Transfer Learning with Pairwise Comparison for Fine-Granular Fruit Shelf-Life Prediction

Jayita Dutta1*, Manasi Patwardhan1, Parijat Deshpande1, Beena Rai1

1TCS Research, Tata Consultancy Services Limited, Pune, India.
jayita.dutta@tcs.com

Abstract
Globally, 30% of the produced foods gets wasted annually across the supply chain. To control this huge wastage and corresponding economic losses, real time prediction of shelf life of foods is essential. This process of reporting shelf life of fruits is often carried out manually using invasive techniques by domain experts, which becomes infeasible when the fruits are getting transported over long distances across the supply chain. To automate this process, we use non-invasive vision-based technique to predict the current age of the fruit from which the shelf-life can be computed. To achieve this we train a model to capture the visual degradation features of fruits. However, such models require large amounts of annotated images for training with fine-granular ‘days-old’ labels. Curating such dataset either by expert annotations or laboratory experiments is expensive. Also, the annotated datasets available for this task are scarce. To address this challenge, in this paper, we avail the accessibility of online time-lapse videos of fruits to auto-synthesise a dataset for the task of pairwise comparison of video frames, whose labels are generated based on their sequence in the video. We transfer-learn the knowledge gained by a model, trained with this data on the comparison task, to improve the performance of the fine-granular classification task where labels are depicting the age of a fruit image. We empirically showcase that with this approach, the performance for the classification task improves by a margin of 16% to 23% percentage in terms of fine-granular classification accuracy for distinct fruits.

Introduction
In an average about 30% of the total produced foods goes waste annually, across the globe. This wastage of perishables such as fruits and vegetables is seen at every node of the food supply chain in between the farm and the fork or even at the consumer end (Grandhi and Appaiah Singh 2016; Grosso and Falasconi 2018) and has a significant negative impact on the world economy as well as on the global environment (Marangon et al. 2014; Koivupuro et al. 2012). To control this huge waste and mitigate the corresponding negative impacts, it is essential that the stake holders of the food supply chain such as farmers, retailers, distributors, consumers etc. can monitor the quality of perishables under any supply chain scenarios and estimate the remaining shelf-life. This would allow them to take decisions dynamically for food repurpose, recycle, reprice or reroute.

Quality of perishables is generally expressed in terms of freshness index or ripeness level (Mitcham, Cantwell, and Kader 1996; Dutta, Deshpande, and Rai 2021). Having a food expert to manually estimate the quality at every step of supply chain is infeasible. In general, quality assessment of perishables is performed by using invasive techniques by measuring variations in chemical composition in terms of starch, sugar, vitamins and minerals (Li et al. 2018; Cadet 2014; Al-Mhanna, Huebner, and Buchholz 2018). However, such testings become infeasible when the produce is getting transported over long distances across supply chain. In such scenarios, quality predictions by non-invasive techniques such as monitoring the emitted gases (Huang et al. 2014; Deisingh, Stone, and Thompson 2004; Simon et al. 1996; Zhang et al. 2012; Hong and Wang 2015) or observing visual degradation in terms of wrinkle formation on skin or growth of black spots, moulds and fungus (Sareen, Chug, and Singh 2022; Surya Prabha and Satheesh Kumar 2015; Pajuelo et al. 2003; Abraham et al. 2011) are preferred. These existing papers have mostly reported coarse grained classification for prediction of good quality of foods. However, in real-life scenarios, more fine-granular classification is essential with ‘days old’ label to take better dynamic decisions.

To predict the remaining shelf life of perishables, its essential to know their current age in terms of ‘number of days old’ starting from the day they are harvested and their expected shelf life under the given supply chain conditions. If these two parameters are known, then remaining shelf life can be calculated by subtracting the current age of the fruit from the expected shelf life. Domain experts have reported the expected shelf life of perishables under distinct environmental conditions (Hassan et al. 2004; Murmu and Mishra 2018; Corradini et al. 2018; Dotto, Vieira, and Pinto 2015; Falah, Nadine, and Suryandono 2015; Parven et al. 2020; Man and Jones 1994; Matar et al. 2018; Chowhan et al. 2016; Li et al. 2019). However, keeping track of the current age of a perishable is extremely challenging when they are getting transported over long distances across the supply chain and in such scenarios prediction of remaining shelf life becomes almost impossible.
Therefore, we develop a non-invasive computer vision based technique to predict the age of the fruits from a given fruit image by modelling it as a fine-granular classification task. We have started with fruits class of perishables as they have a much shorter shelf life and a faster degradation rate, which demands more fine-granular quality annotations than just ‘good’ and ‘bad’. Also, fruits showcase significant visible degradation patterns such as change in color, increased wrinkles on the fruit skin, growing black spots, mould and fungus, which allows the model to capture these features to facilitate the classification task. However, to achieve good performance on such fine-granular task, this vision-based model require large amounts of annotated images. Curating such dataset either by expert annotations or laboratory experiments is expensive and labor-intensive. Also, the annotated datasets available for this task are scarce.

To address this challenge, in this paper we avail the accessibility of online time-lapse videos of fruits to auto-synthesize a dataset for the task of pairwise comparison of video frames, whose labels are generated based on their sequence in a video. We train a model with a shared weight encoder, with this data for the image comparison task. The knowledge gained by this encoder is transfer learned and further fine-tuned to improve the performance of the fine-granular classification task to predict the age of a fruit image in terms of ‘number of days old’ class label.

We use the VGG16 base network (Gulli and Pal 2017; Antonio, Rael, and Buenavides 2021) as our base model which is pre-trained on imagenet and further fine-tune it with ‘Fruits-360’ ( Oltean 2021; Chung and Van Tai 2019) and fruit classification (‘good’ and ‘bad’ classes) dataset to incorporate domain knowledge. In the first experiment, we fine-tune the base model with scarcely available training data with ‘age’ labels. Whereas, in the second experiment we first fine-tune the base model as an encoder in siamese setting (He et al. 2018) for pairwise-comparison task using above mentioned data and further fine-tune for the fine-granular classification task. Experiment 2 performs significantly better on the unseen test dataset for the fine-granular classification task, thus indicating that the pairwise comparison task facilitates the fine-granular fruit age classification task. The fruits used for our experiments are banana, papaya, strawberry and green colored grapes. We observe distinct improvements in the performance of distinct fruits and find it to be a function of the amount of synthetically generated pair-wise comparison data as well as the amount of pre-labelled fine-granular classification data.

**Dataset Collection and Preparation**

We have utilized datasets from three broad sources. For domain adaptability of the base model we use two datasets: (i) ‘Fruits 360’ ( Antonio, Rael, and Buenavides 2021) downloaded from Kaggle \(^2\), which consists of images of 131 classes of fruits with 67692 images for training and 22688 images for validation and (ii) A dataset of images labeled as ‘good’ and ‘bad’ collected for two climacteric fruits (Payasi and Sanwal 2010; Chen et al. 2018) (banana and papaya) and two non-climacteric fruits (Goldschmidt 1997; Paul, Pandey, and Srivastava 2012) (strawberry and green colored grapes). We use web-scraping using similar technique as was reported in (Harimi et al. 2021) for collecting ‘good’ and ‘bad’ fruits dataset. The keywords used for web-scraping are ‘good quality banana’, ‘good banana image’, ‘bad banana image’, ‘bad quality banana’ or ‘banana which has gone bad’, etc. The statistics and label distribution for collected data is depicted in Table 1. We use 80:20 train, validation split of this data for ‘good’ and ‘bad’ classification of fruit.

For synthesizing data for fine-granular classification and pairwise comparison tasks we collect the time-lapse videos from Youtube \(^3\) and other internet sources \(^4\) showcasing sequential and visual fruit patterns under regular environmental conditions.

Out of the available time-lapse videos, very few have ‘age’ (number of days old) labels embedded to the video frames (Statistics in Table 2). To collect data for fine-granular fruit classification task we use these videos. As depicted in Figure 1, the embedded labels are not clearly visible and OCR method (such as Tesseract \(^5\)) for label extraction does not yield good results. Hence, the labels are manually extracted. As different fruits have different perishable timeline and degradation rate, the maximum ‘number of days old’ from the day of harvest label is different for different fruits (depicted in Table 2). Though we use each ‘number of days old’ label as a distinct class for the classification task, for the ease of illustration purpose, in Table 1, we create bins of 5 days to depict the label distribution. We

\(^2\)https://www.kaggle.com/datasets/moltean/fruits

\(^3\)https://www.youtube.com/

\(^4\)https://www.time-lapse-footage.com/

\(^5\)https://github.com/tesseract-ocr/tesseract

Figure 1: Frames of the time-lapse videos with ‘age’ labels capturing the visual degradation patterns of fruits.
Table 1: Label Distribution of data. For fine granular classification each label ranging from 0 to 30 ‘number of days old’ serves as a distinct class, however for illustration purpose the train set statistics are presented in the bins of 5 days. Pairs: Total number of pairs for pairwise comparison including ‘Hard’ and ‘Easy’ pairs. Testset: images in test set.

<table>
<thead>
<tr>
<th>Fruits</th>
<th>'Good'</th>
<th>'Bad'</th>
<th>'0-5'</th>
<th>'5-10'</th>
<th>'10-15'</th>
<th>'15-20'</th>
<th>'20-25'</th>
<th>'25-30'</th>
<th>Pairs</th>
<th>Hard</th>
<th>Easy</th>
<th>Testset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana</td>
<td>95</td>
<td>95</td>
<td>193</td>
<td>246</td>
<td>292</td>
<td>254</td>
<td>142</td>
<td>149</td>
<td>7469</td>
<td>1364</td>
<td>6105</td>
<td>2700</td>
</tr>
<tr>
<td>Papaya</td>
<td>82</td>
<td>82</td>
<td>210</td>
<td>287</td>
<td>215</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>4430</td>
<td>864</td>
<td>3566</td>
<td>1350</td>
</tr>
<tr>
<td>Strawberry</td>
<td>86</td>
<td>86</td>
<td>141</td>
<td>209</td>
<td>81</td>
<td>136</td>
<td>–</td>
<td>–</td>
<td>3428</td>
<td>524</td>
<td>2904</td>
<td>1800</td>
</tr>
<tr>
<td>G. Grapes</td>
<td>55</td>
<td>55</td>
<td>45</td>
<td>53</td>
<td>56</td>
<td>37</td>
<td>32</td>
<td>40</td>
<td>2176</td>
<td>316</td>
<td>1860</td>
<td>2700</td>
</tr>
</tbody>
</table>

Table 2: Training Dataset Statistics. LV: Labeled Videos, LF: Labeled Frames, LDays: number of days labels, VWL: Videos without labels, UF: Unlabelled Frames

<table>
<thead>
<tr>
<th>Fruits</th>
<th>LV</th>
<th>LF</th>
<th>LDays</th>
<th>VWL</th>
<th>UF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banana</td>
<td>4</td>
<td>1276</td>
<td>30</td>
<td>6</td>
<td>4860</td>
</tr>
<tr>
<td>Papaya</td>
<td>2</td>
<td>243</td>
<td>15</td>
<td>4</td>
<td>2867</td>
</tr>
<tr>
<td>Strawberry</td>
<td>4</td>
<td>567</td>
<td>20</td>
<td>6</td>
<td>2064</td>
</tr>
<tr>
<td>G. Grapes</td>
<td>2</td>
<td>263</td>
<td>30</td>
<td>3</td>
<td>1167</td>
</tr>
</tbody>
</table>

We use 80:20 train, validation split for the fine-granular classification, ensuring same label distribution across the splits.

The unlabelled time-laps videos are used to synthesize pairs for pair-wise comparison task (statistics in Table 2). We generate easy and hard pairs. Hard pairs are those pair of fruit frames which do not show significant visual changes as they are picked up from a short time span in the video, whereas the easy pairs are frames with bigger time interval and thus significant visual changes that can be easily be observed. For example, frames with a window size of 1 day does not show significant visual changes and these are referred as hard pairs. These serve as the borderline cases of the decision boundary of our fine-granular classification problem. We use both easy and hard pairs for training for pair-wise comparison task (Table 1). Distinct frame window sizes are used for hard and easy frames for distinct fruits. The average window size for for banana, papaya, strawberry and grapes for easy frames are 100, 20, 15 and 12 and for hard frames are 10, 5, 4 and 4, respectively.

For testing purpose, dataset is generated in our laboratory. 10 samples of each of the above fruits are allowed to perish under normal environmental conditions. The images captured manually with an Android smart phone at three angles (top, left and right), n-times a day (retaining the label distributions of the train-set), over different duration (banana -30, papaya -15, strawberry -20 and green grapes -30 days) in accordance with the 'number of days old' timelines of the training dataset. All the videos and images have varying backgrounds, aspects, and lighting conditions, which simulate real-life conditions and thus eliminate the need for further augmentations. We plan to make this curated dataset available for further research purposes.

**Approach**

The goal is to predict the age of a fruit in terms of 'number of days old' from the time of harvesting, provided an image of the fruit as an input. We perform two experiments.

In the first experiment (Figure 2 (a)), VGG16 network pretrained on ‘imagenet’ dataset is first fine-tuned with the ‘Fruits-360’ dataset and is further fine-tuned on ‘good’ and ‘bad’ fruit classification task, for domain adaptation. We call this trained model ‘VGGFruit’. We add a dense layer (DNNc) on top of ‘VGGFruit’ and further fine-tune the model for the fine-granular classification task using cross entropy loss. This final network is tested on the unseen lab curated test-set to calculate the prediction accuracy.

In the second experiment (Figure 2 (b)), we use ‘VGGFruit’ as a shared weight encoder in Siamese setting. We add a shared weight dense layer (DNNp) and perform Pairwise-Comparison task with the linear comparison layer which considers the order between the compared fruit images in terms of their ‘age’. We use the Bradley Terry as the comparison model with the negative log-likelihood loss as depicted in the equation (Yıldız et al. 2019):

$$L(y(i,j), \hat{y}(i,j)) = \log(1 + e^{-y(i,j)\hat{y}(i,j)})$$

(1)

For pairwise comparison, the tuples are of the form \((i, j, y(i,j))\), where \(y(i,j) = +1\) represents that video frame ‘i’ is older (appears later in the video) than frame ‘j’ and \(y(i,j) = 0\) represents video frame ‘j’ is older than frame ‘i’ in terms of ‘age’ of the fruit. The model generates comparison label \(y^*(i,j) = DNNp(VGGfruit(i)) - DNNp(VGGfruit(j))\) for a pair of fruit images.
### Result and Discussion

For pre-training on ‘Fruits-360’ and the fruit classification (‘good’ and ‘bad’) datasets, 80:20 train, validation split is used. We identify the best model using early stopping criteria with patience value of 5. For domain adaptation, the VGG16 model pretrained on Imagenet dataset is fine-tuned with categorical cross-entropy loss using Stochastic Gradient Descent (SGD) optimizer, 0.9 momentum and 0.0001 learning rate. For Experiment 1, the model is fine-tuned for fine-granular classification task using cross entropy loss. In Experiment 2, the model is trained for pairwise comparison task with Bradley-Terry loss. For both Experiments, SGD optimizer, 0.9 momentum and 0.0001 learning rate are used to train the model.

For both the experiments, the models are tested with the unseen lab curated test-set (Section ‘Dataset Collection and Preparation’). The results are depicted in Table 3. As it can be observed, Experiment 2 performs significantly better than Experiment 1 implying the advantage of transfer learning. The prediction accuracy obtained in case of fine-granular classification task for Experiment 1 (3) is much less due to scarcity of labelled training data. The knowledge embedded in the weights of the ‘VGGfruit’ encoder fine-tuned with the ‘Pairwise-Comparison’ task as the part of Experiment 2 facilitates the fine-granular classification task and thus brings in the improvement in the performance. The improvement is depicted in Table 3 in terms of percentage increase in prediction accuracy. The percentage increase in predicted test accuracy is calculated as (((Experiment 2 accuracy - Experiment 1 accuracy) / Experiment 1 accuracy) * 100). The average percentage increase in the test accuracy for all fruits combined is 18.38%. We perform an ablation where we use VGG16 as the base model with and without domain adaptation by fine-tuning with Fruits 360 and ‘good’, ‘bad’ fruit classification data. It is observed that overall prediction accuracy improves with the domain knowledge (Table 3). Also, the transfer learning form the pairwise comparison task helps more in case of the domain adapted model.

From the Tables 2 and 3, it is observed that the prediction accuracy of Experiment 1 for distinct fruits is proportional to the data available for the fine-granular classification task. Also, the fruit-wise prediction accuracy improvement in case of Experiment 2, is directly proportional to the availability of the pair-wise comparison data per fruit.

We perform error analysis to qualitatively analyze the samples which are mis-classified. We observe that most of the errors in predictions are due to the samples getting mis-classified with the adjacent ‘number of days old’ class labels. We also observe that the improvement in the prediction accuracy in Experiment 2 as compared to the Experiment 1 is mostly due to correction in these mis-classifications. This is specifically because of inclusion of the hard pairs for the pairwise comparison task, facilitating the fine-granular classification to improve the prediction accuracy by reducing the number of samples mis-classified with the adjacent ‘number of days old’ class labels.

From the results reported in Table 3, we also observe that the improvements in test accuracy from Experiment 1 to 2 are more in case of climacteric fruits (banana and papaya) than in case of non-climacteric fruits (strawberry and green grapes). The reason for this is that for the climacteric fruits the degradation changes, which can be observed as they ripen over time are more systematic. On the other hand, the degradation pattern in case of non-climacteric fruits is abrupt and also the visual changes while they degrade are not significant. Thus, the models can better capture the visual degradation pattern shown by climacteric fruits as compared to the non-climacteric fruits, leading to better results.

### Conclusion & Future Work

Our work reports the use of a non-invasive vision-based technique to predict the current age of the fruits in terms of ‘number of days old’ label. We further conclude that the transfer learning with the ‘Pairwise Comparison’ task facilitates the fine-granular ‘age’ classification task having sparsely annotated data, leading to 18.38 % increase in the test accuracy.

To showcase the efficacy of our approach we have started with fruits class as they have a much shorter shelflife, a faster degradation rate and significant visible degradation patterns. However, in future, we plan to use the technique to predict the ‘age’ of other perishable fruits, vegetables, dairy products and different kinds of meat which portrays significant visual degradation pattern under varying temperature and humidity conditions. We also plan to define this task as ordered regression and add other modality inputs such as emitted gases to see if this type of modelling leads to better performance.

### Acknowledgement

We would like to acknowledge our interns Abhijeet Verma and Piyush Gupta for helping in data curation.
References


Grosso, M.; and Falasconi, L. 2018. Addressing food wastage in the framework of the UN Sustainable Development Goals.


