Fruits-360 dataset: new research directions

Mihai Oltean

mihai.oltean@gmail.com
https://github.com/fruits-360

Abstract

Fruits-360 is a database containing images of fruits, vegetables, nuts, and seeds. Here, an improved version of it is introduced. The improvements are focused on the following directions: (1) adding new information about objects, (2) adding new objects with new characteristics (like having diseases and in various stages of growth etc), and (3) enhancing the background removal algorithm. Also, now, the images are stored at their original (captured) size instead of being resized to 100x100 pixels as in the previous versions. New problems and research directions related to this dataset are proposed and discussed.

1 Introduction

Fruits-360 [1] is a dataset with images of fruits, vegetables, and other plants. The first video of a fruit was shot in 2017 and since then, more than one hundred more fruits were filmed and processed. Vegetables, nuts, and seeds were added recently.

Usually, each movie contained about 200-300 frames, which were semi-automatically processed to remove the background. Initially, the images were resized to 100x100 pixels format. The purpose was to establish a uniform standard for easier comparison between trainers. However, this has led to the loss of information (like skin texture and other artifacts). Also, some researchers have requested access to the original size because their training algorithms are better suited for other image sizes than 100x100.

This is why, here, the *Fruits-360 dataset* is improved[1], and several dozen new challenges and problems are proposed and discussed. In a previous paper ([2]) we have discussed only about a single problem: *how to recognize fruits based on their image?* Here, we make several steps forward and we want to know if we can infer more information (about fruits) only by looking at their images. For instance, can we know how much water is inside it? Or fructose? Or, can we

[1]Initial version: 2022.10.1.0. Current version: 2024.02.27.0

make a distinction between fruits, vegetables and nuts only by looking at their pictures? Is the fruit mature enough? All these questions and many others are new problems proposed and discussed in this paper.

Note that the dataset is constantly updated with new images of either existing fruits or of new fruits and the addition of new properties of them.

Currently, there is no training algorithm for solving these new classification problems. For the initial task of recognizing fruits based on 100x100 images, please look at the TensorFlow [6] example described in detail in [2].

The paper is organized as follows: Improvements are listed in section 2. Versioning information is included in section 3. The structure of the archive, folders, and files is described in section 4. Metadata file format is described in section 5. The download addresses for both dataset and metadata are given in section 6. New problems are introduced and discussed in section 7. Future potential developments are discussed in section 8.

2 What is new?

The following aspects are improved here:

- Better image extraction algorithm from videos.
- Better removal of the motor shaft from images.

The following new things are introduced here:

- new problems and research directions related to the dataset.
- Images are now stored at their original captured size. This contrasts with the 100x100 version, where images were stored as 100x100 pixel images.
- images of fruits and vegetables with defects (bugs or diseases)
- images of fruits and vegetables in various stages of growth (not ripen or over-ripen).
- new images of new fruits, vegetables, seeds, etc, and more images of the existing objects.

3 Versioning

The dataset uses a simple format for each version: *Year.Month.Date.Release*.

Usually, *Release* is 0, unless multiple versions are released on the same day.

Please clearly specify the version number when experimenting with the dataset.

Also, apart from the version number, please specify the image size that you work with. This is very important now since the images can have different sizes even for the same fruit.

4 Folders and files

For each fruit, there are multiple files stored in various folders. The folder's name, belonging to a fruit, has the format:

*fruit_name_**,

where * is the index of the fruit. The first index for a fruit is 1. If there are multiple fruits with the same name, they have different indexes (*apple_1*, *apple_2*, *apple_3* etc).

The images for a given fruit have been divided into three subsets: *Training*, *Validation* and *Test*. *Training* contains about 50% of images. *Validation* and *Test* each contains about 25% of images.

Folder *Meta* contains information about the objects in the dataset. This aspect is discussed in section 7 for each particular attribute.

Inside *Training*, *Validation* and *Test*, there is a folder for each object in the dataset. For each object, there are multiple images stored in *jpg* format. The file name starts has the following format:

r?_image_index.jpg,

where ? is a number representing the rotation index (and starts from 0) and *image_index* starts from 0.

Inside *Meta* folder there is also a folder for each object in the dataset. Inside that folder, we have a file named *data.txt*, which contains the attributes of each object.

5 Metadata file structure

The *data.txt* file contains a list of attributes (see section 7 for the meaning of each attribute).

Each attribute is specified as:

Name[*Num_classes*]=*Value*,

where

- *Name* is the name of the attribute,
- *Num_classes* are the number of possible values for that attribute in the entire dataset,
- *Value* is a numerical encoding of the value of that attribute. For instance, if the possible values for an attribute are LOW, MEDIUM and HIGH, then, *Num_classes* is 3 and *Value* can be 0 (for LOW), 1 (for MEDIUM) or 2 (for HIGH).

The first attribute is *VERSION*, which specifies when this file was last updated. This should not be used as a classification problem.

5.1 Example

An example of a metadata file for an apple is given below:

VERSION=2021.09.10.0
CLUSTERS[2]=0
DEFECTS[2]=0
FAT[3]=0
FRUCTOSE[3]=2
GROWTH LOCATION[3]=2
HARDNESS[3]=1
MATURITY[3]=1
NUMBER OF SEEDS[4]=2
PROCESSED[2]=0
PURCHASED[2]=0
SIZE[5]=2
SKIN HARDNESS[3]=0
SKIN ROUGHNESS[3]=0
TYPE[5]=0
YEARLY TREE[2]=1
WATER[3]=2

6 How to download it

The official repository is currently stored at Kaggle:
`https://www.kaggle.com/moltean/fruits`.

The metadata is inside the dataset archive too (see folder *Meta*), but they are maintained with a versioning system at:
`https://github.com/mihaioltean/fruits-360-meta`.

7 Metadata and new problems

So far [2], only one problem has been addressed: recognize the fruit based on its picture. But we can ask more. For instance, can you tell how much water is inside a fruit only by looking at its picture? Or how much fat or fructose? Or if it grows underground or above ground? Or if it has defects or diseases?

Thus, new binary or multi-class classification problems can be devised by adding extra information (also called attributes or metadata) about the objects in the dataset. Another very simple example is to decide the type of object (fruit, vegetable, nut, or seed) based on its image.

A list of attributes is discussed here in detail. Note that some of them are not included in the dataset due to difficulties in quantification or due to big differences caused by the maturity stage.

Most of the data is collected from personal experience (it is easy to find how much water or sugar is inside a fruit) or from various sources like [7, 8, 9, 10, 4, 11] and from several internet websites (a list will be added soon).

For each attribute, the following details are given in this document:

- *Problem type* - can be *Binary* or *Multi-class* (more than 2 classes),
- *Number of classes*,
- *Possible values*,
- *Included in the current dataset?* - *No* or *Yes*. If it is *No*, then is not included in the *data.txt* from *Meta* folder.

When the absolute value for an attribute was not actually measured, a range is provided. For instance, the water content can be *LOW*, *MEDIUM* or *HIGH*. For most of the attributes there are 3 ranges (*LOW*, *MEDIUM* and *HIGH*) or 5 ranges (*LOW*, *low-MEDIUM*, *MEDIUM*, *MEDIUM-HIGH* and *HIGH*). A value of NULL can also be specified in the information is not known.

Attributes in this document are sorted lexicographically except for the first one (*Type*). But, inside the dataset archive, all attributes are listed in lexicographic order.

7.1 Type

Problem type	Multi-class
Number of classes	5
Possible values	FRUIT (0), VEGETABLE (1), NUT (2), SEED (3), OTHER (4)
Included in the current dataset?	Yes

Comments: Is the object fruit, vegetable, nut, seed, or something else? At the moment, the OTHER class should be very rare.

Examples: Apple is a fruit. The potato is a vegetable. Walnut is a nut. Seeds are inside other fruits or vegetables.

7.2 Acidity

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	No

Comments: Acidity can depend on the state of ripeness.

Examples: Lemons have a low pH level. Mushrooms have a high pH level.

7.3 Clusters

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	Yes

Comments: Do, usually, grow in groups?
Examples: Grapes and bananas grow in clusters. Pears do not.

7.4 Crops per year

Problem type	Multi-class
Number of classes	3
Possible values	MULTI-YEAR (0), ONE (1), MORE (2), CONTINUOUSLY (3)
Included in the current dataset?	No

Comments: How many crops per year? *MULTI-YEAR* means that the fruits are not produced yearly or require more than one year to mature (like onions). *MORE* means that multiple distinct crops are produced, CONTINUOUSLY means that the production never stops
Examples:

7.5 Defects / diseases

Problem type	Binary
Number of classes	2
Possible values	ALMOST NO DEFECTS (0), WITH OBVIOUS DEFECTS (1)
Included in the current dataset?	Yes

Comments: Has diseases, bugs, or defects? All 3 aspects have been included in a single attribute.

This is difficult to quantify. Some fruits can have only a small (bug-made) hole, and some others can be pretty messy. Over-ripe in later stages can lead to putrefaction and is considered a defect. This attribute should be expanded into multiple attributes.
Examples:

7.6 Edible

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	No

Comments: The majority of them are edible, so this attribute will not be included in the dataset currently because the problem will be very unbalanced (too much data in one class).

Examples:

7.7 Eaten processed

Problem type	Multi-class
Number of classes	4
Possible values	NOT APPLY (0), NO (1), YES (2), NO AND YES (3)
Included in the current dataset?	No

Comments: How is this fruit usually eaten? Is it usually cooked before eaten? NOT APPLY means that the object is not eaten. "NO AND YES" means that the fruit is eaten equally, either processed or in crude form.

Examples: Potatoes are usually cooked (fried or boiled) before eating. Apples are usually eaten in a crude state (but are also boiled or fried for cookies).

7.8 Fat

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	Yes

Comments: Fat or oil content.

Examples: Apples have LOW fat content. Seeds have HUGE fat content.

7.9 Fertilized

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	No

Comments: Was the soil artificially fertilized? This is important because it can affect the quality of the vegetable and also its size. It is difficult to provide an answer to this question for the purchased objects. Fruits and vegetables from my own garden (see section 7.21) are not artificially fertilized.

Examples:

7.10 Fructose

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	Yes

Comments: Fructose levels.
Examples: Nuts have LOW fructose content. Bananas have HUGE fructose content.

7.11 Glycemic index

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	No

Comments: Glycemic index.
Examples:

7.12 Greenhouses

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	No

Comments: Are the objects grown in a greenhouse? It is difficult to provide an answer to this question for the purchased objects. Fruits and vegetables from my own garden (see section 7.21) are not grown into a greenhouse.
Examples:

7.13 Growth location

Problem type	Multi-class
Number of classes	3
Possible values	UNDERGROUND (0), GROUND LEVEL (1), ABOVE GROUND (2)
Included in the current dataset?	Yes

Comments: Where it grows related to the soil level? If it is a side growing inside a vegetable, the growth location is equal to the growth location of its parent.
Examples: Potatoes grow UNDERGROUND. Watermelon grows at GROUND LEVEL. Tomatoes grow ABOVE GROUND.

7.14 Hardness

Problem type	Multi-class
Number of classes	3
Possible values	SOFT (0), MEDIUM (1) , HARD(2)
Included in the current dataset?	Yes

Comments: How do you feel when you eat them? Shell hardness is not taken into account here.

Examples: Grapes are SOFT. Apples are MEDIUM. Nuts are HARD.

7.15 Maturity stage

Problem type	Multi-class
Number of classes	3
Possible values	UNRIPE (0), RIPEN (1) or OVERRIPE (2).
Included in the current dataset?	Yes

Comments Most of them are in *ripen* stage because, in this state, they are readily available for purchase.

Examples:

7.16 Minerals

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	No

Comments: Minerals content.

Examples:

7.17 Number of seeds

Problem type	Multi-class
Number of classes	4
Possible values	NONE (0), ONE (1), FEW (2), MANY (3)
Included in the current dataset?	Yes

Comments: How many seeds are inside the fruit?

Examples: Seeds have 0 seeds inside. Nuts have 1 seed. Apples have few seeds. Watermelons have many seeds.

7.18 Parent of the seed or nut

Problem type	Multi-class
Number of classes	4
Possible values	NOT APPLY (0), TREE (1), FRUIT (2), VEGETABLE (3)
Included in the current dataset?	No

Comments: It is known that "Nuts are actually the seeds of plants." [4]. So, where the seed or nut is developed? Inside a fruit, vegetable, or directly on a tree? This applies to seeds and nuts only- all others belong to the NOT APPLY category.

Examples: For all nuts, the parent is a tree. Some seeds grow inside/on fruits or inside/on fruits.

7.19 Pesticide

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	No

Comments: Was the fruit or the soil treated with pesticides? This includes herbicides, bactericide, fungicide, etc. It is difficult to provide an answer to this question for the purchased objects. Fruits and vegetables from my own garden (see section 7.21) are not treated with pesticides.

Examples:

7.20 Processed

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	Yes

Comments: Was the fruit processed or not? *YES* can mean boiling, frying, drying, fermentation, refrigeration, pickling, or other types of processing [9]. Fruits treated for fungus, bugs, bacteria are not considered as being processed. Currently, this feature is not included, because not many fruits in this dataset are processed.

Examples: Some cucumbers were fermented, some apples were fried, some figs were dried and some potatoes were boiled.

7.21 Purchased

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	Yes

Comments: Was it purchased or from my own/friends' garden? This is important because, in our garden, we do not use insecticides and most of our fruits have holes (from bugs) and spots or lesions on the skin.

Examples:

7.22 Shell (skin) is edible?

Problem type	Binary
Number of classes	2
Possible values	NO (0), YES (1)
Included in the current dataset?	No

Comments: Is the skin edible or not? This is difficult to specify. In my country, the potato skin is not eaten, even if it is edible and contains many vitamins.

Examples: Coconut shell is not edible.

7.23 Shell (skin) hardness

Problem type	Multi-class
Number of classes	3
Possible values	SOFT (0), MEDIUM (1), HARD (2)
Included in the current dataset?	Yes

Comments: Hardness can also depend on whether the fruit has been processed or not.

Examples: Apples have soft skin. Nuts have a hard shell.

7.24 Shell (skin) roughness

Problem type	Multi-class
Number of classes	3
Possible values	SOFT (0), MEDIUM (1), ROUGH (2)
Included in the current dataset?	Yes

Comments: How the shell feels when touched?

Examples: Apples are SOFT. Cucumbers are MEDIUM. Walnuts are ROUGH.

7.25 Size

Problem type	Multi-class
Number of classes	5
Possible values	SMALL (0), SMALL-MEDIUM (1), MEDIUM (2), MEDIUM-HUGE (3), HUGE (4)
Included in the current dataset?	Yes

Comments: The regular size of the object. This can be very subjective in some cases. For instance, a regular carrot is usually of medium size at most, but, in highly fertilized crops, it can become huge.

Examples: Seeds are small. Nuts are SMALL-MEDIUM. Apples are MEDIUM. Watermelon is HUGE.

7.26 Sugar level

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	No

Comments: Can depend on the level of ripeness. During ripening the starch is transformed to sugar [5].

Examples:

7.27 Vitamins

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	No

Comments: Vitamins content.

Examples:

7.28 Yearly tree

Problem type	Binary
Number of classes	2
Possible values	ANNUALLY (0), MULTI-YEAR (1)
Included in the current dataset?	Yes

Comments: Does the tree having the fruit grows multiyear or starts from seeds every year?

Examples:

7.29 Water

Problem type	Multi-class
Number of classes	3
Possible values	LOW (0), MEDIUM (1), HIGH (2)
Included in the current dataset?	Yes

Comments: Water content.
Examples: Seeds have lower water content. Watermelons have huge water content.

7.30 Weight

Problem type	Multi-class
Number of classes	5
Possible values	SMALL (0), SMALL-MEDIUM (1), MEDIUM (2), MEDIUM-HUGE (3), HUGE (4)
Included in the current dataset?	No

Comments: The weight of the object. This can be different from size because some vegetables can have huge holes inside (for instance melons and zucchini).
Examples:

8 Future development directions

Future work will be focused on:

- adding new fruits and images to the dataset.
- adding new attributes.
- splitting some attributes into multiple attributes. For instance, the DEFECTS should be split into DISEASES, HAS BUGS, HAS FUNGUS, SHAPE DEFECTS, SKIN DEFECTS, GROWTH DEFECTS, etc. This attribute is of great practical importance and deserve a more detailed attention.
- measuring the exact value for some attributes. For instance, the size and weight can be easily measured. pH also. But, this measurement cannot be made for the fruits already filmed in the past.
- adding medical recommendations for fruits. What can you eat if you have some disease and what you cannot eat?
- adding more details for vitamins and minerals content.
- adding details about growth location (country, continent)

- adding details about the temperatures, soil composition and other requirements for optimal growth.
- adding information about production cost and selling price.
- adding world records for size.

References

[1] Oltean M., Fruits-360 dataset, 2017-.

[2] Mureşan H., Oltean M., Fruit recognition from images using deep learning, Acta Universitatis Sapientiae, Informatica, Vol. 10, Issue 1, pp. 26–42 2018.

[3] Mellor C., What's the difference between nuts and seeds? `https://www.woodlandtrust.org.uk/blog/2019/08/difference-between-nuts-and-seeds/`, last accessed on 2021.09.01

[4] Harvard Health Publishing, Quick-start guide to nuts and seeds, `https://www.health.harvard.edu/staying-healthy/quick-start-guide-to-nuts-and-seeds`, last accessed on 2021.09.01

[5] Campbell M., List of Non-Starchy Fruits, SFGate, `https://healthyeating.sfgate.com/list-nonstarchy-fruits-9794.html`, 2008, last accessed on 2021.09.01

[6] Tensorflow website, `https://www.tensorflow.org/`, last accessed on 2021.09.01

[7] Bojnanský V., Fargašová A., Atlas of Seeds and Fruits of Central and East-European Flora: The Carpathian Mountains Region, Springer, 2007

[8] Salunkhe D.K., Kadam S.S., Handbook of Fruit Science and Technology: Production, Composition, Storage, CRC Press, 1995

[9] Barbosa-Cánovas G. V., et al, Handling and Preservation of Fruits and Vegetables by Combined Methods for Rural Areas, FAO Agricultural Services Bulletin 149, 2003.

[10] Prasad S., Kumar U., A Handbook Of Fruit Production, Agrobios, 2010.

[11] Chandra G.R., Engineering for Storage of Fruits and Vegetables: Cold Storage, Controlled Atmosphere Storage, Modified Atmosphere Storage, Academic Press, 2015.