Collection of Multimodal Environmental Data for Air Quality Estimation

Double Blind
Address Line1
Address Line2
Address Line3
mail@cc.org

Abstract

This paper presents an open platform, which collects multimodal environmental data related to air quality from several sources including official open sources, social media and citizens. Collecting and fusing different sources of air quality data into a unified air quality indicator is a highly challenging problem, leveraging recent advances in image analysis, open hardware, machine learning and data fusion and is expected to result in increased geographical coverage and temporal granularity of air quality data.

Introduction

Environmental data is very important for human life and the environment. Especially, the environmental conditions related to air quality are strongly related to health issues (e.g. asthma) and to everyday life activities. Such data are measured by dedicated stations established by environmental organizations, which are usually made available through web sites and services. Furthermore, the availability of low cost hardware sensors allowed for the establishment of personal environmental stations by citizens. In parallel, the increasing popularity of social media has resulted in massive volumes of publicly available, user-generated multimodal content that can often be valuable as a sensor of real-world events (Aiello et al. 2013). This fact coupled with the rise of citizens interest in environmental issues, has triggered the development of applications that make use of social data for collecting environmental information and creating awareness about environmental issues.

To this end, this paper presents a new platform for gathering and fusing environmental data and specifically Particulate Matter (PM) measurements from official open sources and user generated content (including social media communities). This platform aims to contribute towards individual and collective awareness about air quality and to stimulate sustainable behavior with respect to air quality.

Relevant Initiatives

There are several initiatives including projects and applications that attempt to provide citizens with environment-oriented information collected from different data sources.

Table 1: Relevant initiatives

Type	Name of Initiative
Project	iSCAPE (iscape)
	Amsterdam Smart Citizens Lab (amsterdam)
	PASODOBLE (pasodoble)
	AirTick (airtick)
	PESCaDO (pescado)
	iSPEX (ispex)
	CITI-SENSE (citisense)
	Plume Labs (plumelabs)
Application	Ubreathe (ubreathe)
	World Air quality (worldairquality)
	Banshirne (banshirne)
	AirForU (airforu)
	TZOA (tzoa)
	Clean Air Nation (cleanairnation)
	Air Visual (Air Visual)

Table 1 contains a detailed list of such initiatives. For the sake of space, we shall briefly present only a limited number of the relevant projects and applications: a) iSCAPE that encapsulates the concept of smart cities by promoting the use of low cost sensors and by engaging citizens in the use of alternative solution processes to environmental problems, b) the Amsterdam Smart Citizens Lab that uses smartphones, smart watches, and wristbands, as well as open data and DIY sensors for collecting environmental data, c) AirTick, which estimates air quality in Singapore by analysing large numbers of photos posted in the area, d) PESCaDO that focused on open environmental sources and provided users with personalized information, and e) Plume Labs which develops mobile applications providing personalized air quality forecasts using air quality models.

As far as the applications are concerned, the most interesting are: a) Ubreathe that provides current and forecast air quality as well as health advise for UK, b) World Air quality that reports Air Quality Index for 500 cities around the world, c) AirForU that provides Air Quality Index, hourly updates, one day forecast, historical exposure, and personalised tips, and d) Air Visual that presents historical, real-time and forecast air quality data, including PM10, SO2, temperature using indoor and outdoor sensors.

Compared to the aforementioned initiatives, the proposed platform combines data from various sources in an effort to benefit from the reliability of open official data, the abundance and high coverage of publicly available images posted through social media, the quality and consistency of images captured by users of the platform-oriented mobile app and the reliability of the measurements of low-cost open sensor devices for relatively large numbers of community users.

System architecture

The proposed system will collect particulate matter measurements from various sources that will be processed according to their type (i.e. text or image). The sources that are foreseen are: a) web-based official sources, b) image-based sources, and c) hardware-based sources. The aim of having several different sources is to address the need for both reliable and large in number measurements.

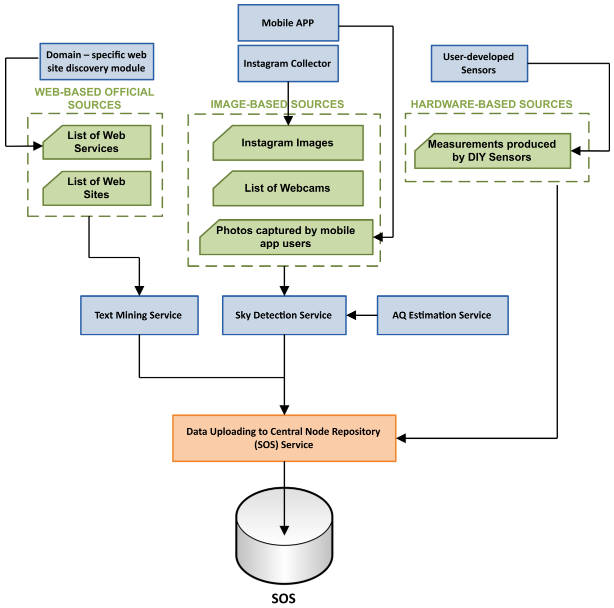

Figure 1: System architecture.

As far as the web-based official sources are concerned, these involve publicly available open data found in environmental web sites, web services and data from webcams. Regarding image-based sources, they include publicly available geotagged images posted through platforms such as Instagram, images captured by the users of the dedicated mobile app and webcams. Finally, hardware-based sources involve low-cost open sensor devices assembled by citizens to monitor the concentration of PM.

The diversity of sources results in obtaining multimodal in nature input data that include images, unstructured and structured text and numeric values. Depending on the type of data, different analysis procedures are foreseen. Specifically, images (coming from social media, the mobile app, and webcams) will be processed using image analysis techniques and image-based air quality estimation that provides an air quality index (e.g. low, high). In the case of web sites, the data is provided in unstructured format and thus text mining is required to extract the target information. Moreover, in the case of web services, the data is provided in structured format and thus no complex text processing is required. Finally, in the case of user-developed sensors, the sources provide numeric values.

Eventually, all collected data are stored into a Sensor Observation Service (SOS) server repository. Figure 1 depicts and overview of the system's architecture.

In the remaining of the paper, we present the data sources used and the techniques for retrieving data from these sources and the post-processing techniques applied.

Data Sources & Retrieval techniques

Web-based official sources

These include web services and web sites that contain environmental measurements. The discovery and indexing of web services and webcams is conducted with the help of air quality experts, while the web sites are discovered using domain specific web search and crawling, which can be divided into two main categories: the first is based on using existing general search engines to access the web and retrieve a first set of results, which are subsequently filtered with the aid of post processing techniques ((Oyama, Kokubo, and Ishida 2004), (Chen et al. 2001)); the second is based on using a set of predefined web sites and expanding them using focused crawling and with the help of machine learning techniques (Zheng, Kang, and Kim 2008).

Image-based data sources

Those include images found in: 1) media sharing platforms such as Instagram, 2) images captured from the mobile app and 3) webcams.

Regarding the images found in media sharing platforms such as Instagram, they offer the advantage of abundance and high geographic coverage. Initially, only geotagged images will be collected and thus for a known rectangle enclosing each city, a set of geo-targeting queries are submitted to the Instagram API is made. However, in case the number of geotagged images is not sufficient (large-scale studies on Instagram suggest that approximately 20 of the images posted to Instagram are geotagged (Manikonda, Hu, and Kambhampati 2014), additional images will be collected by retrieving images tagged with the city name or tagged with known landmarks in the city.

As far as images captured from the mobile app are concerned although they are handled in the same way as social media images, they are expected to be of much higher quality and consistency since their capturing parameters will be controlled by the platform's app.

Finally, webcams can be used as another source of images that depict parts of the skyline of an area of interest. It should be noted that a simple preprocessing of the initial video is required in order to extract frames at specific rate that will be used for representing the video.

Hardware-based sources

These include air quality estimations produced by low-cost open sensor devices assembled by citizens. These estimations will be performed using open hardware and software equipment. More specifically, the widely used Arduino development platform along with the newly introduced PSOC 4 Bluetooth Low Energy (BLE) kits will be programmed for series of widely available PM sensors in order to enable any potential user to perform and contribute air quality measurements. Preconfigured and open software modules will be provided to users along with suggested hardware configurations in order to enable them to implement low-budget air monitoring stations. Data transmission will take place by means of BLE enabled smartphones and an associated Android application. An additional data collection system oriented to users less familiar with electronics will be used. The proposed system will be built around common off-the-shelf materials. The operation principle is to force air (using aquarium or camping pumps) to pass through a paper filter and after a predefined period of exposure to take a photo (by smartphone) of the paper. Then by using computer vision algorithms, developed specifically to handle such images, the colorization of the filter will be translated to PM estimations. More specific, the colorization of the examined filter that will be caused by particulate matters will form unique or small clusters on the top layer of the filter. Thus, these clusters will be projected in image as blobs. Then a blob detector can be applied that will provide the number of identified and marked blob regions which can be correlated with PM concentration (after calibration). The algorithm has the following implementation steps: Thresholding, grouping, merging and blob radius calculation.

Data Analysis

Web Information Extraction

This module involves the extraction of environmental content from web sites and web services. In case of web services the format of the data is well defined and data can be retrieved by simple JSON/XML parsing. As far as web sites are concerned, information extraction from environmental web sites cannot be realized using deep semantic analysis given that most of the information to be retrieved is not reported in text and often there is little linguistic context available. Thus, information extraction can be based on the extraction and transformation of semi-structured web content, typically in HTML format, into structured data. This task typically involves acquiring the page content, processing it by parsing the HTML structure, and extracting the relevant information using regular expressions.

Sky Detection

The module involves two image processing operations: 1) visual concept detection based on low level feature generation and classification for detecting images that contain substantial regions of sky, and 2) localization of the sky regions within the image. As far as visual concept detection is concerned different detectors will be studied including images representation with SIFT, SURF, aggregation using VLAD and training using Logistic Regression (Markatopoulou et al. 2015). Another technique involves representation of images using a pre-trained Deep Convolutional Neural Networks and training with Linear SVMs (Krizhevsky, Sutskever, and Hinton 2012). As far as sky localization is concerned, the selective search technique will be tested that combines the strength of both an exhaustive search and segmentation (Uijlings et al. 2013).

Air Quality Estimation

The module involves the estimation of air quality from user-generated photos or webcams. In general, several studies ((Zerefos et al. 2007), (Zerefos et al. 2014), (Saito and Iwabuchi 2015)) have shown that the color ratio R/G in digital images can be used to derive information about the aerosol content of the atmosphere. Specifically, the system will use the Santa Barbara DISTORT Atmospheric Radiative Transfer Model (SBDART, (Ricchiazzi et al. 1998)) to simulate the R/G ratios for a set of solar zenith angles (SZA) and a set of Aerosol Optical Depths (AOD), the latter approximating the Particulate Matter (PM) load. The R/G ratio will be approximated using the ratio of the diffuse irradiance of two wavelengths (550 nm and 700 nm) rather than the radiance. The resulting R/G ratios will be used to create a 3-D lookup table (Table RG) containing R/G, AOD and SZA. Further, the SZA for each day of the year and each hour of the day will be computed for the geographical latitude/longitude of the urban areas of interest and 3-D lookup tables will be created for each urban area of interest containing SZA, Day of Year (DoY) and Time of Day (ToD). Each geotagged image whose coordinates are within the area of interest will be processed for extraction of information on the mean sky R/G ratio that will be computed automatically for the image, the image coordinates, the DoY and ToD the image was taken.

The table for the respective coordinates will be accessed, receiving as input the DoY and ToD and giving as output the SZA. The Table RG will be accessed, receiving as input the SZA and the image R/G ratio and giving as output the AOD, which, together with the image coordinates will be used to plot the AOD value on a city map.

Data Storage and Indexing: SOS Repository

In order to store and index efficiently the information retrieved from the previously described sources, it is essential to employ a database, which can store efficiently the measurements along with the related information (i.e. date-time, area coverage, and source). Thus, each source can be considered as a sensor, which provides measurements. Hence, a natural option for handling this information is through a Sensor Observation Service (SOS) infrastructure, which provides a generic and flexible means for accessing data produced by sensors (Na and Priest 2007). This includes access to measurements of the sensors, as well as access to information about the observed features of interest and information about the sensor (sensor metadata). The flexibility of the Observation and Reference Model (O&M) can be used for accessing heterogeneous data via a single standard service interface (Cox and others 2011).

Conclusions

The proposed system builds upon the concept of monitoring and fusing heterogeneous and user-generated air quality monitoring resources towards providing reliable measurements (Epitropou et al. 2011). Towards fusing observational data from the aforementioned sources we plan to evaluate methods based on geostatistics, which build upon previous studies demonstrating its feasibility (Denby et al. 2008). More specifically, residual kriging is used to combine the sensor observations with a static base map obtained from a geophysical or statistical model. The resulting map is a combined, value-added product that merges the detailed spatial patterns provided by the base map with the up to-date dynamic information provided by the deployed sensor nodes. Other fusion techniques that could be applied and evaluated include combination of land-use regression techniques with statistical air quality modelling (Johansson et al. 2015). Eventually the fused data will be used to provide personalised services with respect to environmental issues that will raise the awareness of the citizens on air quality and engage them actively in measuring and publishing air pollution levels (Johansson et al. 2015).

References

Aiello, L. M.; Petkos, G.; Martin, C.; Corney, D.; Papadopoulos, S.; Skraba, R.; Goker, A.; Kompatsiaris, I.; and Jaimes, A. 2013. Sensing trending topics in twitter. *Multimedia, IEEE Transactions on* 15(6):1268–1282.

Air Visual. https://airvisual.com/.

AirForU. https://www.uclahealth.org/Pages/AirForU-App.aspx/.

AirTick. https://www.youtube.com/watch?v=l11abvYgvBY.

Amsterdam Smart Citizens Lab. https://waag.org/en/project/amsterdam-smart-citizens-lab.

Banshirne. http://banshirne.com/.

Chen, H.; Fan, H.; Chau, M.; and Zeng, D. 2001. Metaspider: Meta-searching and categorization on the web. *Journal of the American Society for Information Science and Technology* 52(13):1134–1147.

CITI-SENSE. http://www.citi-sense.eu/.

Clean Air Nation (greenpeace India). http://www.greenpeace.org/india/clean-air-nation/.

Cox, S., et al. 2011. Observations and measurements-xml implementation. *OGC document*.

Denby, B.; Schaap, M.; Segers, A.; Builtjes, P.; and Horlek, J. 2008. Comparison of two data assimilation methods for assessing {PM10} exceedances on the european scale. *Atmospheric Environment* 42(30):7122 – 7134.

Epitropou, V.; Karatzas, K. D.; Bassoukos, A.; Kukkonen, J.; and Balk, T. 2011. A new environmental image processing method for chemical weather forecasts in europe. In *Information Technologies in Environmental Engineering*. Springer. 781–791.

iSCAPE. http://horizon2020projects.com/sc-climate-action/h2020-making-cities-sustainable/.

iSPEX. http://ispex.nl/en/ispex/introductie-ispex/.

Johansson, L.; Epitropou, V.; Karatzas, K.; Karppinen, A.; Wanner, L.; Vrochidis, S.; Bassoukos, A.; Kukkonen, J.; and Kompatsiaris, I. 2015. Fusion of meteorological and air quality data extracted from the web for personalized environmental information services. *Environmental Modelling & Software* 64:143–155.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Manikonda, L.; Hu, Y.; and Kambhampati, S. 2014. Analyzing user activities, demographics, social network structure and user-generated content on instagram. *arXiv preprint arXiv:1410.8099*.

Markatopoulou, F.; Mezaris, V.; Pittaras, N.; and Patras, I. 2015. Local features and a two-layer stacking architecture for semantic concept detection in video. *Emerging Topics in Computing, IEEE Transactions on* 3(2):193–204.

Na, A., and Priest, M. 2007. Sensor observation service. *Implementation Standard OGC*.

Oyama, S.; Kokubo, T.; and Ishida, T. 2004. Domain-specific web search with keyword spices. *Knowledge and Data Engineering, IEEE Transactions on* 16(1):17–27.

PASODOBLE. http://lap.physics.auth.gr/pasodoble.asp.

PESCaDO. http://www.iosb.fraunhofer.de/servlet/is/30549/.

PlumeLabs. https://plumelabs.com/.

Ricchiazzi, P.; Yang, S.; Gautier, C.; and Sowle, D. 1998. Sbdart: A research and teaching software tool for plane-parallel radiative transfer in the earth's atmosphere. *Bulletin of the American Meteorological Society* 79(10):2101–2114.

Saito, M., and Iwabuchi, H. 2015. A new method of measuring aerosol optical properties from digital twilight photographs. *Atmospheric Measurement Techniques* 8(10):4295–4311.

TZOA. http://www.tzoa.com/\#homepage.

Ubreathe. http://ee.ricardo.com/ubreathe/.

Uijlings, J. R.; van de Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.

World Air quality. https://itunes.apple.com/us/app/world-air-quality/id396958256?mt=8.

Zerefos, C.; Gerogiannis, V.; Balis, D.; Zerefos, S.; and Kazantzidis, A. 2007. Atmospheric effects of volcanic eruptions as seen by famous artists and depicted in their paintings. *Atmospheric Chemistry and Physics* 7(15):4027–4042.

Zerefos, C.; Tetsis, P.; Kazantzidis, A.; Amiridis, V.; Zerefos, S.; Luterbacher, J.; Eleftheratos, K.; Gerasopoulos, E.; Kazadzis, S.; and Papayannis, A. 2014. Further evidence of important environmental information content in red-to-green ratios as depicted in paintings by great masters. *Atmospheric Chemistry and Physics* 14(6):2987–3015.

Zheng, H.-T.; Kang, B.-Y.; and Kim, H.-G. 2008. An ontology-based approach to learnable focused crawling. *Information Sciences* 178(23):4512–4522.