider multiple image features and image recording conditions.

Here we report a method to detect and quantify PM pollution by extracting a combination of six image features, including transmission, sky smoothness and color, whole image and local image contrast, and image entropy. We further consider the time, geographical location, and weather condition of each photo, to determine the correlation between PM level and various factors. Based on these features, we build a regression model to predict PM level using photos collected in three different cities, Beijing, Shanghai and Phoenix, about 1 year. Many of today's smartphones are equipped with high quality imaging and powerful computing capabilities, which could be used to detect and quantify $PM_{2.5}$ in air by analyzing the photographs of outdoor scenes.

We arrange the present paper in the following orders. First, the optical model of a hazy image formation was described. Second, according to the model analysis, several features were extracted from hazy images, and the support vector regression was applied to train and predict the PM index. Finally, we evaluate the performance and discuss possible ways to improve the accuracy of the present method.

Principle

PM in air affects an optical image in different ways, but they are all originated from the interactions of light with the airborne particles, mainly via light scattering, including Rayleigh scattering and Mie scattering [11]. Light scattering causes an attenuation of light transmission in air, which can be expressed by the Beer-Lambert law,

$$t = e^{-\beta d} \tag{1}$$

where β is the medium extinction coefficient, which depends on particle size and concentration, and d is the distance of light propagation. This equation indicates that if the extinction coefficients at different wavelengths are determined, then PM concentration can be estimated. The extinction coefficient may be determined from an observed image according to [1,12–14],

$$I(x,y) = t(x,y)J(x,y) + (1 - t(x,y))A \tag{2}$$

where I is the observed hazy image, t is the transmission from the scene to the camera, J is the scene radiance, A is the airlight color vector (see explanation below). As shown in Fig 1, the first term of Eq 2 is the direct transmission of the scene radiance into the camera, which is light reflected by the object surfaces in the scene and attenuated by air before entering the camera. The second term $(1\text{-}t(x,y))A$ is called airlight, which is the ambient light scattered by air molecules and PM into the camera [12–15]. Wang et al. [9] applied the above formula to estimate light attenuation. In the present work, the relationship between transmission value and PM density was evaluated by analyzing ROIs at difference distances. Eq 2 assumes constant atmospheric and lighting conditions, which, in practice, may both change with the weather and position of the sun that vary with the time of the day and season. Additionally, both J and A depend not only on the weather and position of the sun, but also on PM distribution and concentration. The present work considers these varying factors as additional features to improve the accuracy of PM estimation based on images. Fig 1.

The above discussion did not consider color information explicitly, which can also serve as important features for PM estimation based on light scattering consideration. Rayleigh scattering dominates when the particles (including air molecules) are much smaller than the

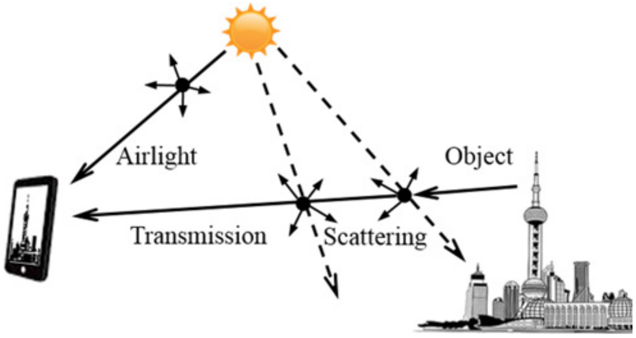

Fig 1. The radiance reaching the smartphone camera is the summation of the transmitted light from the object and airlight from the sun after scattering by air, water and PM in atmosphere.

doi:10.1371/journal.pone.0145955.g001

wavelength of light. It is strongly wavelength dependent, and varies with wavelength (λ) according to λ^{-4}, which is responsible for the blue color of the sky. In contrast, Mie scattering occurs when the size of the particles is comparable to the wavelengths of light, which tends to produce a white glare around the sun when particles are present in air. The combination of Rayleigh and Mie scattering affect the brightness and color saturation of an outdoor image. Conversely, the color and brightness information contains particle concentration and size information, and can be used as distinct features to estimate PM. The present work includes color information as important image features for PM estimation, in addition to light attenuation.

Materials and Method

Data acquisition

To evaluate the capability and accuracy of PM estimation based on image analysis, it is critical to build a database. In the present work, we collected images, as well the date and time of each image, $PM_{2.5}$ index, weather data and geographic location from fixed scenes in three cities, Beijing and Shanghai (China), and Phoenix (U.S.). The Beijing dataset consists of 327 photos (Taken by one of the co-author Yi Zou) of a fixed scene, featuring Beijing Television Tower, captured at almost the same time every morning in 2014. The Shanghai dataset contains 1954 photos of the Oriental Pearl Tower, the icon of Shanghai, from Archive of Many Outdoor Scenes (AMOS) dataset, captured every hour from 8:00 a.m. to 16:00 p.m., from May to December in 2014 [16]. The Phoenix dataset includes 4306 images from AMOS dataset [16], captured every half hour from 9:00 a.m. to 16:30 p.m. in 2014. The $PM_{2.5}$ indices of Beijing and Shanghai were from published documents by the U.S. consulates, which monitor the air quality of the two cities. The air quality of Phoenix was from the published data by U.S. Environmental Protection Agency [17]. Fig 2 shows the $PM_{2.5}$ index range in the three cities. The weather data of the three cities were obtained from Weather Underground (http://www.wunderground.com/) and Weather Spark (https://weatherspark.com/). Precise geographical locations,

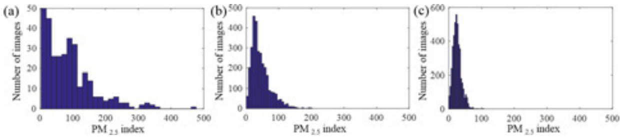

Fig 2. The histogram of $PM_{2.5}$ in different cities. (a) Beijing; (b) Shanghai; (c) Phoenix.

doi:10.1371/journal.pone.0145955.g002

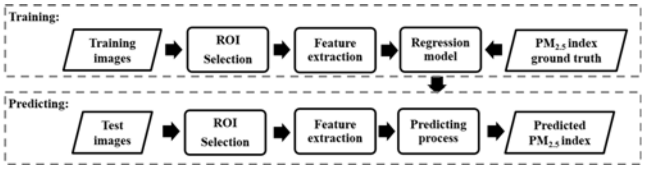

Fig 3. PM estimation via outdoor image analysis.

doi:10.1371/journal.pone.0145955.g003

including longitude, latitude and altitude, were from Google map (https://www.google.com/maps) and elevation map (http://elevationmap.net). Fig 2. (a) Beijing; (b) Shanghai; (c) Phoenix.

Method

After building the database described above, we applied the following image processing algorithm to estimate PM$_{2.5}$ index. As shown in Fig 3, the algorithm mainly consists of the following steps: regions of interest (ROI) selection, feature extraction, regression model training and predicting. We describe the details of these steps below. Fig 3.

ROI selection. The first step is to remove the watermarks in these photos. The watermarks indicate the date and time stamp in our images, which appear in white characters in the first or last few rows. The second step is to build a mask of the sky region, which appears in the images of all the three cities. Fig 4A shows three representative images, one from each city. Both the buildings and background sky are clearly visible. The color images were converted into gray scale images, and then further into binary images with the Otsu method. The Otsu method converts gray scale to binary images by selecting a threshold that minimizes the intra-class variance or maximizing the inter-class variance [18]. In these images, the intensity of the sky is higher than that of the buildings, so the upper part of the binary image is mainly the sky. Fig 4B shows blue lines that mark the boundary between the sky and buildings. To remove the noise caused by the white buildings, we applied the opening operator with a 4×4 disk structuring element, and then filled the holes in the binary image. The third step is to draw the ROIs for the distant buildings manually as shown in Fig 4C, which were used to examine the transmission difference at different distances and PM densities. The ROIs were selected in one image in each dataset and applied to the rest. Fig 4. a) Photos captured at Beijing, Shanghai and

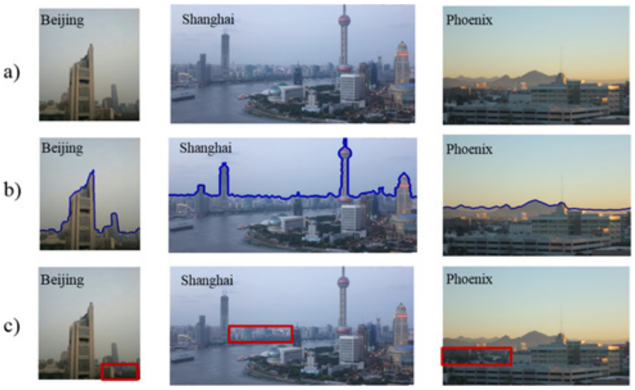

Fig 4. Sample photos in our haze detection database. a) Photos captured at Beijing, Shanghai and Phoenix respectively. b) Boundary lines (blue lines in b) between distant buildings and sky. c) Selected ROIs (red boxes). Reprinted under a CC BY license, with permission from [Yi Zou], original copyright [2014].

doi:10.1371/journal.pone.0145955.g004

Phoenix respectively. b) Boundary lines (blue lines in b) between distant buildings and sky. c) Selected ROIs (red boxes). Reprinted under a CC BY license, with permission from [Yi Zou], original copyright [2014].

Feature extraction. According to the model described in the principle section, transmission can be used to describe the attenuation of scene radiance. To solve for the transmission and thus the attenuation with a single hazy image, the concept of dark channel has been introduced, which assumes the existence of some pixels with zero or very low intensity at least for one color channel in all the outdoor images [13–16]. For a haze–free image *J*, the dark channel is,

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

where J^c is one of the color channels of *J*, and $\Omega(x)$ is a local patch centered at *x*. The airlight can be estimated from the sky or the brightest region, so the transmission can be obtained by,

$$\sim t(x) = 1 - \min_{y\in\Omega(x)}(\min_{c\in\{r,g,b\}} \frac{I^c(y)}{A}) \tag{4}$$

where $I^c(y)/A$ is the hazy image normalized by air-light A, and the second term on the right is the dark channel of the normalized hazy image.

An important assumption in the present model is that the transmission decreases exponentially with the distance between the object in the scene and the camera. We evaluated the transmission by analyzing images of objects at different distances (Fig 5A). Fig 5B shows four ROIs (marked by red boxes) for buildings located at different distances from the camera ($r_1<r_2<r_3<r_4$). The transmission map (Fig 5C) shows the four ROIs at different distances. The average transmission values obtained for the four ROIs are plotted in a semi-logarithmic scale, showing exponential decrease of the transmission with distance, which confirms the validity of the Beer-Lambert law. Fig 5. (a) Schematic illustration of transmission variation with distance. (b) Four ROIs (r_1~r_4) located at increasing distances. (c) The estimated transmission map. (d) Semi-logarithmic plots of transmission curves vs. distance under different haze conditions.

Image contrast is another feature related to PM concentration in air. In fact, human visual perception of air quality is related to image contrast, or visibility [19, 20]. The effect of PM on image contrast can be understood based on Eq 2. As PM concentration increases, the airlight term (second term of Eq 2) arising from light scattering by PM increases. Airlight does not contain information of the scene, which leads to a decrease in the image contrast due to PM.

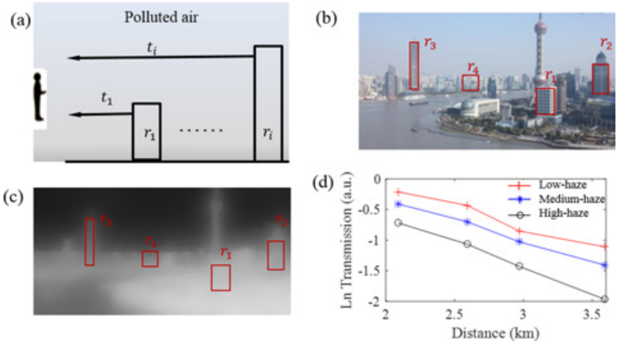

Fig 5. The transmission decreases as the distance or $PM_{2.5}$ index increases. (a) Schematic illustration of transmission variation with distance. (b) Four ROIs (r_1~r_4) located at increasing distances. (c) The estimated transmission map. (d) Semi-logarithmic plots of transmission curves vs. distance under different haze conditions.

doi:10.1371/journal.pone.0145955.g005

Because transmission decreases with the distance between an object and camera, the airlight term contribution also increases with the distance, so the higher is the PM concentration, the lower is the image contrast.

There are many ways to quantify the contrast of an image. A simple way is to use the root mean square (RMS) of an image to describe image contrast. This approach has been found to match with human perception of image contrast [21]. RMS contrast is defined as the standard deviation of the image pixel intensities,

$$RMS = \sqrt{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(I_{ij} - avg(I))^2} \tag{5}$$

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. RMS contrast does not depend on the spatial frequency content, nor convey any information about the spatial distribution of image contrast.

Another image feature that can possibly provide PM information is image entropy, which quantifies information contained in an image, and is related to image texture. Image entropy is defined as,

$$entropy = -\sum_{i=1}^{M} p_i \log_2 p_i \tag{6}$$

where p_i is the probability that the pixel intensity is equal to i, and M is the maximum intensity of the image. As the PM concentration increases, the image increasingly loses its details, and the image entropy decreases.

To determine the image contrast and entropy, we first converted a color images into a gray scale image, and then calculated the image entropy and RMS contrast for the entire image. For comparison, we also determined image contrast of distant buildings by calculating RMS of the selected ROI. Fig 6A–6D show the image of the Oriental Pearl Tower of Shanghai recorded on different days with increasing $PM_{2.5}$. As the $PM_{2.5}$ level increases, the visibility deceases, which is especially clear in a zoomed-in region (red box). The RMS contrast and image entropy both decreases with $PM_{2.5}$ index (Fig 6E). Note that the RMS values of the entire image and the ROI show similar results.

Sky region also carries useful information, such as weather condition. Due to light scattering, the sky is blue on a clear day and gray or white on a hazy or cloudy day. The presence of cloud in the sky can be directly detected from the image, which can be used to differentiate it from the hazy sky. By combining the color and smoothness features, we have attempted to determine clear, partly cloudy, cloudy and hazy days. This information, together with the online weather data, help minimize errors in the estimation of PM due to weather conditions.

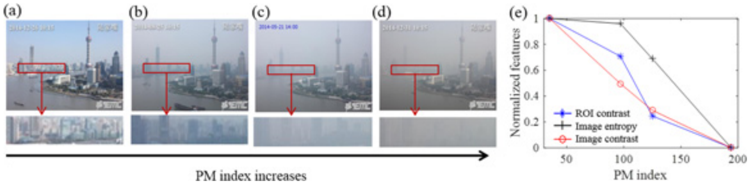

Fig 6. Image features variation with PM index. (a~d): Hazy images showing that the contrast of the building region decreases with the PM index, where the lower panel shows the zooming-in images of the regions marked by the red boxes. (e) The normalized features vs. $PM_{2.5}$ index plot, including ROI RMS contrast (blue), image entropy (black), and image RMS contrast (red).

doi:10.1371/journal.pone.0145955.g006

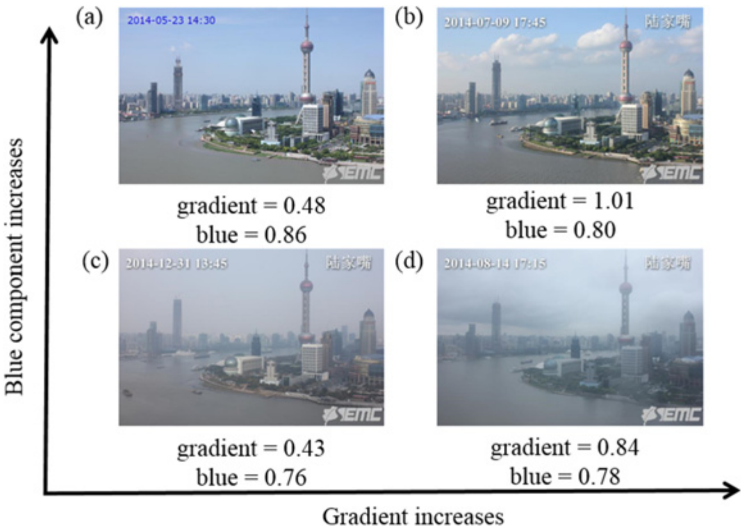

Fig 7. Sky gradient and blue component provide weather information, such as cloud formation. (a) Sunny day; (b) Partly cloudy/sunny day; (c) Hazy day; (d) Cloudy day.

doi:10.1371/journal.pone.0145955.g007

The color of the sky in our study is presented by the average value of the blue component of RGB channels in sky region. The blue channel and sky mask were used to extract the blue component of the sky image (Fig 7). The average of the blue component describes the color of the sky. Fig 7. (a) Sunny day; (b) Partly cloudy/sunny day; (c) Hazy day; (d) Cloudy day.

The smoothness of the sky is defined by the average of the gradient amplitude in the sky region. The image gradient is defined as,

$$\nabla I = \frac{\partial I}{\partial x}\hat{x} + \frac{\partial I}{\partial y}\hat{y} \tag{7}$$

where I is the intensity of the sky region in the image, $\partial I/\partial x$ is the gradient in x direction, and $\partial I/\partial y$ is the gradient in y direction. The average of the gradient amplitude is defined as,

$$avg(|\nabla I|) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2} \tag{8}$$

where $avg(^*)$ is the average value of the two-dimensional image of size M by N. As shown in Fig 7, the blue component in sunny and partly cloudy images is higher than those in the cloudy and hazy images. The averages of the gradient amplitude in the sky regions are higher in cloudy day images than those in sunny and hazy day images.

Both Rayleigh and Mie scattering depend strongly on the angle of sunlight reaching the object, which is determined by the position of the sun for a given scene (Fig 8A). For example, different angles produce different images, such as sunrise and sunset shown in Fig 8B. In the present study, both the scenes and cameras are fixed in positions, so the variation of the scattering is mainly due to the angle of incident illumination, which is determined by the position of the sun. Fig 8. (a) Definition of solar zenith angle. (b) Sample images show that the sky near horizon is red during sunrise and sunset on the same day compared with noon time.

The zenith angle of the sun as shown in Fig 8A indicates the elevation of the sun above the horizon. It is a function of the observer local time, date, longitude, latitude and altitude. In our study, we used solar position algorithm to calculate the solar angle [22]. The main steps include calculations of (1) the earth heliocentric longitude, latitude, and radius with local information,

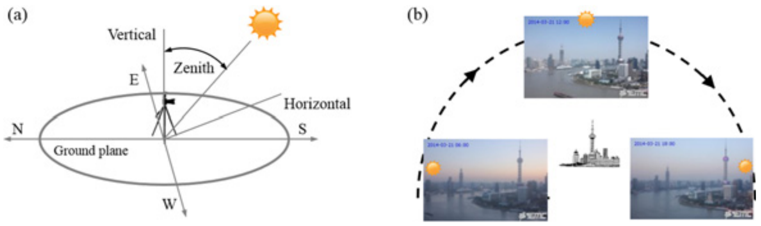

Fig 8. Sky color dependence on solar zenith angle. (a) Definition of solar zenith angle. (b) Sample images show that the sky near horizon is red during sunrise and sunset on the same day compared with noon time.

doi:10.1371/journal.pone.0145955.g008

(2) the geocentric longitude, latitude and the aberration correction, (3) the Greenwich apparent sun longitude and sidereal times, (4) the geocentric sun declination and observer local hour angle, and (5) the topo-centric sun position and solar angle. The solar zenith angle is defined as,

$$\theta_s = \arccos(\sin\varphi\sin\delta + \cos\varphi\cos\delta\cosh) \tag{9}$$

where φ is the local latitude, δ is the sun declination, and h is the local hour angle. In this study, we obtained the longitude, latitude, altitude, coordinated universal time, local date and local time of the captured image. This information is available online, and can also be obtained from the built-in GPS and gyroscope features of smartphones.

Support vector regression. After defining the image features that are possibly related to PM concentration, we determine the relationship between the extracted features and PM concentration using nonlinear support vector machine and kernel to predict the PM concentration. Support vector machines (SVM) [23] have been widely applied in a large number of fields [24–26], including prediction and regression [27, 28]. SVM can also be used to solve nonlinear regression estimation problems, called support vector regression (SVR). In this paper, we used SVR to predict $PM_{2.5}$ index.

The basic idea of SVR is to map input data to a higher dimensional feature space via a function, Φ. A linear function f in the high dimensional feature space formulates nonlinear relationship between input and output data. The regression function can be expressed as,

$$f(w,b) = w\cdot\Phi(x) + b \tag{10}$$

where $f(w,b)$ is the forecasting values, w and b are the function parameter vectors, and Φ is a nonlinear transformation from x to high-dimensional space. The goal of SVR is to minimize function,

$$R_{reg}(f) = \frac{1}{N}\sum_{i=1}^{N}\Theta_\varepsilon(y_i, w^T\Phi(x) + b) \tag{11}$$

where Θ_ε is the ε-insensitive loss function and defined as,

$$\Theta_\varepsilon(y, f(x)) = \begin{cases} |f(x) - y| - \varepsilon, if|f(x) - y| \geq \varepsilon \\ 0, if|f(x) - y| < \varepsilon \end{cases} \tag{12}$$

where ε is a measure of training error, called the radius of the insensitive tube.

In addition, Θ_ε is used to determine the optimal hyper plane in the high dimensional space and minimize the training error between the input data and the ε-insensitive loss function.

Then, SVR minimizes the overall errors,

$$\min_{w,b,\xi^*,\xi}\left(\frac{1}{2}w^T w\right) + C\sum_{i=1}^{N}(\xi_i^* + \xi_i) \tag{13}$$

with the constraints, y_i-$(w\cdot\Phi(x)+b)\leq\varepsilon+\xi_i$, $(w\cdot\Phi(x)+b)$- $y_i\leq\varepsilon+\xi_i^*$, ξ_i^*,$\xi_i\geq 0$, i = 1,2,. . .,N, where ξ_i and ξ_i^* are slack variables, and C is the cost constant. The training vector x_i are mapped to a higher dimensional space with Φ. The radial basis function (RBF) kernel is a popular kernel function used in regression and classification, which can handle the nonlinear relationship well, which is defined as,

$$K(x_i, y_j) = \exp(-\gamma|x_i - x_j|^2) \tag{14}$$

where γ is a kernel parameter. The parameters that dominate SVR are the cost constant, C, and kernel parameter, γ. We performed grid search [28] to determine the optimal values of C and γ ($C = 2^8$, $\gamma = 2^2$). We used the toolbox LIBSVM in MATLAB R2013a and 2-fold cross validation as the regression strategy. We have also performed leave one out cross validation, and found similar results.

Results and Discussion

To predict the $PM_{2.5}$ index with the regression model, we randomly selected half of the samples as training data, and the other half as the testing data, and then we considered the second half as training data and first half as the testing data. For each city's data, we used the 2-fold cross validation and obtained the prediction results. Fig 9 plots the real $PM_{2.5}$ index vs. predicted $PM_{2.5}$ index. The prediction error was evaluated with root mean square error (RMSE), R-squared and F-test. RMSE is defined as,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{15}$$

where $\hat{y}_i$ is the i^th^ forecast value, and y_i is the i^th^ observed value, i = 1,2,. . .,N. R-squared is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - avg(y))^2} \tag{16}$$

where $\hat{y}_i$ is the i^th^ forecast value, $avg(y)$ is the average value, y_i is the i^th^ observed value, i = 1,2,. . .,N. R-squared increases with the agreement between the model prediction and actual result with a maximum value of 1, corresponding to perfect match of the prediction and the actual result. F-test evaluates the null hypothesis that all regression coefficients are equal to zero vs. the alternative that at least one does not. A significant F-test indicates that the observed R-squared is reliable. Fig 9. (a) Beijing; (b) Shanghai; (c) Phoenix.

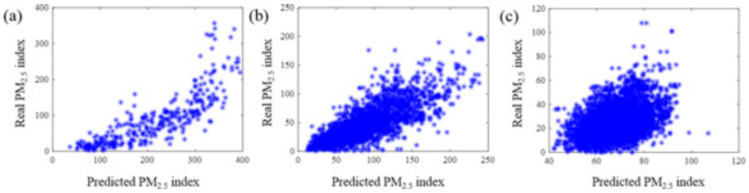

Fig 9. Real $PM_{2.5}$ index vs. predicted $PM_{2.5}$ index plot. (a) Beijing; (b) Shanghai; (c) Phoenix.

doi:10.1371/journal.pone.0145955.g009

Table 1. Assessment of the support vector regression.

dataset	RMSE	R squared	F test
Beijing	42.69	0.65	$p<0.0001$
Shanghai	19.23	0.57	$p<0.0001$
Phoenix	2.89	0.22	$p<0.0001$

doi:10.1371/journal.pone.0145955.t001

Fig 9 shows that the predictions for the Beijing and Shanghai data correlate well with the actual $PM_{2.5}$ indices. In contrast, the correlation for Phoenix is less obvious, which is mainly caused by that $PM_{2.5}$ index in Phoenix falls within a narrow range (0–40). The $PM_{2.5}$ indices in Beijing and Shanghai can reach as high as 340 and 204, respectively, while the $PM_{2.5}$ index in the Phoenix can only reach 38, far below those in Beijing and Shanghai. Table 1 lists RMSE values, which shows that RMSE in Beijing is larger than that in Shanghai, which is caused by the prediction error for high $PM_{2.5}$ index data (over 120), as shown in Fig 9A. R-squared values in Beijing and Shanghai are better than that in Phoenix due to the reason discussed above. F-test in Table 1 shows that R-squared values are reliable.

Weather conditions

In the present work, the effect of different weather conditions was also taken into consideration in $PM_{2.5}$ index prediction. Rainy and snowing days were rare in these datasets, we thus focused on two weather conditions: clear and cloudy days. In the Beijing dataset, 139 and 181 photos were captured on clear and cloudy days, respectively. In the Shanghai dataset, 548 and 1275 photos were captured on clear and cloudy days, respectively. For each weather condition and city, we used the same regression method and 2-fold cross validation as described above.

As shown in Table 2, the prediction error on cloudy days is larger than that in clear days. This is mainly caused by the water droplets in the air, which also scatter light. High humidity can significantly increase the effect of air pollution on visibility. For example, PM attracts water molecules leading to hygroscopic growth in ambient atmosphere [29]. When relative humidity reaches 80%, particles can grow to sizes that cause large increase in light scattering [30].

Since the relationships between particle concentration, relative humidity and visibility are complicated, relative humidity was added as a feature to build the regression model in the following test. Results with and without humidity taken into account as one of the features are shown in Table 3. We can see that the prediction improved after adding the humidity feature for both datasets, especially in the case of Shanghai. This observation correlates with the fact that there are more images captured on cloudy days than that on clear days in Shanghai than in Beijing, and also Shanghai is usually more humid than Beijing. Considering that many of today's newer smartphones are incorporating humidity sensors (e.g. Samsung S4), it is possible to include the humidity as a key feature to estimate $PM_{2.5}$ index.

Table 2. Regression results for different weather conditions.

dataset	RMSE		R squared		F test	
	Clear	Cloudy	Clear	Cloudy	Clear	Cloudy
Beijing	38.90	58.52	0.55	0.45	$p<0.0001$	$p<0.0001$
Shanghai	13.11	25.18	0.58	0.48	$p<0.0001$	$p<0.0001$

doi:10.1371/journal.pone.0145955.t002

Table 3. Regression results with and without humidity as a feature.

dataset	RMSE		R squared		F test	
	Without	With	Without	With	Without	With
Beijing	42.69	40.43	0.65	0.68	p<0.0001	p<0.0001
Shanghai	19.23	14.05	0.57	0.72	p<0.0001	p<0.0001

doi:10.1371/journal.pone.0145955.t003

Feature assessment

To evaluate the features used in the current method, we calculated the distance correlation (DC) and Pearson correlation (PC) between each extracted feature and $PM_{2.5}$ index. The distance correlation is a measurement of dependence between random vectors. DC varies between 0 and 1, representing low and high correlation between an extracted feature and $PM_{2.5}$ index. PC is a measure of linear relationship between two vectors, which varies from -1, indicating a perfect negative linear relationship, to 1, indicating a perfect positive linear relationship.

As shown in Table 4, the transmission has one of the largest correlations with $PM_{2.5}$ index, which is an important indicator of PM concentration. The ROI contrast, whole image contrast and image entropy also show good correlations with $PM_{2.5}$ index, supporting human visual perception that the visibility decreases with increasing $PM_{2.5}$ index. The sky smoothness and color analysis have some correlations with $PM_{2.5}$ index. As for solar zenith angle, the DC and PC values are low, indicating little correlation of the quantity with $PM_{2.5}$ index. The statistical analysis shows that transmission, image contrast and sky features are good features for $PM_{2.5}$ estimation. We also calculated the PC between solar angle and other features. The DC and PC between solar zenith angle and sky smoothness are 0.38 and 0.22 respectively, and the DC and PC between solar zenith angle and sky color are 0.24 and 0.23 respectively. The solar angle shows correlations with sky smoothness and color. The normalization of the related image features with the solar zenith angle could improve the prediction accuracy.

In this study, two feature optimization methods were used to evaluate the feature redundancy and achieve the optimized regression performance. The first one is principle component analysis (PCA) [31]. The second one is sequential backward feature selection (SBFS) [32], and the criterion is RMSE. All eight features were included in the datasets: transmission, ROI contrast, image entropy, image contrast, sky smoothness, sky color, solar zenith angle and relative humidity. We used PCA on the datasets to reduce the dimensions, and then perform the training and regression with 1~7 principle components (PC) respectively. The PCA-SVR results are shown in Fig 10. Compared with the previous results, the regression error in Beijing's dataset (RMSE: 39.08, R squared: 0.69) is smaller using the first 5 PCs. For Shanghai's dataset, the

Table 4. The features and their correlations with $PM_{2.5}$ index in our dataset.

Features	Beijing		Shanghai		Phoenix	
	DC	PC	DC	PC	DC	PC
Transmission	0.81	-0.78	0.60	-0.60	0.32	-0.32
ROI contrast	0.82	-0.76	0.40	-0.40	0.28	-0.29
Image entropy	0.63	-0.54	0.42	-0.46	0.24	-0.24
Image contrast	0.43	-0.43	0.52	-0.55	0.13	-0.12
Sky smoothness	0.34	-0.29	0.32	-0.31	0.28	-0.30
Sky color	0.43	-0.43	0.20	-0.21	0.09	-0.08
Solar zenith angle	0.13	-0.01	0.12	-0.11	0.10	-0.04

doi:10.1371/journal.pone.0145955.t004

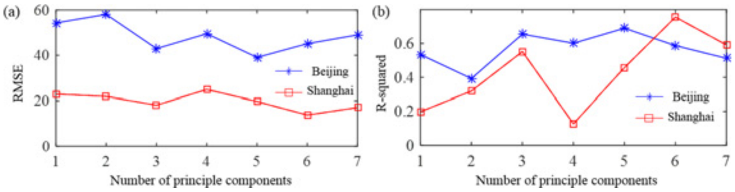

Fig 10. PCA-SVR results for Beijing and Shanghai's dataset. (a) RMSE. (b) R-squared.

doi:10.1371/journal.pone.0145955.g010

regression performance is also better using the first 6 PCs than that without PCA optimization (RMSE: 13.76, R squared: 0.76). From the above results, we can see that with more information, by including combined image features, weather and geographic information, the $PM_{2.5}$ index prediction can be improved. To optimize the feature combination, the PCA-SVR method is efficient to reduce the computation burden and improve the $PM_{2.5}$ index prediction. Fig 10. (a) RMSE. (b) R-squared.

To search for an optimal feature subset, SBFS method was used. Considering there are 8 features in the initial dataset, the method mainly includes: (1) remove the 1st feature in the dataset D_8, and obtain the regression error e_{81}; (2) remove the ith feature in D_8, and obtain the regression error e_{8i}; (3) repeat the above process, so we can get the errors $\{e_{81},e_{82},...,e_{88}\}$; (4) the minimum error e_{8j} in $\{e_{81},e_{82},...,e_{88}\}$ corresponds to the optimized feature subset of 7 dimensions, called D_7, and the jth feature is considered as the least important one in D_8; (5) remove the feature sequentially and repeat steps (1)~(4), so we can get the feature subsets, $\{D_1,D_2,...,D_8\}$, which are considered the optimized feature subset for each dimension; (6) in these feature subsets $\{D_1,D_2,...,D_8\}$, the optimal subset is the one with the minimized regression error. In Beijing and Shanghai's datasets, the optimal features include transmission, ROI contrast, image entropy and sky smoothness. Shanghai has a humid subtropical climate, thus, the relative humidity is one of the most important features. From the regression errors, we can see the feature selection method improves the prediction accuracy. Both PCA and SBFS methods can reduce the feature redundancy and improve the regression with comparable regression errors (Table 5). The PCA method converts the original features into PCs with orthogonal transformation, so PCs not more than the original number of variables can be chosen to be used in the $PM_{2.5}$ index prediction. SBFS method selects an optimal feature subset in the original feature space, so we can improve our understanding in feature contribution and potentially develop new features based on the optimal subset.

The method presented here has several limitations that may be improved in the future. 1) More image features can be included and analyzed, and an optimal combination of different features can be developed with genetic algorithms or particle swarm optimization methods [33], 2) better algorithm, such as deep convolutional neural network can take advantage of the two-dimensional structure of an input image and be used to perform machine learning task, 3) Additional information, such as magnetometer, gyroscope, and thermometer, could be used to determine the camera angle and meteorological parameters.

Table 5. The performance comparison between all the features and the optimized feature subsets for Beijing and Shanghai's dataset.

Dataset	All the features		PCA		SBFS	
	RMSE	R squared	RMSE	R squared	RMSE	R squared
Beijing	40.43	0.68	39.08	0.69	38.28	0.70
Shanghai	14.05	0.72	13.76	0.76	13.65	0.76

doi:10.1371/journal.pone.0145955.t005

Conclusions

We have developed an image-based method to estimate $PM_{2.5}$ index in air. We have extracted various image features, including transmission, image contrast and entropy, sky smoothness and color, and studied their correlations with the reported $PM_{2.5}$ indices in Beijing, Shanghai and Phoenix. We have also examined the effects of solar zenith angle, and weather conditions on the accuracy of the predictions. Using the image and non-image features, we have analyzed a large number of images captured in Beijing (327 images, one per day for 327 days), Shanghai (1954 images, and 8 images per day for 245 days), and Phoenix (4306 images, and 16 images per day for 270 days), and concluded that the method can provide reasonable prediction of $PM_{2.5}$ index over a wide $PM_{2.5}$ index range. We do not expect that the present method will replace the gold standard particle counting apparatus, however, its simplicity and smartphone readiness can help promote air pollution awareness, and alert people with serious respiratory diseases to stay away from suspected polluted air.

Acknowledgments

We thank Yi Zou for assistance with the image data. We thank Dangdang Shao, Yuting Yang and Hui Yu for their discussion.

Author Contributions

Conceived and designed the experiments: NJT. Performed the experiments: CBL NJT. Analyzed the data: CBL NJT. Contributed reagents/materials/analysis tools: YZ FT. Wrote the paper: NJT FT CBL.

References

1. Narasimhan SG, Nayar SK. Vision and the atmosphere. Int J Comput Vision. 2002; 48(3): 233–254. doi: 10.1023/A:1016328200723
2. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: a review of the effects of particulate matter air pollution on human health. J Med Toxicol. 2012; 8(2): 166–175. doi: 10.1007/s13181-011-0203-1 PMID: 22194192
3. Hauck H, Berner A, Gomiscek B, Stopper S, Puxbaum H, Kundi M, et al. On the equivalence of gravimetric PM data with TEOM and beta-attenuation measurements. J Aerosol Sci. 2004; 35(9):1135–1149. doi: 10.1016/j.jaerosci.2004.04.004
4. Ruppecht E, Meyer M, Patashnick H. The tapered element oscillating microbalance as a tool for measuring ambient particulate concentrations in real time. J Aerosol Sci. 1992; 23: 635–638. doi: 0.1016/0021-8502(92)90492-E.
5. Macias ES, Husar RB. Atmospheric particulate mass measurement with beta attenuation mass monitor. Environ Sci Technol. 1976; 10(9): 904–907. doi: 10.1021/es60120a015
6. Anderson TL, Ogren JA. Determining aerosol radiative properties using the TSI 3563 integrating nephelometer. Aerosol Sci Tech. 1998; 29(1): 57–69. doi: 10.1080/02786829808965551
7. Smith JD, Atkinson DB. A portable pulsed cavity ring-down transmissometer for measurement of the optical extinction of the atmospheric aerosol. Analyst. 2001; 126(8): 1216–1220. doi: 10.1039/B101491I PMID: 11534583
8. Muir D, Laxen DPH. Black smoke as a surrogate for PM 10 in health studies. Atmos Environ. 1996; 29 (8): 959–962. doi: 10.1016/1352-2310(94)00370-Z.
9. 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. doi: 10.1109/VCIP.2014.7051572
10. Hyslop NP. Impaired visibility: the air pollution people see. Atmos Environ. 2009; 43(1): 182–195. doi: 10.1016/j.atmosenv.2008.09.067
11. McCartney EJ. Optics of the atmosphere: scattering by molecules and particles. John Wiley and Son, 1975.
12. Fattal R. Dehazing using color-lines. ACM T Graphics. 2014; 34(1): 13. doi: 10.1145/2651362

13. Carr P, Hartley P. Improved single image dehazing using geometry. IEEE DICTA: Digital Image Computing: Techniques and Applications; 2009 Dec 1–3; Melbourne, Australia. 2009. p. 103–110. doi: 10.1109/DICTA.2009.25

14. Fattal R. Single image dehazing. ACM T Graphics. 2008; 27(3): 72. doi: 10.1145/1399504.1360671

15. Koschmieder H. Theorie der horizontalen Sichtweite: Kontrast und Sichtweite. Keim & Nemnich; 1924. 171–181 p.

16. Jacobs N, Roman N, Pless R. Consistent temporal variations in many outdoor scenes. CVPR 2007: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition; 2007 Jun 17–22; Minneapolis, USA; 2007. p. 1–6. doi: 10.1109/CVPR.2007.383258

17. Noble CA, Vanderpool RW, Peters TM, McElroy EF, Gemmill DB, Wiener RW. Federal reference and equivalent methods for measuring fine particulate matter. Aerosol Sci Tech. 2001; 34(5): 457–464. http://dx.doi.org/10.1080/027868201750172914.

18. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979; 9(1):62–66. doi: 10.1109/TSMC.1979.4310076

19. Malm WC, Leiker KK, Molenar JV. Human perception of visual air quality. J Air Pollut Control Assoc. 1980; 30(2): 122–131. doi: 10.1080/00022470.1980.10465927

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. doi: 10.1016/j.scitotenv.2009.02.019 PMID: 19275954

21. Olman CA, Ugurbil K, Schrater P, Kersten D. BOLD fMRI and psychophysical measurements of contrast response to broadband images. Vision Res. 2004; 44(7): 669–683. doi: 10.1016/j.visres.2003.10.022 PMID: 14751552

22. Reda I, Andreas A. Solar position algorithm for solar radiation applications. Sol Energy. 2004; 76(5): 577–589. doi: 10.1016/j.solener.2003.12.003

23. Vapnik V N, Vapnik V. Statistical learning theory. New York: Wiley; 1998.

24. Byun H, Lee S W. Applications of support vector machines for pattern recognition: A survey. In: Lee SW, Verri W, editors. Pattern recognition with support vector machines. Berlin: Springer; 2002. p. 213–236. doi: 10.1007/3-540-45665-1_17

25. dos Santos EM, Gomes HM. Appearance-based object recognition using support vector machines. Proceedings of XIV Brazilian Symposium on. Computer Graphics and Image Processing; 2001 Oct 15–18; Florianopolis, Brazil. Washington, DC, IEEE; 2001. p. 399.

26. Ni KS, Nguyen TQ. Image superresolution using support vector regression. IEEE T Image Process. 2007; 16(6): 1596–1610. doi: 10.1109/TIP.2007.896644

27. Wu CH, Ho JM, Lee DT. Travel-time prediction with support vector regression. IEEE T Intell Transp. 2004; 5(4): 276–281. doi: 10.1109/TITS.2004.837813

28. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM T Intel Syst Tech. 2011; 2(3): 27. doi: 10.1145/1961189.1961199

29. Swietlicki E, Zhou J, Berg OH, Martinsson BG, Frank G, Cederfelt SI, et al. A closure study of submicrometer aerosol particle hygroscopic behaviour. Atmos Res. 1999; 50(3): 205–240. doi: 10.1016/S0169-8095(98)00105-7

30. Day DE, Malm WC. Aerosol light scattering measurements as a function of relative humidity: a comparison between measurements made at three different sites. Atmos Environ. 2001; 35(30): 5169–5176. doi: 10.1016/S1352-2310(01)00320-X

31. Jolliffe I. Principal Component Analysis. John Wiley & Son; 2014.

32. Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recogn Lett. 1994; 15(11): 1119–1125. doi: 10.1016/0167-8655(94)90127-9

33. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19): 2507–2517. doi: 10.1093/bioinformatics/btm344 PMID: 17720704