CNN-Based Non-Contact Detection of Food Level in Bottles from RGB Images

Yijun Jiang, Elim Schenck, Spencer Kranz,
Sean Banerjee, and Natasha Kholgade Banerjee ✉

Clarkson University, Potsdam, NY 13699
{jiangy,schencej,kranzs,sbanerje,nbanerje ✉}@clarkson.edu

Abstract. In this paper, we present an approach that detects the level of food in store-bought containers using deep convolutional neural networks (CNNs) trained on RGB images captured using an off-the-shelf camera. Our approach addresses three challenges—the diversity in container geometry, the large variations in shapes and appearances of labels on store-bought containers, and the variability in color of container contents—by augmenting the data used to train the CNNs using printed labels with synthetic textures attached to the training bottles, interchanging the contents of the bottles of the training containers, and randomly altering the intensities of blocks of pixels in the labels and at the bottle borders. Our approach provides an average level detection accuracy of 92.4% using leave-one-out cross-validation on 10 store-bought bottles of varying geometries, label appearances, label shapes, and content colors.

Keywords: food, level detection, deep convolutional neural networks, training set augmentation

1 Introduction

The propagation of ubiquitous technologies in the consumer space has enabled a wide range of applications in kitchen environments to provide user-centric smart assistance [4, 34]. The pervasion of ubiquitous sensing devices and intelligent monitoring of consumer activity has provided a further boost to smart kitchen applications, motivated by the need to provide nutrition awareness [10]. Successful monitoring of user food consumption in kitchens by understanding food levels in containers, recognizing food item counts, and detecting the age of food items has the potential to enhance intelligent kitchens by providing automatic person-centric shopping lists, and recommending user-aware diet choices.

However, existing approaches to detect food quantity are largely contact-based, making propagation of the approaches to average consumer spaces difficult. The approach of Chi et al. [10] requires weight sensors built into a countertop that sense weight change when the object makes contact with the countertop. Approaches that use capacitive sensors [5, 32, 37] only work with liquids, and depend upon full immersion into the liquid, which can induce contamination. Work that detects content-based modulation of vibrational characteristics of objects [12,40] requires installation of sensors on the surface of the container. Non-contact approaches on food use character recognition to detect the expiry date [29], which is rarely visible in frontal viewpoints of containers. They estimate quantities of plated food from top-down cameras [14, 38], which prove infeasible to install in multi-shelf environments, or provide a binary response on presence or absence of a food item [22, 33] as opposed to estimating quantity.

In this paper, we provide the first approach to perform fully non-contact detection of the level of food such as salad dressing in store-bought bottles using deep convolutional neural networks (CNNs) on frontal images of the containers. The input to the CNNs consists of images of bottles with labels of a variety of shapes and appearances, while the output is one of four classes representing four different levels per bottle. We use bottles made of clear glass or plastic to enable visual level detection. While one method of level detection is to count the pixels in an image segment representing food contents, traditional image segmentation algorithms such as k-means clustering [30] or mean shift [11] yield incorrect segment boundaries in the presence of soft edges typical of real-world lighting and camera noise. While deep learning based segmentation algorithms show higher accuracy [9, 26], real-world containers contain highly textured labels in addition to the food contents and may reveal various backgrounds, requiring a two-step process to first segment the image, and then recognize which segment represents food. Instead, our approach takes inspiration from at-a-glance approaches for recognition [7, 31] and identifies food level directly from the image in one step.

Our work addresses three challenges to estimate food level from store-bought containers. First, store-bought containers show diversity in 3D geometry. Second, labels affixed to store-bought containers show a large variability in shape and appearance. Third, the contents of the containers demonstrate a range of color variations. To address these challenges, we augment the training sets used in learning the CNNs by (i) attaching physically printed labels with synthetic textures to the training bottles to provide invariance to label shape and texture, (ii) interchanging the contents of the training bottles to strengthen the invariance of the CNN to food color, and (iii) altering the intensities of images in random blocks in regions of the label and bottle border to prevent overfitting to bottle geometry, label shape, and label appearance. The random intensity alteration is inspired by the work of [41] which reduces overfitting in CNNs by changing pixel values in random rectangles in training images for object and person detection.

We use leave-one-out cross-validation on a set of 10 store-bought bottles with varying geometries and textures, where the training set contains no label, bottle, or food content from the test set. We use patches containing single containers extracted from bottle line-ups typical of real-world shelves and countertops. Our approach provides an average food level detection accuracy of 92.4%.

2 Related Work

Our work falls in the area of intelligent approaches to monitor food use and human behavior in kitchens. One approach of monitoring usage of food items is to use food identity recognition on an image of the item when a user scans the item before a camera after removing it from a storage location. Recognizing food identity requires pre-training of classification systems on a large group of food items. The success of deep learning approaches has motivated a number of approaches to perform food identity recognition with high accuracy. Liu et al. [24] and Hassanejad et al. [13] use CNNs to classify food images for dietary assessment. Kagaya et al. [17] use data expansion [21] to improve classification of food images using deep CNNs. Since pre-trained neural networks may not be tailored to food images, Martinel et al. [25] train deep residual networks and obtain 90.27% accuracy on the Food-101 dataset. Kawano and Yanai [18] combine features obtained from deep CNNs and conventional hand-crafted features. The approach of Sandholm et al. [33] builds food identity recognition into a cloud-based system to monitor food usage from a fridge. These identity recognition approaches require external accounting mechanisms to keep track of food counts, and do not inherently address level detection unlike our work.

To avoid external tracking of counts, several approaches perform holistic 'at-a-glance' estimation of discrete object counts in an image. Regressors [6, 8, 20] and CNNs [3, 27, 28, 39] have been used to perform estimation of counts of people [6, 8, 28, 39] and animals [3, 27]. The work of Chattopadhyay et al. [7] uses CNNs trained on entire images and gridded cells in images to estimate discrete object counts. The work of Laput et al. [22] uses support vector machines (SVMs) trained on features obtained using correlation-based selection on subregions from images in a kitchen environment. The SVMs are used to perform classification of discrete quantities of objects in a sink, presence or absence of a food item, and general clutter on a countertop. Unlike discrete quantity estimation, our task handles estimation of the quantity of continuously varying food items. Approaches exist to use top-down cameras to estimate the volume of plated solid food [14, 38] and the level of solid waste in trash cans [2]. Such approaches are impractical for consumer estimation of food level in containers, since the containers are required to be open, the camera may not be installable directly above the container when the containers are in multi-level shelves, and the approach may yield low accuracy for narrow-mouthed bottles which may occupy few pixels in the image space. In contrast to top-down camera approaches where the contents are unobscured, our task is rendered challenging by the significant obscuration of liquid content induced by the label, and the variation in this obscuration due to differences in label shape and appearance.

There exist a number of contact-based approaches on detecting the quantity of the contents in a container. Several approaches use capacitive sensors immersed in liquids in containers to detect levels based on the differences in dielectric constants of the liquid and the surrounding air [5,32,37]. Such immersive sensors can prove intrusive, potentially unsafe, and impractical to detect content levels in large container line-ups typical of home and store environments. Approaches also exist to measure the differences in vibrational characteristics of containers due to the presence of varying quantities of contents. Zhao et al. [40] induce physical vibrations of waste bins using a DC motor, and measure the

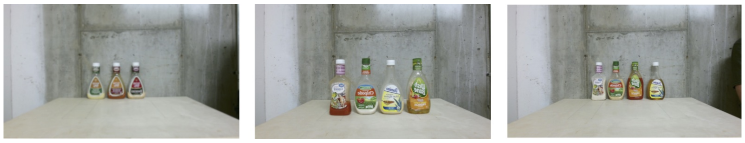


Fig. 1. Original images captured for a variety of bottle line-ups composed from the ten bottles used in this work.

effect of vibration damping due to varying levels of garbage contents. Fan and Khai [12] provide a device that emits a sine wave probe sound using a speaker and classifies the impulse response received by a microphone to estimate food quantity. Both approaches require installation of sensors in contact with the container. Chi et al. [10] use a countertop-installed weight sensor in contact with a container to measure weight changes due to content reduction. Unlike the capacitive, vibrational, and contact-based weight detection approaches discussed here, our work provides fully non-contact level detection using an off-the-shelf RGB camera, improving the portability of our system to consumer environments.

3 Data Collection

We use an off-the-shelf RGB camera of resolution 1920×1080 to capture an image dataset of 10 store-bought bottles with six different geometries. The camera is part of the Kinect sensor that flips images horizontally; however, to avoid overhead of extra operations, we do not perform unflipping. Five of the geometries correspond to bottles with salad dressings, while one corresponds to agave syrup. One geometry represents Ken's Steakhouse dressings, under which, we capture one set of three dressing types—Country French, Honey Mustard and Russian. Another geometry represents Kraft dressings, under which we capture a second set of three dressing types, namely Thousand Island, Honey Mustard and Italian. The remaining four geometries separately represent one bottle of Wish-Bone Caesar Dressing, one bottle of Hidden Valley Farmhouse Originals, one bottle of Southwest Chipotle Salad Dressing, and one bottle of Domino Light Organic Agave Nectar. We pour out varying quantities of liquid and leave 25% of liquid for Level 1, 50% for Level 2, 75% for Level 3, and 100% for Level 4.

To perform the capture, we place groupings of 3 or 4 of the 10 bottles at various levels in rows on a wooden plank against a concrete wall with texture. We perform between 15 to 21 small random real-world translations from left-to-right and from front-to-back to represent minute changes in position that occur when users have repeated interactions with containers on shelves or countertops. We use the camera to capture one RGB image per real-world translation at a resolution of 1920×1080. Figure 1 shows examples of images captured by the HD camera demonstrating the groupings captured in our work.

While the image captured by the camera contains several bottles, our objective in this paper is to use CNNs to perform level detection on a bottle-by-bottle

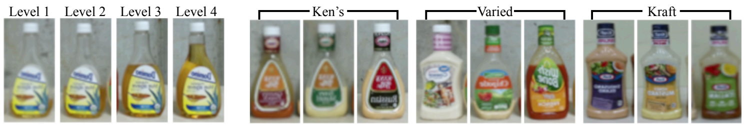


Fig. 2. Left: Four patches representing four different levels for a one bottle. Right: patches for remaining nine bottles. Note the differences in geometry within the 'Varied' group, and with respect to the 'Ken's' and 'Kraft' groups.

basis. We perform a manual extraction of image patches containing individual bottles by specifying a region containing each bottle. While we do not perform automatic bottle extraction in this work, our goal is to make our approach directly pluggable into bottle extraction performed using off-the-shelf object detection algorithms. Since off-the-shelf algorithms may yield bounding boxes that are not perfectly centered around the bottle, we simulate offsets in bounding boxes by sliding the manually specified region left-to-right and top-to-bottom in the image to yield 36 translated patches per bottle instance. To ensure that all patches provided to the CNNs are of the same size, we resize them to a low resolution of 120 × 60, which accelerates training and prevents overfitting. Figure 2 shows examples of patches for four levels of one bottle on the left, and for various levels for the remaining nine bottles on the right.

4 Classification using CNNs

Network Architecture. Figure 3 shows the architecture of the CNNs used in our work. The network is made of three blocks. The first block includes two repeated Conv-BN-ReLU layers, that perform convolution using 32 filters, batch normalization (BN) of the feature maps [16], and activation of the normalized feature maps using the rectified linear unit (ReLU). We compared the accuracies of 3×3, 4×4, and 5×5 filters, and determined that filters of size 4×4 yielded the highest accuracy. The second block consists of another set of two repeated Conv-BN-ReLU layers where the number of 4 × 4 filters is doubled to 64 as recommended in [21], [35], and [15]. The third block consists of two Conv-BN-ReLU layers using 128 filters, the first of which performs 4×4 convolution, and the second of which performs 1×1 convolution. The penultimate layer compresses 128 feature maps using four 1×1 convolutional filters for the four classes. At the output layer, we use global average pooling (GAP) [23] to minimize overfitting by reducing the number of parameters, and we use the softmax function to convert the GAP pooling results into classification probabilities.

Training Data Augmentation. To improve the invariance of our work to differences in bottle geometry, label shape, label appearance, and color of food contents, we use three strategies to augment the data used to train the neural network—expanding the label diversity by attaching new physically printed labels with synthetic texture to the training bottles termed 'Syn', interchanging

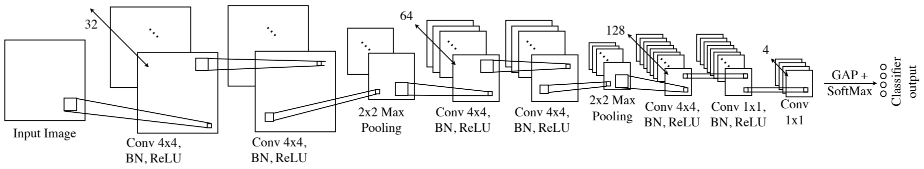


Fig. 3. CNN architecture used in our work. ‘Conv’ represents convolution, ‘BN’ represents batch normalization, and ‘GAP’ represents global average pooling.

the contents in the training bottles termed ‘Int’, and performing random image-based alterations to the training images termed ‘Ran’. For the ‘Int’ approach, we interchange each liquid once to double the size of the training data, which enables the CNNs to avoid overfitting to liquid color. For the ‘Syn’ approach, we design and print 3 sets of labels for each bottle. The synthetic labels are of different shapes and colors, which reduces overfitting of the CNNs to the labels.

For the ‘Ran’ training strategy, we randomly choose half of all the training patches and augment each patch 20-fold by performing two types of transformations at random: domain-based transformations, including horizontal translation up to 3 pixels, horizontal flip, and scaling up to ±0.05, as recommended in [21] and [35], intensity transformations by performing global shifting of each RGB channel up to 30 in intensity values [35], and intensity alterations in random rectangles in half of the patches selected at random as suggested in [41]. Figure 4 shows examples of the training data augmented by the three strategies.

We train five CNNs using various combinations of the three training data augmentation strategies—‘Int+Syn’ that uses interchanging and physically printed labels with synthetic texture, ‘Ran’ that uses random intensity alterations only, ‘Ran+Syn’ that uses random intensity alterations with printed labels, ‘Ran+Int’ that uses random intensity alterations with liquid interchanging, and ‘Ran+ Int+Syn’ that combines all three augmentation strategies. We also train two baseline CNNs for comparison—one based on the original training data without any augmentation strategy termed ‘Orig’, and one with the labels peeled off the bottles termed ‘Bare’ trained with the ‘Ran’ augmentation strategy.

Training and Testing. We generate train and test datasets by performing 1-fold cross validation based on bottles. We train the CNNs using Adam [19] as the adaptive gradient optimizer with cross entropy as the loss function. After each max pooling layer, we include dropout [36] with probability of 25% to prevent overfitting. We choose a batch size of 32 and train for 10 epochs. Our CNN architecture is implemented using the Keras API platform wrapped around the TensorFlow [1] library with GPU support. We perform training and testing using an Asus ESC4000-G3 server containing a single Intel Xeon E5-2660 v3 2.6GHz 10-core processor, 256 GB of RAM, and two NVIDIA GeForce GTX 1080 Ti GPUs. The training takes 1.5 hours per fold, while testing takes 0.24 milliseconds per image. The small level detection runtime per image in comparison to 33.33

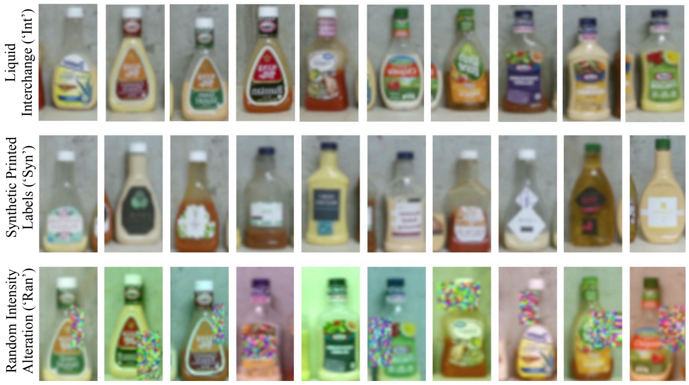

Liquid Interchange ('Int')
Synthetic Printed Labels ('Syn')
Random Intensity Alteration ('Ran')

Fig. 4. Training data augmentation performed in this work. Top row: interchange of liquids (bottles correspond to their original countertops in Figure 2), middle row: attachment of printed labels with synthetic texture, last row: random image-based alterations of intensities in randomly chosen rectangular patches in the images.

milliseconds for 30 fps frame-rate enables our work to be readily deployed into real-time applications.

5 Results

Table 1 shows results of classification accuracies for all the CNNs. While the 'Orig' version receives an average accuracy of 69.9%, the various training augmentation strategies provide improvements in accuracy to 77.1% for 'Int+Syn', 78.9% for 'Ran', 81.7% for 'Ran+Syn', 85.2% for 'Ran+Int', and 92.4% for the combined 'Ran+Int+Syn' strategy. Figure 6 shows examples of the actual level and predicted class probabilities for a variety of bottles using the training approaches 'Ran+Int+Syn', 'Ran+Int', 'Int+Syn', and 'Orig'. As a baseline, the 'Bare' CNNs, where labels are peeled off the bottles in the training and testing set, provide 100% classification when trained with the 'Ran' strategy. Figure 5 shows the overall confusion matrices for CNNs trained without augmentation, and with the five augmentation approaches discussed in this work. Using randomized intensity alterations provides a boost in performance in Level 1, while improvement in classification of Level 2 to Level 4 is obtained using physical interactions of liquid interchanging and printed labels. This may be attributed to the ability of the synthetically printed labels placed in locations of the actual labels to learn the label appearance distribution, and for interchanging to boost invariance to color of the liquid behind the label.

For Bottles 3 and 5, the color similarity of the lower part of the label and the liquid prevents the CNNs from performing correct level prediction, even when trained with the 'Ran' strategy in the case of Bottle 5. The accuracy increases to 100% and 83.6% for Bottles 5 and 3 respectively when we train with 'Ran+Syn', since the diverse array of synthetic labels used in our approach ensures that the

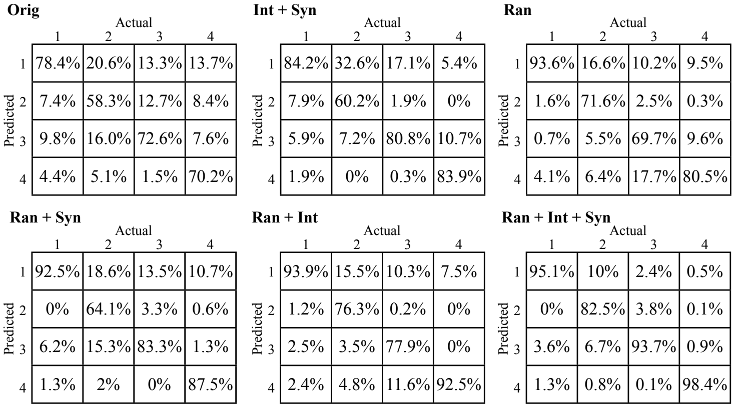
Table 1. Classification accuracy as percentages using CNNs trained with various combinations of augmentation strategies, as compared to CNNs trained with no augmentation ('Orig') and CNNs trained on label-free bottles ('Bare').

ID	1 	2 	3 	4 	5 	6 	7 	8 	9 	10 	Mean
Orig	68.8	50.0	73.8	100.0	54.2	100.0	25.0	91.1	68.1	67.7	69.9
Bare	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Int+Syn	31.9	81.6	75.0	100.0	88.8	100.0	39.6	91.8	81.3	81.2	77.1
Ran	72.9	93.6	82.8	100.0	77.0	100.0	27.1	86.6	80.0	69.2	78.9
Ran+Syn	64.8	66.3	83.6	100.0	100.0	100.0	30.9	99.0	73.1	99.3	81.7
Ran+Int	83.2	88.3	93.3	100.0	98.6	100.0	31.9	99.6	90.1	67.0	85.2
Ran+Int+Syn	79.9	92.8	98.9	100.0	97.3	100.0	60.1	100.0	96.3	98.9	92.4

color similarities are modeled by combinations of synthetic labels and training bottles. For Bottle 2, although Bottles 1,2 and 3 have the same geometry, the label of Bottle 2 shows higher differences in label location and logo appearance compared to Bottles 1 and 3, due to which the 'Orig' strategy shows a low performance on Bottle 2. When trained with the 'Ran' approach, label occlusion improves the accuracy for Bottle 2 to 93.6%. For Bottles 4 and 6, similarity in liquid color and geometry enables all CNNs to predict 100% despite differences in label appearance.

In the combined augmentation strategy, i.e., 'Ran+Syn+Int', we observe a mis-prediction of Level 2 as Levels 1 or 3, and of Level 3 as Levels 3 and 4, due to the proximity of these levels. Our investigation reveals that 97.6% of the Level 3 mis-classifications as 2 or 4 and 56.1% of the Level 2 mis-classifications as Level 1 and 3 are due to Bottle 7, which shows a maximum of 60.1% correct average prediction. A small amount of confusion is also observed between Levels 1 and 3. This is due to the fact that the viscosity of the liquid causes it to stick to the container, inducing the appearance in moderate everyday lighting conditions at lower levels to resemble the appearance at higher levels. In future work, we will investigate the use of scene-specific illumination to resolve optical differences of liquids sticking to container walls with respect to the rest of the contents.

The similarity between agave syrup color on the label and in the contents of Bottle 10 influences level prediction in the 'Orig' strategy due to the closeness of the liquid color to part of the label appearance. The 'Ran+Syn' strategy improves the accuracy to 99.3% since synthetic labels enhance the invariance of the CNNs to the label contents. However, while the performance is likewise improved for Bottle 1 which shows color similarity within the white label writing and the liquid, the accuracy of Bottle 1 reaches a maximum of 83.2% and drops with synthetic label. As future work, we will investigate creating synthetic labels that model color similarities to the bottle liquid.



Orig

Actual

Predicted	1	2	3	4
1	78.4%	20.6%	13.3%	13.7%
2	7.4%	58.3%	12.7%	8.4%
3	9.8%	16.0%	72.6%	7.6%
4	4.4%	5.1%	1.5%	70.2%

Int + Syn

Actual

Predicted	1	2	3	4
1	84.2%	32.6%	17.1%	5.4%
2	7.9%	60.2%	1.9%	0%
3	5.9%	7.2%	80.8%	10.7%
4	1.9%	0%	0.3%	83.9%

Ran

Actual

Predicted	1	2	3	4
1	93.6%	16.6%	10.2%	9.5%
2	1.6%	71.6%	2.5%	0.3%
3	0.7%	5.5%	69.7%	9.6%
4	4.1%	6.4%	17.7%	80.5%

Ran + Syn

Actual

Predicted	1	2	3	4
1	92.5%	18.6%	13.5%	10.7%
2	0%	64.1%	3.3%	0.6%
3	6.2%	15.3%	83.3%	1.3%
4	1.3%	2%	0%	87.5%

Ran + Int

Actual

Predicted	1	2	3	4
1	93.9%	15.5%	10.3%	7.5%
2	1.2%	76.3%	0.2%	0%
3	2.5%	3.5%	77.9%	0%
4	2.4%	4.8%	11.6%	92.5%

Ran + Int + Syn

Actual

Predicted	1	2	3	4
1	95.1%	10%	2.4%	0.5%
2	0%	82.5%	3.8%	0.1%
3	3.6%	6.7%	93.7%	0.9%
4	1.3%	0.8%	0.1%	98.4%

Fig. 5. Confusion matrices for CNNs trained without augmentation ('Orig'), and with various combinations of the three augmentation strategies discussed in this work.

6 Discussion

We have presented an approach in this paper to detect the level of food in store-bought food containers such as salad dressing bottles using convolutional neural networks trained on RGB images. To enable the neural networks to obtain invariance to bottle geometry, label shape, and label texture, we augment the training sets used to train the neural networks using printed labels with synthetic textures and random alteration of intensity blocks on the borders of the bottle, and the interior of the label. Our approach provides an average accuracy of 92.4% using a leave-one-out cross-validation with bottles containing opaque and semi-transparent liquids of several colors. While we have tested our approach with liquids, it can be readily extended to containers with solid contents.

One limitation of our approach is that it requires the optical properties of the container and food to be distinct, thereby preventing level detection in opaque containers. However, a large category of household containers fall within the realm of our approach, including translucent containers such as milk cans, and containers with microscopic perforations in the label that arise due to the process of printing label contents on plastic. While wrap-around labels preclude fine-grained level detection in the region of the label, our method can still be used to detect 'near full' if contents exist in the upper portion of the container above the label, and 'approaching empty' if the region below the label shows depleting contents. Another limitation is that while our approach handles bottles with variations in geometric structure, it requires them to be nearly the same height in order for consistent image sizes as input to the CNNs. In future work, we will investigate image resizing combined with container category detection to perform level percentage detection for containers of varying height.

Since our work performs food level detection on crops of containers from bottle line-ups found in shelves and on countertops, it can be deployed into consumer

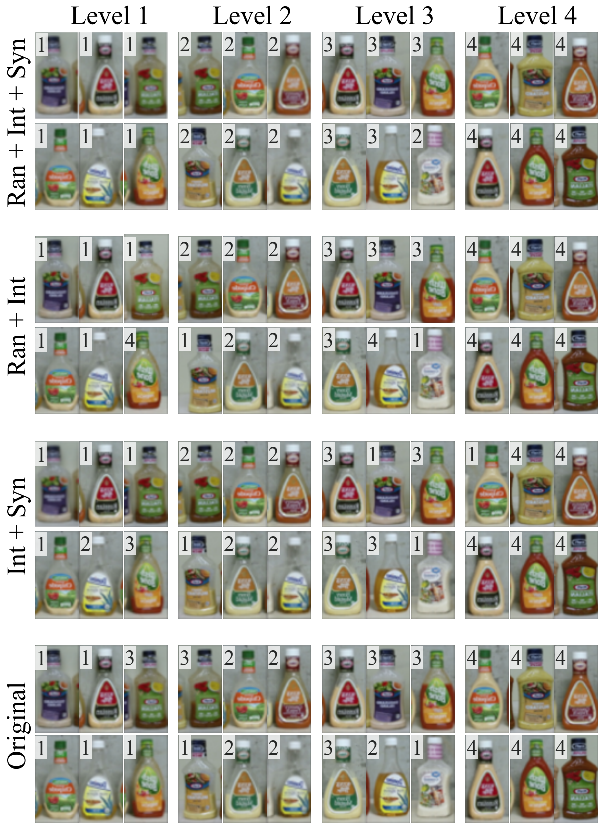


Fig. 6. Results using various training strategies in this work. Each row provides a training strategy, each column of represents the actual level, while the number in each image provides the predicted level.

systems by combining sliding-window bottle detection with food level detection in the sliding window. As part of future work, we are expanding our dataset to contain wider array of container geometries, and solid and liquid food items with a range of opacities and mixture homogeneities, captured under varying illumination. We will also include slight rotations of containers which arise when users interact with them. To eliminate training dependence on physical activities such as attaching printed labels and interchanging liquids, we will investigate virtual approaches to alter liquid color and label appearance in the training set.

Acknowledgements

This work was partially supported by the National Science Foundation (NSF) grant #1730183.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI (2016)
2. Arebey, M., Hannan, M., Begum, R.A., Basri, H.: Solid waste bin level detection using gray level co-occurrence matrix feature extraction approach. Journal of environmental management **104**, 9–18 (2012)
3. Arteta, C., Lempitsky, V., Zisserman, A.: Counting in the wild. In: European conference on computer vision. pp. 483–498. Springer (2016)
4. Bonanni, L., Lee, C.H., Selker, T.: Counterintelligence: Augmented reality kitchen. In: ACM SIGCHI (2005)
5. Canbolat, H.: A novel level measurement technique using three capacitive sensors for liquids. IEEE Trans. on Instrumentation and Measurement (2009)
6. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: IEEE CVPR. pp. 1–7 (2008)
7. Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D., Parikh, D.: Counting everyday objects in everyday scenes. CoRR, abs/1604.03505 **1**(10) (2016)
8. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localised crowd counting. In: BMVC. vol. 1, p. 3 (2012)
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**(4), 834–848 (2018)
10. Chi, P.Y.P., Chen, J.H., Chu, H.H., Lo, J.L.: Enabling calorie-aware cooking in a smart kitchen. In: International Conference on Persuasive Technology (2008)
11. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE TPAMI **24**(5), 603–619 (2002)
12. Fan, M., Truong, K.N.: Soqr: sonically quantifying the content level inside containers. In: ACM UbiComp (2015)
13. Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., Cagnoni, S.: Food image recognition using very deep convolutional networks. In: MADiMa (2016)
14. Hassannejad, H., Matrella, G., Ciampolini, P., Munari, I.D., Mordonini, M., Cagnoni, S.: A new approach to image-based estimation of food volume. Algorithms **10**(2), 66 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR (2016)
16. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
17. Kagaya, H., Aizawa, K., Ogawa, M.: Food detection and recognition using convolutional neural network. In: ACMMM (2014)
18. Kawano, Y., Yanai, K.: Food image recognition with deep convolutional features. In: ACM UbiComp (2014)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Kong, D., Gray, D., Tao, H.: A viewpoint invariant approach for crowd counting. In: IEEE ICPR. vol. 3, pp. 1187–1190 (2006)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)

22. Laput, G., Lasecki, W.S., Wiese, J., Xiao, R., Bigham, J.P., Harrison, C.: Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In: ACM SIGCHI. pp. 1935–1944 (2015)
23. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
24. Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y.: Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In: ICOST (2016)
25. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. arXiv preprint arXiv:1612.06543 (2016)
26. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: IEEE CVPR. pp. 1520–1528 (2015)
27. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences (2018)
28. Onoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: ECCV. pp. 615–629. Springer (2016)
29. Peng, E., Peursum, P., Li, L.: Product barcode and expiry date detection for the visually impaired using a smartphone. In: DICTA (2012)
30. Ray, S., Turi, R.H.: Determination of number of clusters in k-means clustering and application in colour image segmentation. In: Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. pp. 137–143. Calcutta, India (1999)
31. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE CVPR. pp. 779–788 (2016)
32. Reverter, F., Li, X., Meijer, G.C.: Liquid-level measurement system based on a remote grounded capacitive sensor. Sensors and Actuators A: Physical (2007)
33. Sandholm, T., Lee, D., Tegelund, B., Han, S., Shin, B., Kim, B.: Cloudfridge: a testbed for smart fridge interactions. arXiv preprint arXiv:1401.0585 (2014)
34. Sato, A., Watanabe, K., Rekimoto, J.: Mimicook: a cooking assistant system with situated guidance. In: TEI (2014)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. JMLR (2014)
37. Terzic, E., Nagarajah, C., Alamgir, M.: Capacitive sensor-based fluid level measurement in a dynamic environment using neural network. Engineering Applications of Artificial Intelligence (2010)
38. Xu, C., He, Y., Khannan, N., Parra, A., Boushey, C., Delp, E.: Image-based food volume estimation. In: Proceedings of the 5th international workshop on Multimedia for cooking & eating activities. pp. 75–80 (2013)
39. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: IEEE CVPR. pp. 833–841 (2015)
40. Zhao, Y., Yao, S., Li, S., Hu, S., Shao, H., Abdelzaher, T.F.: Vibebin: A vibration-based waste bin level detection system. ACM IMWUT (2017)
41. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)