A Framework for Biodiversity Image Analysis using Machine Learning and Crowdsourcing Knowledge

Loukas Chatzivasili[1,*], Georgia Charalambous[1,*], Maria Papoutsoglou[1,*], Georgia Kapitsaki[1], Ioustina Harasim[1], Eva Chatzinikolaou[2], Georgia Sarafidou[2], Ioannis Rallis[2] and Markos Digenis[2,3]

[1]Department of Computer Science, University of Cyprus, Nicosia, Cyprus
[2]Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Heraklion, Crete, Greece
[3]Department of Environment, Faculty of Environment, Ionian University, Zakynthos, Greece

Abstract
This paper proposes a data-driven methodology framework to enrich existing biodiversity data by collecting crowd-sourced knowledge contributed by citizens in the form of photographs. The framework aims to provide the design of an easy-to-use web interface tool that clusters biodiversity images from multiple data sources. Through its design it will provide the users the ability to identify species by uploading photos to this tool.

Keywords
Machine Learning, Big Data, Citizen science, Biodiversity,

1. Introduction

"All we need is a smartphone with an internet connection." In our days, an Internet-connected mobile phone is an indispensable, highly useful tool for all citizens. A crucial advantage of ubiquitous connectivity is the ability to dynamically create and share content. Our daily experience testifies to the widespread sharing of content that users carry out just for fun; many people however take the extra step and contribute data they capture around them towards wider community causes and into *crowd-sourcing knowledge* [1]. In our work, we focus on crowd-sourced knowledge in the form of biodiversity data provided by citizens. More specifically, visual content captured by citizens, such as the photographs they take from visits to different ecosystems, have significant value and can prove beneficial for various environmental purposes, such as understanding and protecting species, particularly endangered ones.

Collecting crowd-sourcing knowledge towards a scientific goal is directly linked to the concept of *citizen science*, namely the voluntary process by which a person acts as a "human sensor" that collects and/or processes data, mainly for ecological and environmental purposes [2]. In recent years the field advanced further into using machine learning algorithms to automate image processing and analysis for environmental purposes. In our approach, we introduce a framework that uses an annotated dataset of biodiversity images from coastal areas (acquired through pertinent data sources and from citizen scientists in the field) and proposes the use of best-practice image processing tools to automate key processes. We also describe a dynamic web interface through which users can visualize a categorization of images or upload their images to be classified in clusters using the images from the aforementioned dataset.

Why is such a framework needed? The proposed biodiversity image processing framework, part of the SocioCoast project,[1] is part of a larger effort that aims to drastically improve participation in environmental issues by motivating users to share their biodiversity-related images captured in visits to the coast through a versatile, user-friendly mobile app developed and made available by the project. By realizing value through sharing of their images, citizens and tourists will be further incentivized to increase the collected crowd-sourcing knowledge and enhance the image database, one of the main components of the project's overall framework [3]. As more and more images are uploaded into the database we will face the challenge of handling a very large volume of data. In order to provide accurate and dynamic data analytics as the volume of image-storing and processing gets bigger, the proposed framework leverages scalable (big-data) image analytics technologies.

Published in the Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece).
*These authors contributed equally.
lchatz01@ucy.ac.cy (L. Chatzivasili); gchara04@ucy.ac.cy (G. Charalambous); mpapou02@ucy.ac.cy (M. Papoutsoglou); gkapi@ucy.ac.cy (G. Kapitsaki); iharas01@ucy.ac.cy (I. Harasim); evachatz@hcmr.gr (E. Chatzinikolaou); g.sarafidou@hcmr.gr (G. Sarafidou); i.rallis@hcmr.gr (I. Rallis); m.digenis@hcmr.gr (M. Digenis)
© 2023 Copyright © for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://sociocoast.eu/

CEUR-WS.org/Vol-3379/DATUM_2023_6.pdf

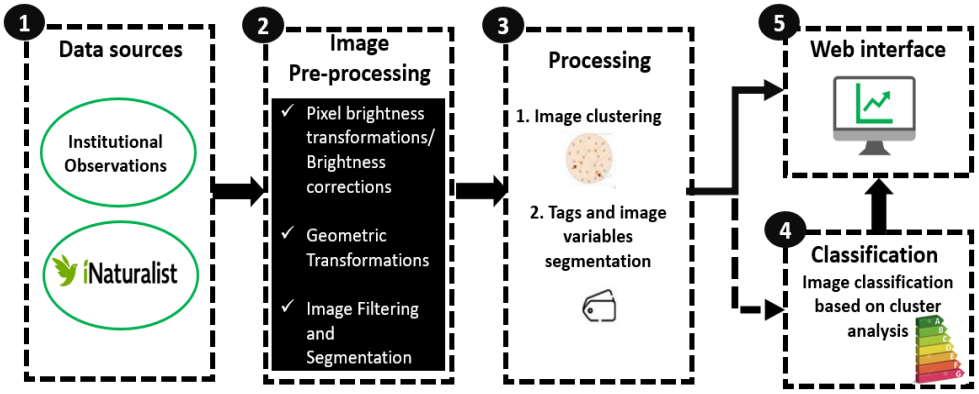

Figure 1: Big data proposed framework

2. Proposed Framework

In this section, we analyze the main components of our proposed framework, which is depicted in Figure 1.

Data Sources

Our proposed framework employs a combination of two types of data sources: (1) institutional (expert) observations and (2) biodiversity data from open-source platforms such as iNaturalist.[2]

The institutional observations utilized in this framework relate to biodiversity species detected in coastal areas and beaches, as recorded by expert personnel from the Hellenic Centre for Marine Research (HCMR) using established measurement tools. The dataset includes 130 species, represented by 196 photographs, collected from 6 distinct coastal areas in Crete. For a subset of the species, multiple photographs were obtained. Along with the species name, a list of properties including: kingdom, phylum, class, category, habitat, and pressures was also provided for each species. These properties were added as tags/labels to the corresponding images and will be used in later stages of the analysis.

To get an overview of our dataset and its important characteristics, we performed a brief descriptive analysis. For example, we identified that 80.76% of the species belong to the "Animalia" kingdom while 15.38% and 3.84% belong to "Plantae" and "Chromista" kingdoms respectively. It can be observed that species that belong to the kingdom "Animalia" are the most diverse in our dataset. We also found that 39% of our species belong to the "Chordata" phylum which suggests that this Phylum has the most diverse representation of species in our dataset.

[2]http://www.inaturalist.org/

The iNaturalist API was used as an additional data source for biodiversity data as mentioned above. In our analysis, we included only citizen observations with a species photograph. In addition with the photographs, we also collected the taxonomic information of the species including the kingdom, phylum, and class. Kingdom, phylum, and class are among the variables that are common between the two different datasets we considered.

Image Pre-processing

The second stage in our framework is the image preprocessing. This stage is important for improving the quality of images we will use in the next phase of the framework, the data processing. We will use a variety of techniques such as image resizing, normalization, enhancement, and filtering to guarantee that the images are consistent in size and format, and that any noise or distortions are removed. Additionally, we will perform "Randaugment", a recent approach which has been implemented to improve accuracy in image data [4].

Processing

The third phase of our proposed framework is the analysis of the collected data. We will start by grouping our species images into clusters based on their similarity. We will apply different clustering methods and then we will compare their performance. The performance of the examined clustering approaches will be evaluated by comparing the HCMR dataset's results to the results collected from the iNaturalist dataset. We will use accuracy, precision, and recall metrics to evaluate the performance of the algorithms. Based on the evaluation results, the best clustering method will be selected and will be fine-tuned to improve its performance.

Once the image clustering algorithms have been applied to the two datasets, we will use the results to perform tags and image variables segmentation following these steps: Firstly, we will identify the clusters by comparing the clustering results of the images to the labels provided in the HCMR's dataset. Next, we will assign tags to each cluster based on the labels assigned to the images. For example, a cluster of images of fish could be assigned the tag "fish". Then, we will analyze the features of the images within each cluster and we will perform image segmentation using variables such as color and shape. After that, we will compare the results of the segmentation to the HCMR's dataset to see if there are any similarities or differences in the clusters. Finally, we will be able to draw useful conclusions about the accuracy and reliability of the clustering algorithm's performance.

Classification

As added value to the image clustering, our framework will offer an optional phase for image classification, using the results of the cluster analysis and photos uploaded from users optionally. As a matter of fact, unsupervised learning techniques in machine learning like clustering, can be used as a prep step to supervised learning techniques like classification. This can be achieved by using the resulting clusters to train a classifier e.g. an artificial neural network or a K-nearest neighbor classifier.

Web Interface

As a final step, the proposed framework will provide a dedicated web interface to visualize the results of image clustering. The interface will enable users to easily browse and view the resulting information. We will have a variety ways of displaying each cluster including photo galley, scatter plots and heatmaps. Users can navigate through the clusters, and they will able to view the images within each cluster along with their associated labels. Users can also upload photos of species via the web interface, which will be analyzed and assigned to a cluster based on their similarity. Once a photo has been assigned to a cluster, users will be able to successfully identify the species' category and access additional information based on its cluster's labels. A mobile version of this interface will also be accessible via the SocioCoast's mobile app to facilitate participation by citizen scientists in the field.

3. Conclusions

In this paper we describe a framework proposed within the SocioCoast project that aims to bring together crowdsourcing knowledge, image sharing, and institutional (expert) image observations and annotation. Through this framework we highlight the value of an annotated dataset from biodiversity experts and the importance that human categorization brings to enhance the value of well-known biodiversity data sources such as the iNaturalist. Furthermore, the design of the proposed framework shows, after detailed image pre-processing steps, how the use of machine learning clustering algorithms could group two different image sources and enhance the one with the other's variables.

Our future research directions follow mainly four directions: Firstly, we plan to test different machine-learning clustering algorithms in order to detect the most accurate for our datasets not only to cluster the available images but also to connect the variables from experts with the variables from iNaturalist's database. Secondly, using the results from the cluster analysis we aim to propose some additional possible variables which could be added to iNaturalist's database. Additionally, through the implementation of the web interface component, we will give the ability to experts or simple users to upload their own datasets and categorize their new content with the currently available one. Finally, another important future aim is to evaluate the implemented proposed framework with a user experience questionnaire during the training sessions which will take place under the dissemination results near the end of SocioCoast project.

Acknowledgments

Research for this paper was undertaken in the course of the SocioCast project of the Cooperation Programme Interreg V-A Greece-Cyprus 2014-2020, co-funded by the European Union (ERDF) and national funds of Greece and Cyprus under the grant agreement No 5050709.

References

[1] M. G. Martinez, Solver engagement in knowledge sharing in crowdsourcing communities: Exploring the link to creativity, Research Policy 44 (2015) 1419–1430.

[2] J. Silvertown, A new dawn for citizen science, Trends in ecology & evolution 24 (2009) 467–471.

[3] M. Papoutsoglou, K. Markakis, L. Chatzivasili, G. Kapitsaki, K. Magoutis, L. Katelaris, C. Bekiari, A framework to enhance smart citizen science in coastal areas, in: Companion Proceedings of the Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1260–1265.

[4] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 702–703.