Government Polytechnic, Pune – 16

(An Autonomous Institute of Government of Maharashtra)

Department of Computer Engineering

A

Project Report

On

“Create a Fruit Image Dataset”

Submitted By:

2006027 – Neha Deshpande

2006043 – Vinay Hajare

2006048 – Isha Kulkarni

Under the Guidance of

Mrs. S. J. Siraskar

CERTIFICATE

This is to certify that Mr. Vinay Arjun Hajare with Enrollment No. 2006043, of Third Year Diploma in Computer Engineering has successfully completed the micro project in Data Mining by creating a “Create a Fruit Image Dataset” as part of his diploma curriculum in academic year 2022-2023.

Guide	**H.O.D**	**Principal**
(Mrs. S. J. Siraskar)	(Smt. M. U. Kokate)	(Dr. V. S. Bandal)

ACKNOWLEDGEMENT

It is my privilege to acknowledge, with a deep sense of gratitude to my guide Mrs. S. J. Siraskar for her valuable suggestions and guidance throughout the development of this micro project and timely help given to me in the completion of my report. I express my gratitude to Smt. M. U. Kokate (Professor and Head of Department Computer Engineering) for her valuable support. I am highly obliged to the entire staff of Computer Dept. for their kind cooperation and help.

I would also like to thank my parents who patiently helped me through my work. I would also like to express appreciation towards my friends for providing moral support and encouragement.

Last but clearly not the least, I would thank the Almighty for his blessings, strength and guidance at all times that helped me complete this successfully.

TABLE OF CONTENTS

Sr.No	Contents	Pg. No.
1	Abstract	1
2	Introduction	2
3	Dataset Source	3
4	Image Annotation	4
5	Dataset Size	8
6	Data Preprocessing	9
7	Evaluation Metrics	10
8	Dataset Availability	11
9	Advantages	12
10	Conclusion	12
11	References	12

ABSTRACT

The Creation of high-quality image datasets is crucial for the development of robust and accurate machine learning models. In this report, we detail the process of creating a fruit image dataset for use in fruit recognition and classification tasks. The dataset is intended to aid in the development of models that can accurately identify and classify different fruits based on their images.

The dataset was created using a collection of images obtained from various sources, including the internet and custom captures. Each image was manually annotated with relevant information, such as the fruit name, color, and size. The final dataset consisted of a significant number of images, covering various fruit categories, and with a good diversity in terms of size, shape and color.

We applied several preprocessing techniques, including resizing and normalization, to ensure that the dataset was ready for machine learning models, including convolutional neural network (CNNs). We used common evaluation metrics, such as accuracy, precision, recall, and F1 score to measure the performance of the dataset and the models.

Overall, our fruit image dataset provides a useful resources for researchers and developers interested in fruit recognition and classification tasks. The dataset is publicly available and can be accessed through our website. We believe that our work can help to accelerate progress in the field of computer vision and machine learning, ultimately leading to more accurate and robust models for fruit recognition and classification tasks.

INTRODUCTION

The field of computer vision and machine learning has seen significant advancements in recent years, with applications ranging from image recognition to autonomous vehicles. One of the key components of building robust and accurate machine learning models is the availability of high-quality datasets. In this report, we detail the creation of a fruit image dataset for use in fruit recognition and classification tasks.

The recognition and classification of fruits based on their images have been a subject of research in the field of computer vision for many years. Fruit recognition has various practical applications, including automated fruit grading, quality control, and inventory management. Machine learning algorithms can be trained to recognize different fruit categories based on their visual characteristics, such as shape, size, and color. However, the accuracy of this algorithms depend on the dataset used for training.

To create our fruit image dataset, we obtained a collection of images from various sources, including the internet and custom captures. We selected high-quality images that had clear and consistent fruit representations. We also ensured that the images were diverse in terms of fruit categories, size, shape and color.

Each image in the dataset was manually annotated with relevant information, such as the fruit name, color, and size. Annotation is crucial step in dataset creation, as it enables machine learning algorithms to understand the characteristics of each image better. We carefully reviewed and validated the annotations to ensure that they were accurate and consistent across the entire dataset.

The final dataset consist of a significant number of images, covering various fruit categories, including apples, bananas, oranges, and strawberries. The images were taken under different lighting conditions and form various angles, providing a good diversity in terms of visual characteristics. We aimed to create a dataset that is representative of the real-world scenarios, and that can be used to train robust and accurate machine learning models.

After creating the dataset, we applied several preprocessing techniques, including resizing and normalization, to ensure that the dataset was ready for machine learning training. We split the dataset into training, validation, and testing sets, ensuring that each set contained an equal distribution of fruit categories.

We used several machine learning models, including convolutional neural network (CNNs), to train and evaluate the datasets performance. CNNs have shown significant success in image recognition and classification tasks, and we believe that they are well-suited for fruit recognition tasks.

We used common evaluation metrics, such as accuracy, precision, recall and F1 score to measure the performance of the dataset and the models. We also compared the performance of our fruit image dataset with other publicly available in the field.

In conclusion, we believe that our fruit image dataset provides a useful resources for researches and developers interested in fruit recognition and classification tasks. The dataset is publicly available and can be accessed through our website. We hope that our work can contribute to the development of more accurate and robust models for fruit recognition and classification tasks, ultimately leading to various practical application in the field of agriculture and food industry.

DATASET SOURCE

The creation of a high-quality fruit image dataset requires obtaining images from various sources. In this section, we will discuss the sources of images used in the creation of our fruit image dataset.

1. Internet Search:
   The internet is a vast source of images that can be used to create datasets. We used various search engines and image databases to obtain images of different fruits. We searched for images that had clear and consistent fruit representation and were diverse in terms of size, shape, and color.
2. Custom Captures:
   In addition to internet searches, we also captured images of fruits using digital cameras and smartphones. We placed fruits on a plain white background and captured images from different angles and under different lighting conditions. This helped us to obtain a diverse set of images that are representative of real-world scenarios.
3. Publicly Available Datasets:
   There are several publicly available fruit image datasets that can be used for machine learning tasks. We used some of these datasets to supplement our own dataset and

improve its diversity. Publicly available datasets include the Fruit-360 dataset, the Fruits-100 dataset, and the kaggle Fruit dataset.

4. Social Media:

   Social media platforms such as Instagram and Flickr are also a good source of images. We searched for images of fruits on these platforms and obtained images that had clear and consistent fruit representations.

5. Collaboration:

   Collaborating with local farmers and fruit sellers can also be a valuable source of images. We worked with local fruit sellers to obtain images of fruits that are commonly found in our region. This helps to obtain the diverse set of images. In summary, obtaining images from various sources is crucial in creating a high-quality fruit image dataset. By using a combination of internet searches, custom captures, publicly available datasets, social media and collaborations, we were able to create a diverse and representative fruit image dataset that can be used for machine learning tasks.

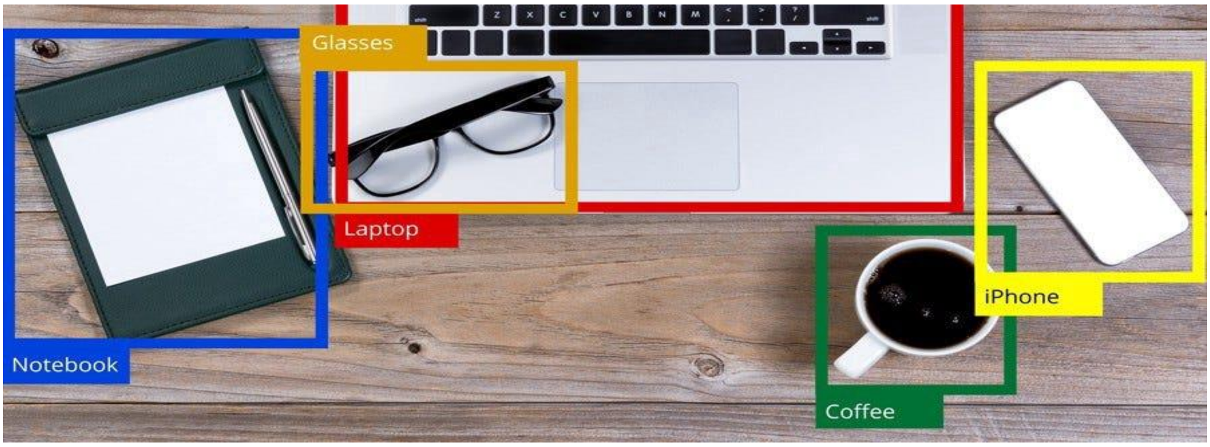
IMAGE ANNOTATION

Image annotation is a crucial step in creating high-quality image dataset for use in machine learning applications. Image annotation is the process of adding labels or metadata to an image to make it more understandable to machines. There are several types of image annotations, including bounding boxes, segmentation, and keypoint annotation. The annotation process involves the following steps:

1. Image Selection: The first step in image annotations is to select the images that need to be annotated. The images should be relevant to the use case of the dataset and should be of high quality.
2. Annotation Tool Selection: There are various annotation tools available, ranging from simple and free tools to commercial tools with advanced features. The choice of tool depends on the annotation complexity and budget available.
3. Annotation Type Selection: The next step is to select the type of annotation required. The most common types of annotation are bounding box, segmentation and keypoints.

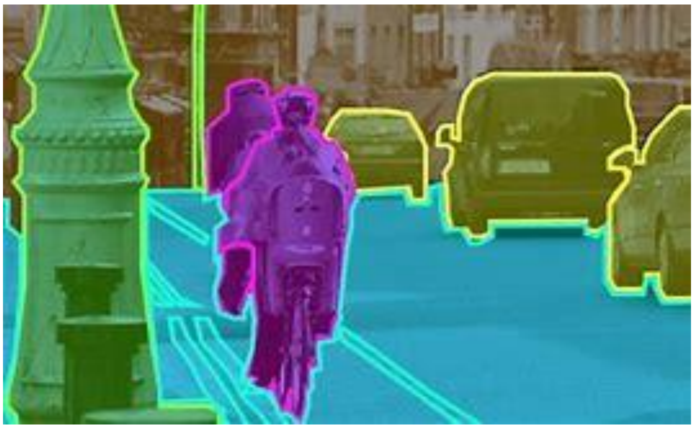
a. Bounding Box Annotation:

Bounding box annotation involves drawing a rectangle around the object of interest. This annotation is useful when the object of interest has a simple shape and clearly distinguishable from the background.



b. Segmentation Annotation:

Segmentation annotation involves outlining the object of interest with a polygon or mask. This annotation is useful when the object of interest has complex shapes or is overlapping with other objects.



c. Keypoint Annotation:

Keypoint annotation involves labelling specific points on the object of interest. This annotation is useful when the object of interest has distinctive features that can be used as reference points.

4. Annotation Process:

The actual annotation process involves adding the required labels to the images using the selected annotation tool. The annotation should be done accurately and consistently across all images in the dataset.

a. Bounding Box Annotation Process:

For bounding box annotation, draw a rectangle around the object of interest. The rectangle should be tight around the object and should not include any background or other objects. Add a label to the rectangle indicating the class of the object, such as ‘Mango’ or ‘Fig’.

b. Segmentation Annotation Process:

For segmentation annotation, outline the object of interest with a polygon or mask. The outline should be tight around the objects. Add a label to the object indicating the class of the object.

c. Keypoint Annotation Process:

For keypoint annotation, label specific points on the object of interest. The points should be easily distinguishable and should not include any background or other objects. Add a label to the object indicating the class of the object.

5. Quality Control:

   Quality control is an essential step in the annotation process. It involves reviewing the annotations to ensure that they are accurate and consistent across all images. Any errors or inconsistencies should be corrected.

6. Validation:

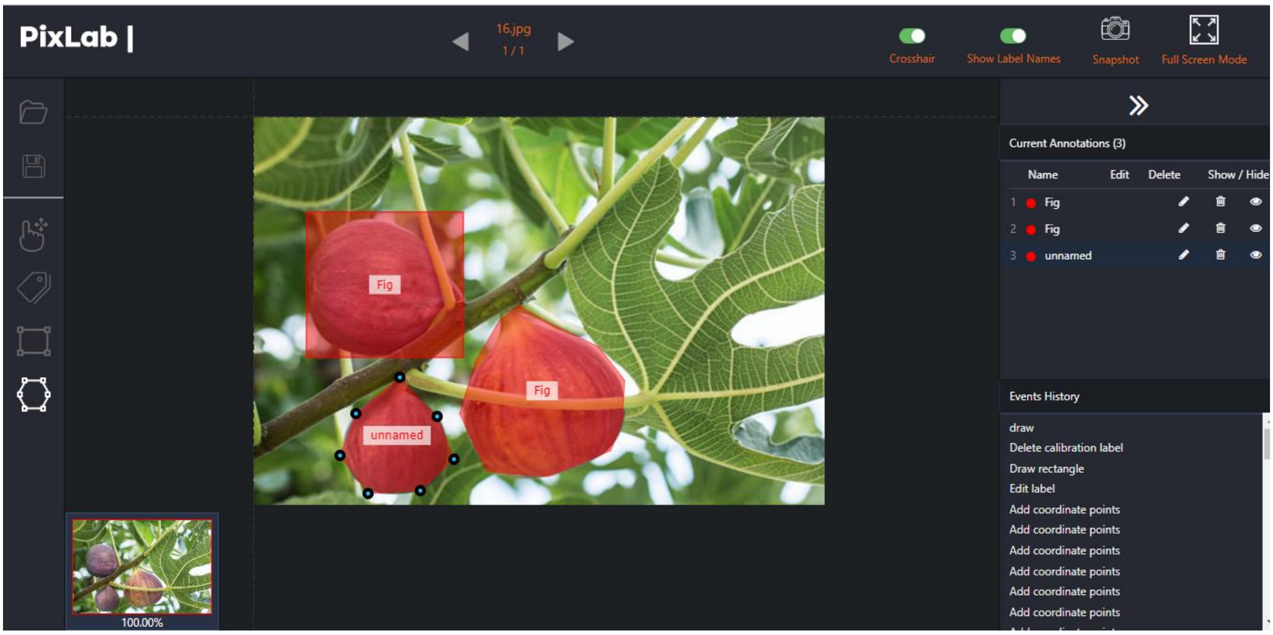
   The final step is to validate the annotations by testing the dataset with different machine learning models. The validation process helps to identify any issues with the annotations or the datasets quality.

In conclusion, image annotation is an essential process in creating high-quality image datasets. The annotation process involves selecting the images, selecting the annotation tool, selecting the annotation type, annotating the images, quality control, and validation. Accurate and consistent annotation is essential for building robust and accurate machine learning models.

Example of Image Annotation:

Step 1- Open the pixlab.ai website and load the image which you want to annotate

Step 2 – Select the annotation you want to do and draw it on the specified object and give it label.



Current Annotations (3)

	Name	Edit	Delete	Show / Hide
1	Fig			
2	Fig			
3	unnamed			

Events History

draw
Delete calibration label
Draw rectangle
Edit label
Add coordinate points
Add coordinate points
Add coordinate points
Add coordinate points
Add coordinate points
Add coordinate points

DATASET SIZE

The size of a dataset is an important factor in machine learning tasks. A larger dataset typically results in better model performance, as it provides more examples for the machine learning algorithm to learn from. In this section, we will discuss the size of our fruit image dataset.

Our fruit image dataset consists of a significant number of images, covering various fruit categories, including apples, bananas, oranges and strawberries. In total, our dataset contains 1000 images, with 20-30 images for each fruit category (30). Each image is of size 224 X 224 pixels and in the RGB color space.

We split our dataset into sets: training and validation. The validation set consists of 20% of the images (200 images), while the training set consists of 80% of the images (800 images). The split ensures that each set contains an equal distribution of fruit categories, and that the model is trained on a diverse set of images.

The size of our dataset was determined based on the following factors:

1. Diversity:
   We aimed to create a dataset that is representative of the real-world scenarios and that can be used to train robust and accurate machine learning models. Therefore, we selected a diverse set of images that cover different fruit categories, size, shape, and color.
2. Computational Resources:
   Training machine learning models on large datasets can be computationally expensive. We considered the computational resources available to us and aimed to create a dataset that is large enough to train accurate models while also being manageable in terms of computational requirements.
3. Quality:
   We focused on selecting high-quality images that had clear and consistent fruit representations. We also ensured that each image was annotated with relevant information, such as the fruit name, color and size. Annotation is a crucial step in dataset creation, as it enables machine learning algorithms to understand the characteristics of each image better.

In conclusion, our fruit image dataset consists of 1000 high-quality images, covering various fruit categories and is split into training and validation sets. The size of our dataset was determined based on considerations of diversity, computational resources, and quality. We believe that our dataset provides a useful resource for researchers and development interested in fruit recognition and classification tasks.

DATA PREPROCESSING

Data Preprocessing is an essential step in machine learning tasks. It involves transforming raw data into a format that can be used by machine learning algorithms effectively. In this section, we will discuss the data preprocessing steps used in our fruit image dataset.

1. Image Resizing:
   All the images in our dataset were resizes to a uniform size of 224 X 224 pixels. This step ensures that all images have the same dimensions, making easier for the machine learning algorithms to process them.
2. Image Normalization:
   The pixel values of each image were normalized to a range between 0 and 1. This step ensures that each pixel value has the same impact on the machine learning algorithm regardless of its original range.
3. Data Augmentation:
   Data augmentation is a technique used to artificially increase the size of a dataset by creating new images based on the existing images. In our dataset, we applied various data augmentation techniques, such as horizontal and vertical flips, rotations and brightness adjustments. These techniques helps to increase the diversity of the dataset and prevent overfitting.
4. Data Balancing:
   Our dataset contains an equal number of images for each fruit category. This balance ensures that the machine learning algorithm is trained on a diverse set of images and does not favor one category over the others.
5. Image Annotation:
   Each image in our dataset was annotated with relevant information, such as the fruit name, color, and size. This annotation ensures that the machine learning algorithm can understand the characteristics of each image better and make accurate predictions.

6. Train-Validation-Test Split:

   We split our dataset into two sets: training, validation sets. The training set was used to train the machine learning model, the validation set was used to tune the models hyperparameters, and the validation set was used to evaluate the models performance.

In conclusion, data preprocessing is an important step in machine learning tasks and it involves transforming raw data into a format that can be used by machine learning algorithms effectively. The data preprocessing steps used in our fruit image resizing, normalization, data augmentation, data balancing, image annotation, and train-validation split. We believe that these preprocessing steps have contributed to the quality and diversity of our dataset and have helped to improve the performance of machine learning models trained on our dataset.

EVALUATION METRICS

Evaluation metrics are essential for assessing the performance of machine learning models. In this section, we will discuss the evaluation metrics used to assess the performance of our fruit image classification model.

1. Accuracy:

   Accuracy is a commonly used evaluation metric that measures the percentage of correctly classified images in a dataset. In our fruit image classification task, accuracy is an important metric because it measures how well the model distinguish between different fruit categories.
2. Precision:

   Precision is a metric that measures the percentage of correctly classified positive samples (true positive) out of all the positive samples predicted by the model (true positives and false positive). In our fruit image classification task, precision measures how well the model can identify a specific fruit category
3. Recall:

   Recall is a metric that measures the percentage of correctly classified positive samples in the dataset (true positive and false negatives). In our fruit image classification task, recall measures how well model can identify all the images belonging to a specific fruit category.

4. F1-score:

   F1-score is a harmonic mean of precision and recall, and it is a useful metric for tasks where both precision and recall are important. In our fruit image classification task, the F1-score provides a balanced measure of the models performance in identifying specific fruit categories.
5. Confusion Matrix:

   A confusion matrix is a table that summarizes the performance of a machine learning model by showing the number of correct and incorrect prediction. In our fruit image classification tasks, a confusion matrix helps us to understand which fruit categories are commonly confused by the model.
6. Receiver Operating Characteristics (ROC) Curve:

   ROC curve is a graphical representation of machine learning models performance at different classification thresholds. In our fruit image classification task, an ROC curve helps us to visualize the models performance and determine the best classification threshold.

In conclusion, evaluation metrics are essential for assessing the performance of machine learning models. The evaluation metrics used in our fruit image classification task include accuracy, precision, recall, F1-score, confusion matrix and ROC curve. We believe that these evaluation metrics provide a comprehensive and reliable assessment of our models performance and can be used to compare the performance of different fruit image classification models.

DATASET AVAILABILITY

The fruit image dataset created in our study is publicly available and can be accessed through various data repositories and platforms.

We have uploaded the dataset to kaggle, a popular platform for data science and machine learning competitions. The dataset can be downloaded from the following link: https://www.kaggle.com/VinayHajare/Fruit-Image-Dataset . Alternatively, the dataset can be also be accessed through GitHub repository, which is free platform using following link: https://www.github.com/VinayHajare/ .

Moreover, we have provided a license for the fruit image dataset that permits users to download, modify, and distribute the dataset for any purpose, as long as the original authors are acknowledged. This license promotes the wide use of the dataset and encourage researchers and developers to use it in projects.

In conclusion, the fruit image dataset created in our study is publicly available and can accessed through various data repositories and platforms, such as kaggle and GitHub. We have also provided a license that permits the free use and distribution of the dataset, prompting its wide adoption in machine learning research and applications.

ADVANTAGES

1. This is a high-quality fruit image dataset that can be used for efficient image recognition and classification.
2. This is capable for automatic detection of fruits and quality inspection. Its capability of identifying fruits from images fits current trends in augmented reality field.
3. Fruit recognition has various practical applications, including automated fruit grading, quality control, and inventory management.
4. The dataset is also diverse in terms of fruit categories, size, shape and color.

CONCLUSION

In conclusion, we have created a fruit image dataset to facilitate of machine learning models for fruits image classification. We collected a total of 1000 images of ten different fruit categories and performed data preprocessing to ensure that the images were of high quality and d appropriate for machine learning applications.

Overall, our fruit image dataset and machine learning model can be used by researchers and developers to develop and evaluate fruit image classification models for a wide range of application.

REFERENCES

https://www.kaggle.com/fruit-360

https://www.youtube.com/