SUFID: Sliced and Unsliced Fruits Images Dataset

Latifah Abdullah Bin Turayki[1(✉)] and Nirase Fathima Abubacker[2]

[1] DCU @ Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
Laaltraiki@pnu.edu.sa
[2] Dublin City University, Dublin, Ireland
nirase.fathimaabubacker@dcu.ie

Abstract. Given the recent surge in online images of fruit, ever more sophisticated models and algorithms are required to organize, index, retrieve, and interact such data. However, the relative immaturity of the processes in current use is holding back development in the field. The current paper explains how images of ten classes of fruit namely apples, bananas, kiwis, lemons, oranges, pears, pineapples, coconuts, mangos and watermelons, presented both sliced and unsliced, were sourced from the Fruits360, FIDS30, and ImageNet datasets to create a single database, "SUFID", containing 7,500 high-res images. Pre-processing was based on using pixel color distribution to determine whether each image was corrupt, or would fit the database. The paper describes the unique opportunities, principally within computer vision, presented by SUFID's hierarchical structure, accuracy, diversity, and scale, all of which can be of use to food researchers. As the dataset and benchmarks are believed to be of benefit to all researchers in the field, they are offered gratis to support future study.

Keywords: Sliced fruits · Unsliced fruits · Multiclass fruits · Balanced fruits dataset

1 Introduction

Digital imagery is used ever more frequently, with digital images of fruit – in particular, sliced fruit – seeing a surge in number, popularity, and availability. Researchers have therefore proposed increasingly sophisticated algorithms and models to facilitate the use of these images, in particular to encourage better organization, indexing, retrieval of data, and user interaction. Given that fruit is normally traded unsliced, few models or algorithms have, to date, been devised to deal with sliced fruit. Moreover, annotating data to the extent required is costly in terms of time and resources. Therefore, although sliced fruit constitutes a core element of food research, few accurate relevant data are available. A technology which enables the automatic recognition of sliced fruit would thus be of significant use to researchers investigating the causes and treatment of obesity as it would allow them to capture dietary intake, and enable self-recording patients to accurately and easily track their consumption. Fruit trading companies would also benefit from the availability of this type of multiclass dataset [1, 2]. Combined with recent progress in deep learning, it is therefore considered that a good dataset will be of benefit to both researchers and businesses.

H. Badioze Zaman et al. (Eds.): IVIC 2019, LNCS 11870, pp. 237–244, 2019.
https://doi.org/10.1007/978-3-030-34032-2_22

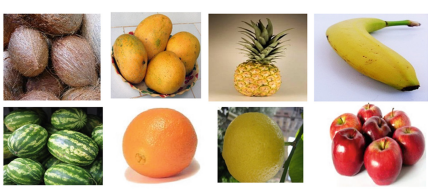

Fig. 1. Sample of the unsliced fruits.

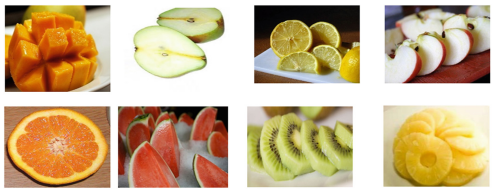

Fig. 2. Sample of the sliced fruits.

The SUFID dataset was designed to both reliable and balanced. It offers images of multiclass fruits (ten varieties of fruit), both sliced and unsliced: bananas, coconuts, lemons, oranges, watermelons, apples, kiwis, pears, mangos and pineapples. This is of obvious benefit to researchers seeking to develop robust image-searching algorithms in which understanding is derived from content, and provides training and benchmarking data against which to test such algorithms. Thus, it will reduce both the time and cost of future research in the area of computer vision. Ultimately, it offers significant potential for the creation of a high-performance deep-learning model.

Recent competition in object recognition has further boosted the development of algorithms. Despite these advances, the fact that datasets tend to be collated with a particular purpose in mind has concentrated research in specific areas, such as sliced or unsliced fruit. The present study aims to progress the automated recognition of fruit by making freely available a comprehensive database of images of common varieties of fruit taken in a range of conditions. Furthermore, it is hoped it will foster the development of more advanced methods of fruit production, for example automated harvesting and yield mapping. In the following sections, the paper sets out the research framework underlying the creation of the SUFID balanced fruits dataset: sourcing the images, challenging the derived dataset, and, finally, creating the data collection and pre-processing protocol.

2 Fruit Images Sources

Fruit-related databases tend to suffer from one or both of two limitations: low numbers, or only containing images of whole (unsliced) fruit. The most commonly used datasets are described below.

Fruits-360[1] [3] a dataset of 100 × 100-sized images of unsliced fruit. Each class is made up of 500–1000 images of a single fruit, pictured from different angles. The Fruits-360 dataset is no has sliced fruits.

Unlike Fruits-360, FIDS30 dataset [4][2] contain images for sliced fruits as well as unsliced. While the database contains 971 images divided into 30 categories, with between 30 to 50 images per category, there is great variety among the images. Some images have white backgrounds, with some having backgrounds that are noisy and

[1] https://www.kaggle.com/moltean/fruits#fruits-360_dataset.zip.

[2] https://www.vicos.si/Downloads/FIDS30.

coloured, and image sizes differ greatly. Some images are of just one fruit, while some contain a variety.

ImageNet [5][3] a large database containing approximately 22,000 different categories described by Deng (2009), each with more than 500 images. ImageNet's size and diversity have made it the 'gold standard' for image recognition purposes; however, due to the wide range of objects covered, there is no focus on food research and so fruit images need to be extracted.

After a database was selected from the three sources, the next stage was pre-processing. Pre-processing is a vital part of the process in preparing images, and can be broken down into several steps, as discussed in more detail in the following sections. The quality of images used in this study varied considerably, as did the degree of pre-processing required.

3 Challenges

Several challenges were encountered when building this dataset. Firstly, it was difficult to find sufficient suitable images as far fewer are available of certain types of sliced and unsliced fruit than of others. Secondly, the process of using algorithms to locate the desired data was lengthy and exacting. Thirdly, images were available in a variety of formats, creating noise, which reduced the degree to which they were useful.

Unified image resolution was therefore used to resize images to a predefined resolution, with the standard 224 × 224 adopted. Table 1 below summarizes the resolutions required for architectures used with transfer learning [6].

Table 1. The Resolution for architectures employed with transfer learning.

Network	Pre-training	Scale
AlexNet	ImageNet	227 × 227
AlexNet	ImageNet	451 × 451
VGG16	ImageNet	224 × 224
VGG19	ImageNet	224 × 224
GoogleNet	ImageNet	224 × 224

It has been proven by multiple studies that augmenting data can enhance deep learning as related to image classification. So, when the first stage of pre-processing was complete, the data were augmented by employing traditional elastic and affine transformations, whereby new images were generated by rotation or reflection of the source, or zooming in on certain parts of it.

[3] http://www.image-net.org.

Pre-processing is essential with a number of steps needing to be taken for image preparation. Images employed in this study will be of various qualities, and some will require more pre-processing than others.

4 Data Preparation

For an image classification model to be considered 'state-of-the-art', it must have a significant amount of labeled, supervised data which, in turn, depends on a significant investment of human time and effort, given that data are scarce and collating those available is an expensive undertaking. The available fruit-related databases tend to offered limited information about the type and condition of the fruit pictured. Searching existing datasets revealed that they tend to privilege images of whole over sliced fruit, and have too few of the latter for any clear outcome to be generated. Thus, for the new database to be properly balanced, it was necessary to obtain data from several sources.

4.1 Dataset Collection Protocol

Data were sourced from three datasets, namely FIDS30, Fruits360, and ImageNet. The multiclass fruits represented were apples, bananas, coconuts, kiwis, lemons, mangos, oranges, pears, pineapples, and watermelons, selected because they were covered by all three chosen datasets, and a total of 7500 usable images were located. First, approximately 40–60 images per class of fruit were taken from Fruits360; thereafter, a further 47–55 per class were taken from FIDS30, and finally, ImageNet was used to ensure a balance of images of sliced and unsliced fruit for each class.

As ImageNet includes multiple classes of different-sized and-colored fruit images, it is suitable for the detection and classification of objects. First, images were acquired from the ImageNet database by running a crawler (python code) to parse the images of the required classes, forwarding the URLs and unique fruit numbers from the source and downloading copies to the destination, as summarized in Table 2. Thereafter, extraction and pre-processing were carried out as described in the next section.

Table 2. Unique fruit numbers in ImageNet

Class	Apple	Banana	Kiwi	Lemon	Orange
ImageNet fruit number	07739125	07753592	07763629	07749582	07747607
Class	Pear	Pineapple	Coconut	Mango	Watermelon
ImageNet fruit number	07767847	07753275	07772935	07764155	07756951

4.2 Dataset Pre-processing

Inputting the data extracted resulted in a tri-colored matrix described as (L) × (W) (3) where: L = number of pixels in X-axis; W = number of pixels in Y-axis; and 3 = number of channels (blue, green, and red). This study adopted the standard size of matrix used for networks trained on ImageNet, which is 224 × 224 × 3.

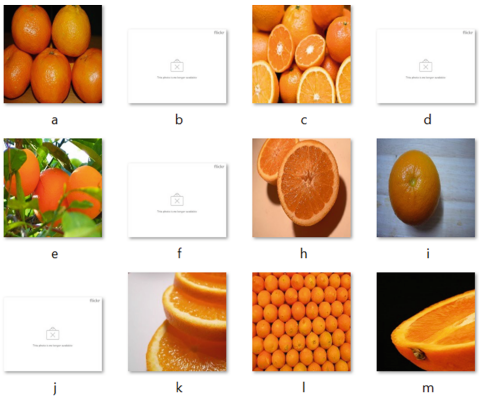

Fig. 3. Sample of images downloaded from ImageNet, (b), (d), (f) and (j) represent invalid images, (a), (c), (e), (g), (h), (i), (k) and (m) represent valid images.

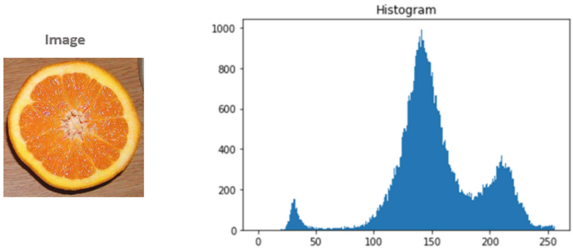

Fig. 4. Pixel colour distribution of a valid image.

Once the images were downloaded, they had to be checked for corruption and to be sure they were valid for the purposes of this study. As can be seen from Fig. 3., it was clear that (b), (d), (f), and (j) were invalid image.

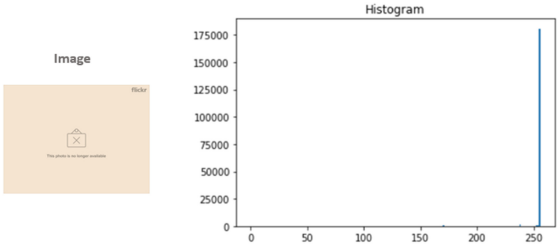

Fig. 5. Pixel colour distribution of a invalid image.

The validity of images was assessed on the basis of distribution of pixel color. Figures 4 and 5 demonstrate how a pixel color distribution plot can determine the fitness or corruption of a given image: if there is too much of a single color within an image, it is generally assumed to be corrupt and therefore of limited use in model training. The filter is designed to evaluate each and every image and then compute the histogram, with the invalidity criterion being a monochrome value of over 0.8. In other words, images which are composed of 80% of more of a single color were excluded as invalid. The filter computes the percentage of monochrome pixels by taking the maximum value of the histogram and dividing it by the size of the image using the formula:

$$percentage\_monochrome = \frac{max\_histogram}{height \times width}$$

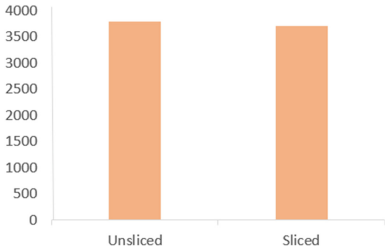

Fig. 6. Number of images of unsliced and sliced fruit.

The next step was to ensure that all imported images had the dimensions required for the extracted database, i.e., 224 × 224 (height x width). After import, filtration, and resizing had been carried out, a balance between images of sliced and unsliced fruit had to be secured to ensure results would be reliable. A thorough manual check revealed that in certain subsets there was an imbalance between the two; therefore, a further search of the source dataset was run in order to extract the necessary number of relevant images. Images of the minority class (sliced or unsliced) were also modified by rotating, cutting, cropping, or zooming as required, and re-inserted into the dataset as an additional balancing step. Figure 6 shows the number of images in the unsliced and sliced fruits.

Once the preparation processes outlined above had been carried out for each and every image in every class of fruit, the dataset was divided into subsets for training, validation, and testing. The first of these, training, comprised 80% (N = 6000) of the total images, with the remaining 20% (N = 1500) assigned to testing. These 1500 images were then withdrawn before the final testing phase.

Despite the laboriousness of the work described above, the set of data ultimately collected is of value as these processes cumulatively balanced out the properties of the images of fruit extracted. This will be of significant help to a range of sections of the research community.

4.3 SUFID Summary

The SUFID balanced fruits dataset comprises a repository of high-quality, 224 × 224-pixel images of ten classes of fruit (apples, bananas, coconuts, kiwis, lemons, mangos, oranges, pears, pineapples, and watermelons). Each class of fruit is represented by 750 images, of which 600 comprise the training set and the remaining 150 comprise the testing set. A proper balance of sliced and unsliced images of each class of fruit has been maintained. Each image presents one example or multiple examples of a single class of fruit. Sample dataset images are reproduced in Figs. 1 and 2.

5 Experiment Environments

This study used Anaconda Distribution, which employs packages from Python, a programming language that can implement deep learning models using Keras and TensorFlow libraries. TensorFlow was created by Google. It is an-open source deep learning framework that uses data flow graphs to perform numerical computations. The notes in the flow graphs signify mathematical operations, while the edges of the graphs signify the tensors, which are multidimensional data arrays [7]. This study also notably used an Nvidia GPU was used for this study's experiments. The GPU requires CUDA libraries to activate TensorFlow.

6 Conclusions

The current paper has presented the SUFID balanced fruits dataset, which is a high-quality database offering balanced numbers of images of sliced and unsliced examples of ten categories of fruits. The objective of this research is to make this dataset freely and widely available to researchers in relevant fields of computer vision, namely fruit classification, recognition and clustering. The images in the database were sourced from three existing datasets ((Fruits360, ImageNet, and FIDS30) and are of apples, bananas, coconuts, kiwis, lemons, mangos, oranges, pears, pineapples, and watermelons. Images are presented of sliced and unsliced examples of each class of fruit, sometimes appearing singly and sometimes with multiple examples in a single image.

The current research makes three contributions to the field. Firstly, it is our hope that, given its careful balance of images of sliced and unsliced fruits, SUFID will become a valuable resource across computer vision studies, particularly within fruit research, as images of sliced fruit have previously been difficult to source. Secondly, SUFID should be of value in training, as the algorithms currently used for object recognition tend to focus on images of unsliced fruit due to their easier availability. Indeed, future researchers could study the possibilities around transferring knowledge of common objects to models learning rare objects. Lastly, SUFID can serve as a more sophisticated and up-to-date benchmark database for future computer vision research, which it is hoped will encourage more research in the specific field of fruit recognition.

Future work will focus on the following points:

i. Develop an effective systems to classify and clustering multiclass fruits images, including both sliced and unsliced fruit using SUFID.
ii. Extending this study to collect other types of fruits to build a dataset that can be used for other research.

Acknowledgment. The financial support by the Deanship of Scientific Research at the Princess Nourah Bint Abdulrahman University. The authors would like to thank anonymous reviewers for their helpful comments. A version of this dataset with multiclass fruits was presented during in the 6th International Visual Informatics Conference 2019 (IVIC'19).

References

1. Bhargava, A., Bansal, A.: Fruits and vegetables quality evaluation using computer vision: a review. J. King Saud Univ. Inf. Sci. (2018)
2. Zhang, Y., Wang, S., Ji, G., Phillips, P.: Fruit classification using computer vision and feedforward neural network. J. Food Eng. **143**, 167–177 (2014)
3. Fruits 360 dataset. https://www.kaggle.com/moltean/fruits#fruits-360_dataset.zip
4. FIDS30 DataSet. http://www.vicos.si/Downloads/FIDS30
5. ImageNet. http://www.image-net.org
6. Ergun, H., Sert, M.: Fusing deep convolutional networks for large scale visual concept classification. In: 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), pp. 210–213 (2016)
7. Hameed, K., Chai, D., Rassau, A.: A comprehensive review of fruit and vegetable classification techniques. Image Vis. Comput. **80**, 24–44 (2018)