

Maltese Flora Classifier

Isaac Tanti

Institute of Information & Communication Technology

Malta College of Arts, Science & Technology

Corradino Hill

Paola PLA 9032

isaac.tanti.a104765@mcast.edu.mt

Abstract—With the help of computer vision as a deep learning method, various types of objects in graphical visuals can be classified. The main aim of this research is to deliver a tool that will enable users to identify local flora found on the Maltese islands. In this study, the use of image processing techniques is proposed, namely Convolutional Neural Networks (CNN) as the model which will be trained using YOLOv3 as a single-stage object detection algorithm with a custom-built dataset containing images of local Maltese flora such as *Opuntia Ficus Indica* and *Cheirolophus Crassifolius*. Google Colabs will be used as a platform to build the tool using Python as the main language, OpenCV library, YOLOv3 with darknet backbone and other tools such as Github and Google Drive as repositories. The proposed tool detected targets with a Mean Average Precision of up to 76.13% and we identify situations where this is challenging and recommend future directions.

Index Terms—MCAST, ICT, CNN, YOLOv3, F R-CNN, SSD, Dataset

I. INTRODUCTION

The Maltese islands are filled with life and vegetation and are home to some very rare and unique plants. Unfortunately, only few are knowledgeable of the subject and this might bring risks that one day these plants are either forgotten or simply becomes extinct. With the help of **computer vision** as a **deep learning method**, an image classifier tool can be implemented using an object detection algorithm such as **YOLOv3**, which is used to **train a model** with a custom built **dataset** that contains images of Malta's unique flora. Once that the prediction's rate of accuracy is high enough, the tool can be later installed on a server and used in conjunction with a number of online services such as an **API**, which can be used in web applications to identify our local flora. The following sections give a detailed explanation of which the ideal algorithms for this project to train the model are, how the most suitable algorithm be identified and how the data-sets were obtained and constructed. Finally, the results obtained during this research will be presented followed by a brief break down of these results to determine their accuracy and efficiency.

II. LITERATURE REVIEW

In recent years, deep learning has led to significant improvements in today's technology, more specifically computer vision tasks such as image classification, object detection, and object segmentation. A deep learning-based model depends on feature extraction from deep learning networks, contrasting

to the traditional machine learning base such as DPM (Deformable Part Model) [1] that makes use of HOG (Histogram of Oriented Gradients) [2] as the main feature of learning. A deep learning object detection model is usually designed as a two-stage detector to highlight precision. A two-stage detector type model such as Faster R-CNN (Region-based Convolutional Neural Networks) [3] has significant improvements when compared to the traditional method. Though, two-stage detectors still have one main flaw; in order to operate, they requires a huge number of resources such as computational power and time. As a solution to this problem, Redmond, et al. [4] proposed YOLO (You Only Look Once) as a one-stage deep learning-based object detection model that focuses on speed whilst maintaining accuracy. The main variance from previous works is that YOLO is a one-stage deep learning-based object detection. This model treats the object detection problem as a regression problem which makes YOLO faster by design. The recent release, YOLOv3 [5], holds a significant number of improvements over previous versions. YOLO can be considered as a huge milestone that makes deep learning-based object detection model get closer to real-time tasks.

Soviany and Ionescu had explained in their research paper "Optimizing the Trade-off between Single-Stage and Two-Stage Deep Object Detectors using Image Difficulty Prediction" [6] that currently there are primarily two types of state-of-the art object detectors: two-stage and single-stage detectors. Two-stage detector, such as Faster R-CNN or Mask R-CNN [7] use a Region Proposal Network (RPN) to create regions of interest during the first stage. They later send this region through the pipeline for object classification and bounding-box regression. These models are able to reach the highest accuracy rates, but are generally slower. On the other hand, single-stage detectors such as YOLO and SSD (Single Shot MultiBox Detector) [8] consider object detection as a simple regression problem by accepting an image as input to understand and learn the class probabilities and bound-box coordinates. Single-stage detectors are very quick and efficient, but they tend to be less accurate than two-stage detectors. With that being said, finding a model that delivers the most ideal trade-off between accuracy and speed is not an easy task. Soviany and Ionescu [6], tackled this issue using image difficulty predictions. They concluded that although both types of detectors have their strengths and weaknesses, their use is recommended for different scenarios. Thus even though two-

stage detectors have the most accurate rate of predictions, single-stage detectors would still provide a reasonable accuracy rate at a much faster speed than a two-stage detector.

Mean Average Precision (mAP) is typically used to evaluate the performance of object detectors, which is based on the ‘Accuracy’ ranking of detection rates for each class. ‘Precision’ calculates the accuracy rate of a prediction, whilst ‘Recall’ calculates the rate of true positives in a number of predictions. In order to obtain the Average Precision for each object class, the area under the precision-recall (PR) curve for the detected objects must be recorded. The PR curve is created by first mapping each detected bounding box to the most overlapping ground-truth bounding box, with regards to the IoU (Intersection over Union) measure. However, such mapping can only take place if the IoU is higher than 50%. Furthermore, the detections are organized in descending order of scores. Precision and recall values are calculated each time a new positive sample is recalled. The PR curve is acquired by plotting the precision and recall pairs as lower scored detections are progressively included.

Joseph Redmon and Ali Farhadi [5], compared their brand new algorithm, YOLOv3 in three different variations to other detectors such as different variants of SSD, FPN and RetinaNet.

TABLE I
ALGORITHM COMPARISON.

Algorithm	Inference time (ms)	mAP
SSD321	61	28.0
DSSD321	85	28.0
SSD513	125	31.2
DSSD513	156	33.2
FPN FRCN	172	36.2
RetinaNet-50-500	90	32.5
RetinaNet-101-500	90	34.4
RetinaNet-101-800	198	37.8
YOLOv3-320	22	28.2
YOLOv3-416	29	31.0
YOLOv3-608	51	33.0

Through these findings, Redmon and Farhadi have concluded that although all YOLOv3 variants did not have the best mAP (mean average precision), it still holds an exceptionally good accuracy rate with YOLOv3 having the best speed amongst tested algorithms.

- 1) **TP = True Positive** - An outcome produced by a model which correctly predicted the positive class.
 - 2) **TN = True Negative** - An outcome produced by a model which correctly predicted the negative class.
 - 3) **FP = False Positive** - An outcome produced by a model which incorrectly predicted the positive class.
 - 4) **FN = False Negative** - An outcome produced by a model which incorrectly predicted the negative class.
- **Accuracy (ACC)** - This shows the measure of effectiveness of the machine learning model.

$$Accuracy(ACC) = \frac{TP + TN}{P + N} \quad (1)$$

- **Precision (PPV)** - Precision is the ratio of correctly predicted positive values to the total predicted positive values. This metric indicates the number of correct positive predictions in relation to the number of overall positive predictions.

$$Precision(PPV) = \frac{TP}{TP + FP} \quad (2)$$

- **Recall (TPR)** - The recall is the ratio of correctly predicted positive values to the actual positive values. Recall highlights the sensitivity of the algorithm i.e. out of all the actual positives how many were detected by the program.

$$Recall(TPR) = \frac{TP}{P} \quad (3)$$

III. RESEARCH METHODOLOGY

Problem & Hypothesis

With the help of computer vision as a deep learning method, it may be proposed that various types of local flora can be classified by training a model using an algorithm and data sets.

Aim & Objectives

The aim of this research is to deliver a tool that allows users to identify local flora found in the Maltese islands. With that being said, the main objectives are:

- To identify whether such data sets exist and if not, delve into how they can be constructed
- Analyse existing algorithms and propose a suitable model
- Examine the proposed model using real data to determine its effectiveness

Research questions:

- 1) How can the required **data sets** be obtained or created?
- 2) Which are the ideal **algorithms** for the purpose of this project and how can the most ideal one be identified?
- 3) Are the **results** accurate and efficient?

Chosen algorithm:

With the research done in the previous section, it has been concluded that YOLOv3 is the most ideal algorithm to use in this project. As stated earlier, literature suggests that although YOLOv3 is a single-stage algorithm, it still provides an outstanding accuracy rate with the quickest response time from all the other state-of-the-art algorithms.

Software Requirements:

- **Windows 10** - Windows 10 is the latest version of Windows operating system. Whilst being very user friendly, it still provides key tools such as terminals.
- **Google Chrome** - Google Chrome was used as currently it is one of the fastest and most optimized internet browsers. Furthermore, it is produced by Google, making it safe and reliable. Finally, as a Google product, it goes hand in hand with other Google services such as Google Drive and Google Colab.

- **Google Colab** - This project was built on a GPU environment provided by Google Colabs. Google Colabs was used as it out performs any of the computers at my disposal and it is also a free to use online cloud service.
- **Google Drive** - Google Drive was used as the main storage location for the project. Apart from this service being available free of charge, it is a cloud service provided by Google, therefore it is highly secured thus making it safe from any attacks or data corruption. Finally, both the environment and the storage are provided by the same vendor, Google, making it slightly more user friendly.
- **Python 3** - Python 3 is currently one of the most popular scripting languages and it has a very low yet powerful learning curve. Google Colab is a Linux based environment, thus it goes hand in hand with Python 3.
- **Github** - Github is used as a data repository and as a way to download libraries for python to install and use.
- **WinZip** - WinZip is a file compression tool which is used to compress files such as images for datasets. This software was used because Google drive has a limited amount of storage, hence datasets were compressed before being uploaded to take the least amount of space possible.
- **Labellmg** - Labellmg is an open source python annotation tool developed by Github user 'Tzutalin'. It was used to create the annotations for the images in the datasets by manually selecting the region for it's corresponding class.
- **OpenCV** - OpenCV is a highly optimized library with focus on real-time applications. OpenCV library was installed on the environment as the algorithm used makes heavy use of it.
- **Nvidia Cuda** - Nvidia Cuda ToolKit is used to fully optimize the GPU so it can perform in the best possible way to achieve the best possible results.
- **YOLOv3 with Darknet backbone** - YOLOv3 algorithm was used by installing its backbone, Darknet. YOLOv3 is a real-time, state-of-the-art object detection system.
- **Fatkun Batch Downloader** - Fatkun Batch Downloader is a Google Chrome extension which was used to scrape images from Google search and finally mass download them. This tool was used to gather images to construct the datasets.

Hardware Requirements:

Image classification algorithms tend to require a lot of resources such as computational and graphical power. Albeit, since both Google Colab and Google Drive are cloud based services, a powerful computer was not required to fully implement the prototype. As an image capturing device, a OnePlus7 Pro was used as it has a high resolution camera.

- Desktop Computer:
 - CPU: Intel Core i5-9400F
 - RAM: 32GB
 - Storage Technology: NVMe SSD
 - GPU: Nvidia GeForce GTX 1070 8GB

- OnePlus 7 Pro
 - Main Camera: 48MP, f/1.6

Research Pipeline

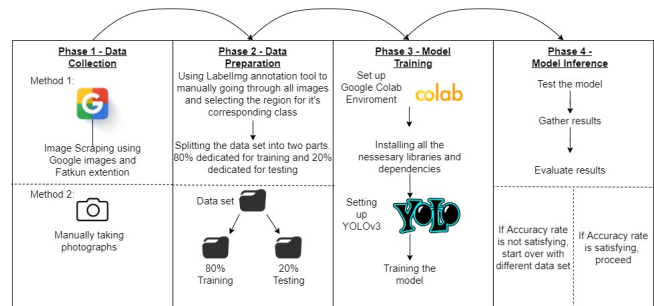


Fig. 1. Research Pipeline

Phase 1 : Data Collection

When it comes to data sets which includes images of flora which is mostly found on the Maltese islands, it is quite impossible to find a ready built data set publicly available on the internet. Therefore, other traditional methods such as scraping images from search engines and manually capturing photos of indigenes' plants were employed. For the purpose of this project, two kinds of plants were selected: 'Opuntia Ficus Indica' which is more commonly known as 'Bajtar' and 'Cheirolophus Crassifolius' which is more commonly known as 'Widnet il-Bahar'. These two kinds of flora were selected for two reasons: because training the model takes a lot of time and as more images and classes are added, the training duration drastically increases and finally, although the leaf of 'Opuntia Ficus Indica' is much larger than that of 'Cheirolophus Crassifolius', they share a very similar shape, thus it brings a challenging task to the chosen algorithm.

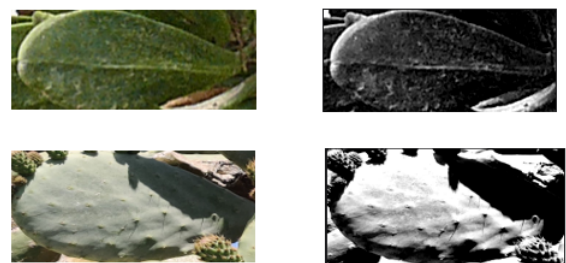


Fig. 2. Leaf comparison

Phase 2: Data Preparation

Once around a hundred images of each plant were obtained, the annotation labeling process commenced. Labellmg was used, using python as a compiler to manually label each and every image to create annotations. This was done by selecting a region on the image and linking it with its corresponding class such as 'Cheirolophus Crassifolius flower' or 'Cheirolophus Crassifolius bulb'. Once all the images were processed,

they were divided into two folders; one for training which contains 80% of the total images, and one for testing which contains the other 20%.



Fig. 3. Image annotation tool

Phase 3 - Model Training

Once the data set was obtained, model training commenced by first creating an environment on Google Colab which uses GPU as the main runtime type. After the environment was configured, Google Drive was connected and authenticated for further use. Using a Git command, darknet was downloaded into the environment solution and the make file was altered with the help of a script, which enables and uses GPU and OPENCV libraries. Darknet was then compiled using the 'make' command whilst the dataset was compressed and uploaded from a local machine to Google Drive, and finally copied and extracted into the Darknet data folder in Google Colab's solution. An additional backup folder was created in Google Drive which served as a location which YOLOv3 will use to save the trained model later on. Two specific python functions were included into the environment which helped to upload and download files directly from and to Colab's solution, and an additional function was added which assisted in displaying images in Colab's terminal. Now that the algorithm was installed and configured, four very important files were uploaded: the first file contained the configuration for the YOLOv3 algorithm, the second included all the classes names whilst the third contained the dataset details and configuration file path, images directory path, image annotations directory path and the backup folder path which was created earlier. Finally, the fourth file contained a python script that used the first three files to generate a text document containing all the paths of each and every image with their corresponding annotation. Once that was done, a pre-trained convolutional layer model provided by Darknet developers, was downloaded using Git. The purpose of this pre-trained model is to serve as a base to the model which was about to be trained. Finally, one small but crucial script was executed in the browser console which stopped Google Colab from timing out after a couple of minutes with no user interactivity, thus stopping the training process. This allowed the environment to train the model for countless hours with no interruption whatsoever. Once everything was in order, the train command was executed to commence the model training. The model training process

duration depends on the size of the data set. During that time, one could monitor the progress by displaying a graph (Fig. 4) of loss over iterations. The total elapsed time for training the model was around 17 hours.

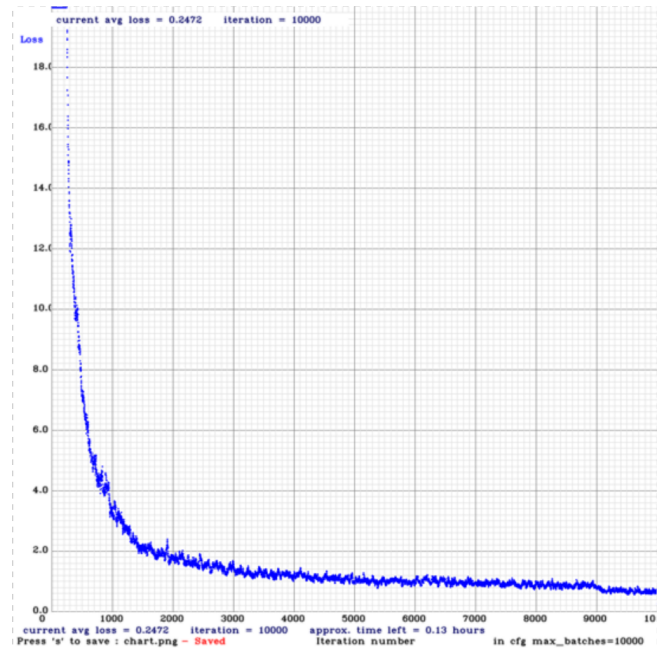


Fig. 4. Graph of Loss over Iteration

Phase 4 - Model Inference

Once that the training was complete, a command was executed to start the testing to obtain the mAP and IoU percentages. A total of ten YOLOv3 weights were produced, and each and every weight had different percentage ratings. After the results were observed, the mAP and IoU of each and every weight was recorded. The initial results were not satisfying enough for image classification, therefore the whole process had to be redone and the dataset was recreated to contain more images with better and clearer graphics than the previous ones. After the second model was trained and results were observed, the mAP and IoU drastically increased, with the highest percentage exceeding the recommended rates which were listed on the official YOLOv3 documentation. Therefore, the training phase was concluded.

```
IoU threshold = 50 %, used Area-Under-Curve for each unique Recall
mean average precision (mAP@0.50) = 0.750353, or 75.04 %
Total Detection Time: 2 Seconds
```

Fig. 5. YOLOv3 Output of mAP calculation

IV. FINDINGS & DISCUSSION OF RESULTS

The previous sections of this research paper identified the requirements needed to construct the proposed solution. Additionally, the steps which show how it was implemented were also documented, with the respective results recorded in the tables below.

TABLE II
FIRST DATASET WEIGHTS MAP COMPARISON.

Weight	mAP %
yolov3_custom_weight_1000	51.00%
yolov3_custom_weight_2000	53.75%
yolov3_custom_weight_3000	53.45%
yolov3_custom_weight_4000	56.13%
yolov3_custom_weight_5000	52.78%
yolov3_custom_weight_6000	54.51%
yolov3_custom_weight_7000	53.52%
yolov3_custom_weight_8000	52.31%
yolov3_custom_weight_9000	52.53%
yolov3_custom_weight_10000	55.04%
Average	53.502%

In table II, the results of the first model are presented and one can clearly observe that they aren't satisfying enough for image classification. With an average mAP of 53.502% and the highest mAP rate of 56.13%, they weren't high enough to deem the tool accurate and efficient. After the whole process was redone with an updated dataset, the following results shown in table III were obtained.

TABLE III
SECOND DATASET WEIGHTS MAP COMPARISON.

Weight	mAP %
yolov3_custom_weight_1000	71.00%
yolov3_custom_weight_2000	73.75%
yolov3_custom_weight_3000	75.15%
yolov3_custom_weight_4000	76.13%
yolov3_custom_weight_5000	71.78%
yolov3_custom_weight_6000	74.91%
yolov3_custom_weight_7000	73.52%
yolov3_custom_weight_8000	72.33%
yolov3_custom_weight_9000	74.93%
yolov3_custom_weight_10000	75.04%
Average	73.854%

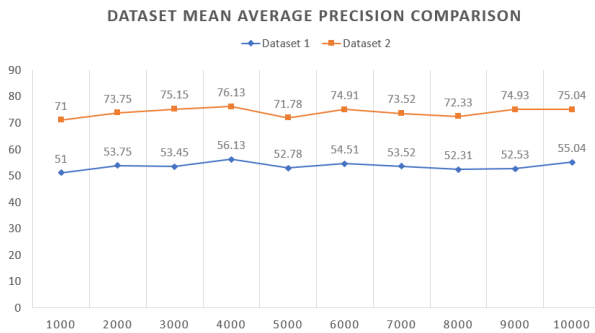


Fig. 6. Dataset comparison

The above graph (Fig. 6) shows that both datasets were compared and a drastic increase in accuracy can be clearly observed. This does not only show that the updated dataset had effected the accuracy of the model, but also it has proven that the larger the dataset, the more accurate the model is.

From the results in the above tables, one can also notice the similarities in both datasets, as the 4th weight from both datasets had the highest mAP percentage. Thus concluding that later variants of weight do not necessarily imply that a better mAP will be obtained. In figure 5 one can also notice that at the final testing stage, the entire test only took 2 seconds to be conducted. This is clear proof that YOLOv3 is indeed a very fast algorithm.

V. CONCLUSION

Finally, with regards to the findings and results in the previous sections, a conclusion answering the initial research questions can be formed.

Research Questions covered:

1: How can the required datasets be obtained or created?

The internet is the largest source of information which one may use in order to gather any type of data. Having said that, pre-constructed datasets are widely available to download for free, but unfortunately, the specific type of dataset that was required was not yet available. Therefore, a dataset was constructed by scraping images from search engines and manually taking photos of the subject class, which would be manually and individually analyzed ensuring that the region of interest which corresponds to a class would then be recorded and labeled.

2: Which are the ideal algorithms for the purpose of this project and how can I identify the most ideal one?

The two main types of image classification algorithms are single-stage and two-stage state-of-the-art algorithms where single-stage algorithms provide a fast result with a moderate rate of accuracy, whilst two-stage algorithms provide the best accuracy rates with a much slower result. For the sake of this research paper and the prototype, a single-stage algorithm, more specifically YOLOv3, was deemed to be the most ideal to use as it is the fastest amongst all the algorithms researched in the section above while still providing a decent accuracy rate.

3: Are the results accurate and efficient?

Taking into consideration the quality of datasets used, an average mAP rate of 73.854% produced by the trained model was obtained. With regards to the algorithm's documentation, the achieved rate exceeded the documented ideal rate for the best results. Finally, by using the fastest single-stage algorithm at this time, one can conclude that the results are both accurate and efficient.

VI. FUTURE DIRECTIONS

This tool has unlimited potential and can be used in several productive ways. The plan is to keep working on new datasets which will include more varieties of local flora. Another future possibility is adding local fauna such as indigenes insects, birds, lizards, mammals and fish. The use of an additional database containing flora and fauna information, combined with this tool and an Augmented Reality approach could result in a fun and educational smartphone application for primary and secondary schools used to educate their students.

ACKNOWLEDGEMENT

The author is truly grateful and would like to thank the following people for assisting in the proof reading of this research paper: Michaela Micallef, Michela Cutajar. The author would also like to thank Christian Cassar, Keane Schembri, Matthew Vella and all other people who contributed in finding necessary research for this paper.

REFERENCES

- [1] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, p. 1627–1645, 2010.
- [2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05)*.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, p. 1137–1149, Jan 2017.
- [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] Redmon, Joseph, Farhadi, and Ali, "Yolov3: An incremental improvement," Apr 2018. [Online]. Available: <https://arxiv.org/abs/1804.02767>
- [6] P. Soviany and R. T. Ionescu, "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2018.
- [7] Gkioxari, Piotr, Girshick, and Ross, "Mask r-cnn," Jan 2018. [Online]. Available: <https://arxiv.org/abs/1703.06870v3>
- [8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, p. 21–37, 2016.