Traditional Maltese Food and Snacks Classifier

Keane Schembri
129, Qroll Street
Birzebbugia
Malta
keane.schembri.a106267@mcast.edu.mt

***Abstract*—With Malta's tourism sector on continuous growth annually and being a very small island with rich culture, a tourist can engage and interest their selves into our culture. Using a classification tool such as this study proposes makes it easier for a foreigner to learn basic concepts in Malta's traditional food. This has been achievable using YOLOv3 neural network to detect and classify different articles of food, from cheeselet to deep-fried date roll's, labelled in Maltese as 'ġbejniet' and 'imqaret' consecutively. On a dataset of 350 images, the model has achieved over 89% classification accuracy on images with multiple objects, and 100% classification accuracy on images with a single object to classify.**

***Index Terms*—Convolutional Neural Network (CNN), YOLOv3, Algorithm, Image Annotation**

I. Introduction

Using computer vision as one of deep learning methods, it is possible to classify various traditional Maltese food. The idea behind this study is to have a more user-friendly introduction and an easier access for foreigners to traditional Maltese food, how to identify different articles of food and their label. Hence this study discusses and explores the field of computer vision and the reasoning behind a regression-based YOLOv3 algorithm high speed's performance. This study also highlights the collection and construction of a dataset to train a YOLOv3 model and what such dataset consists. Finally, this study will question and answer whether or not computer vision is able to reach a high level of accuracy, which algorithm is best suited to classify food and how a reliable dataset is constructed. The significance of this research is to provide a tool that assists people who are not familiar with traditional Maltese food such as tourists to distinguish between various food easily.

II. Literature Review

Computer vision has become a very demanding technology that contributed to various sectors to identify various aspects; for instance, computer vision contributes a lot in the medical field and helps practitioners such as nurses and doctors in the process of identifying, study or treat diseases [1]. This technology is often found in agriculture, the tourism sector, sport, manufacturing amongst others. This technology works by taking an image, understand what the image is and interpret its content with a view to solve tasks and resolve queries.

Study relevant to pedestrian recognition [2] discusses how computer vision includes two operations; to filter the image by reducing noise that could hinder the object segmentation algorithms performance, and the second operation is subtracting the background from the region of interest (ROI). The main concern about this method is that the captured image conditions may have different lighting, shadows and level of reflections that might influence the overall performance of subtracting the background. This step could be quite challenging and plays a very important role in object detection systems, a solution is to use a sliding window that moves across an image and collect window view to the histogram of oriented gradients (HOG) however, this is an expensive operation. Therefore a less expensive solution are algorithms based on deep neural networks that are proven to have a better success rate.

Other relevant studies take into variable shape, size and the colour of the object to be detected. During a study regard vehicle intelligence and understanding of its surroundings [3], the researcher starts to question various aspects in relation to a typical driving scenario. Other variables discussed where the dynamic environment a vehicle is found in, thus the researcher takes the lighting of the environment and the background into consideration.

Throughout, various researchers have done intense study in the field of computer vision, some of which have solved major tasks. One vital study in the medical sector has researched the use of computer vision in medical imaging to localize and examine human long bones. Images also come in the form of sound waves which are emitted by a B-scan sensor and are translated into a digital image more understandable to the human eye. However, a B-scan sensor is not able to distinguish between long bones and other specimens such as muscles, soft tissue, veins and other internal organs under the skin tissue. Hence, this study implements a deep learning system to be able to recognize bone specimens. This study discusses various other algorithms but makes use of YOLOv3 convolution neural network algorithm to differentiate the said specimens in ultrasound images [1]. Similarly, my research makes use of the same concepts this study does which classifies long bones, or in my case traditional Maltese food, in digital images.

YOLOv3 Algorithm

YOLOv3 is a regression-based one-stage target detection algorithm just like SSD [4] which can detect and classify

objects in an image at the same time [5]. YOLOv3 divides a digital image into S x S non-overlapped grids and determines whether each grid has a target to predict bounding boxes around [6]. Considering that YOLOv3 is a single end-to-end network that performs feature extraction, location and classification in a network, makes the algorithm extremely fast i.e. at 320x320 YOLOv3 runs at 22ms at 28.2mAP [7].

A convolution neural network (CNN) is a deep learning network that assigns importance to an image such as weights and can differentiate one aspect of an image from another. YOLOv3 is one of the fastest object detection algorithms that makes use of a CNN in real-time detection without loss of accuracy.

Different Algorithms

A study of vehicle detection [3] highlights and compares other object detection algorithms with a view of identifying which is the better algorithm.

Computer vision is categorized into two types; region-based such as R-CNN and Faster RCNN and regression-based such as SSD and YOLOv3. Region-based target detection algorithms extract region proposals from top to bottom of a given image as a candidate region for the model to analyse, extract features in each proposal, classifies them and performs border regression on the region proposals thus this process takes a lot of time since it is divided into several stages. R-CNN is a region-based algorithm that uses a selective search algorithm that makes the process run too slow therefor making the overall object detection run slow.

Faster RCNN was introduced to improve the quality of R-CNN. This algorithm makes use of region proposal network (RPN) to select the candidate regions instead of the selective search algorithm which results in faster detection and enables end to end detection by a neural network. RPN reduces the number and improves the quality of region proposals [7]. This method has a limited ability to extract other features and generalization hence a neural network-based algorithm using the deep learning model solves several problems.

A one-stage algorithm, SSD, results that the algorithm achieves better detection on smaller objects since the SSD generates more anchor points to make the object position more accurate [3].

Dataset

Study in the medical field using computer vision [5] reseraches cancer tumour in mammogram images using a convolutional neural network. This study explains the use of a digital database for screening mammographs to train and tests the CNN system. The dataset used contains 2,620 cases of breast imaging, including four mammograms for each case made up from non-malignant to critical cases in order to have accurate results. Indeed, deep learning requires a complex dataset to train a model to be more accurate. During this study, researchers have discussed techniques of augmentation to the training dataset where such techniques require for the dataset to generate new instanced using different transformation methods such as rotation of images, translation and scale [8]. Considering that very limited images of traditional Maltese food are available online to scrape which results in a small dataset, this technique could improve the training process of my model by expanding and generating more images. These researchers discuss how the original dataset was augmented by rotating the mammograms three times with angles 90, 180 and 270.

III. Methodology

To start the process of classifying different articles of Maltese food I had to establish which articles I would like my model to classify hence I choose the traditional Maltese cheeselet commonly known as 'ġbejniet' and another Maltese snack deep-fried date roll's commonly known as 'imqaret'. Knowing which article of food my model can classify I could start the training process which requires a large set of annotated images to train. However, since classifying these types of food is not a very popular concept it was difficult to find a ready annotated dataset that consists of hundreds of images of cheeselet and date roll's, thus I had to create my own. I have adopted to use the YOLOv3 algorithm to classify food due to being the better algorithm to classify food, further analysis is found in the literature review section. Thus answering the second research question of this study.

Dataset

To acquire several images I made use of an online tool called *'Fatkun Bach Download'*, this enabled me to search and scrape several images found on Google then filter through each image and decide their relevance under some circumstances i.e. angle of the article of food, lighting, resolution and if other food articles overlapping each other. Filtering through the scraped images I was able to delete irrelevant images downloaded unintentionally or duplicates of the same image. This process causes the dataset to have a smaller number than desired to train the model which results in a less accurate prediction. To resolve I have also populated the dataset with images I took myself of both food articles to classify.

Annotation process enables the images to be labelled for the algorithm of choice to train the model. Before feeding an annotation tool with images to label, YOLOv3 requires images to all be the same size, this step was easily achieved by using *'Batch Resize'* online tool to resize all images at 416 x 416. Making use of *'BBox-Label-Tool'* is a great tool that annotated images ready for YOLOv3 by providing a JPG, PNG or JPEG image as input and highlights a border around the article of food. In this process, I made sure to highlight each individual

article of food alone as grouping several foods altogether will influence the training. Image annotation is another way of filtering through each image and deciding their relevance. After highlighting the regions of food, the tool will provide a text file populated with the image label which includes the index of class (*'ġbejniet', 'imqaret'*), x-axis, y-axis, width and height.

This process answers the first research question of this study; how can a reliable dataset be obtained or constructed.

Configure File and Train

Training a YOLOv3 model requires a great amount of processing power hence I trained my model using Google Colab as this environment provides Googles own GPU's for use. I opted for this method as my personal computer doesn't have a graphics card thus training a model would have taken a long time. Before training the model, colab required the setup of installing darknet and uploading the dataset to darknet where two text files were generated for training and testing. Configuring darknet enabled me to provide the environment with the number classes and the destination to their names, destination of the backup folder to save weights and paths to train and test files. Finally, it was time to train in darknet using *darknet53.conv.74* according to the *yolov3_custom.cfg* file which consists of configurations and arithmetic solutions related to our model.

Testing and Results

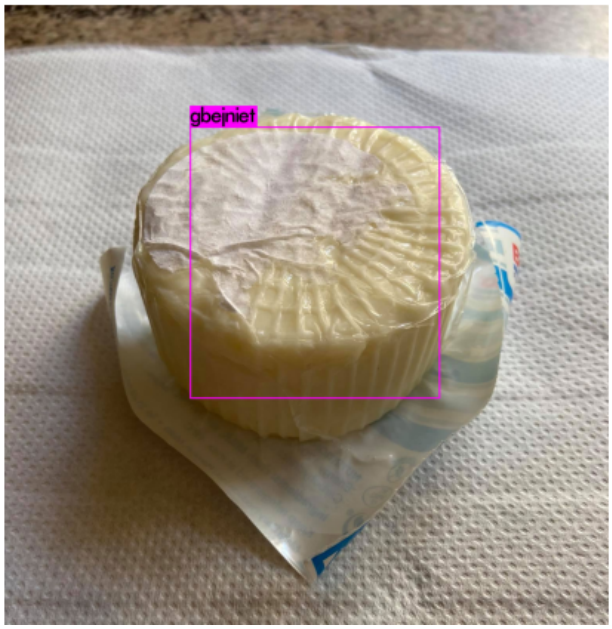
The testing dataset contains a set of both cheeselet and deep-fried date roll's images from different angles and some images made up of multiple objects to detect. Classifying images using *yolov3_custom_final.weights* which are the trained weights using the food dataset. After testing with test sample images, the food classifier model was able to reach over 89% accuracy when an image consists of multiple objects as seen below.

```
test_images/(6).jpg: Predicted in 90.348000 milli-seconds.
gbejniet: 89%
gbejniet: 98%
gbejniet: 100%
gbejniet: 92%
Unable to init server: Could not connect: Connection refused

(predictions:4826): Gtk-WARNING **: 16:38:59.584: cannot open display:
```

When the object to detect is relatively large the model detects good results, in fact the model was able to reach a 100% accuracy rate when classifying images with a single object as seen below. This analysis answers the third research question of this study; will image processing be able to reach a high level of accuracy.

```
Done! Loaded 107 layers from weights-file
test_images/(2).jpg: Predicted in 90.143000 milli-seconds.
gbejniet: 100%
```



When testing with low-resolution images that consist of a too complicated detection scene the model doesn't detect anything.

```
Done! Loaded 107 layers from weights-file
test_images/(85).jpg: Predicted in 89.985000 milli-seconds.
Unable to init server: Could not connect: Connection refused

(predictions:4797): Gtk-WARNING **: 16:38:08.743: cannot open display:
```

IV. Evaluation

The model was trained using 350 images and has reached a level of 89% accuracy. However, the dataset could improve in view of having a more accurate model. Improving a dataset would consist of having a larger number of images made up from different augmentation techniques such as image rotation and scale thus allowing a more diverse dataset. Improving the model would consist of having other relevant food to classify. Noted that bounding boxes could have also been drawn covering a larger area of the object in target, this is done during the annotation process.

V. Conclusion

research paper proposes to build a traditional Maltese food classifier based on the YOLOv3 algorithm combined with a neural network to be applied in scenarios to help the tourism sector in Malta. Results of the model reflect a very high accuracy level and can be improved by training of larger and more complex datasets including different food articles. Hence, for future work, the algorithm could expand on a larger range of food articles and improve accuracy by training on a better dataset to be able to discriminate better between different objects in the same image.

Acknowledgement

Gratitude towards my mentor who assessed and supervised this work Mr Daren Scerri for his excellent guidance, supported by the MCAST ICTAR research team and guided by Mr Ivan Briffa.

References

[1] R. A. F. Lazuardi, T. Karlita, E. M. Yuniarno, I. K. E. Purnama, and M. H. Purnomo, "Human bone localization in ultrasound image using yolov3 cnn architecture," in *2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*. IEEE, 2019, pp. 1–6.

[2] P. Tumas and A. Serackis, "Automated image annotation based on yolov3," in *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*. IEEE, 2018, pp. 1–3.

[3] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Q. Liu, "A comparative study of state-of-the-art deep learning algorithms for vehicle detection," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 82–95, 2019.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[5] K. Djebbar, M. Mimi, K. Berradja, and A. Taleb-Ahmed, "Deep convolutional neural networks for detection and classification of tumors in mammograms," in *2019 6th International Conference on Image and Signal Processing and their Applications (ISPA)*. IEEE, 2019, pp. 1–7.

[6] Y. Zheng, H. Bao, X. Xu, N. Ma, J. Zhao, and D. Luo, "A method of detect traffic police in complex scenes," in *2018 14th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2018, pp. 83–87.

[7] F. Miao, Y. Tian, and L. Jin, "Vehicle direction detection based on yolov3," in *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2. IEEE, 2019, pp. 268–271.

[8] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital mammography*. Springer, 1998, pp. 457–460.