WHEN WILL BREAKFAST BE READY: TEMPORAL PREDICTION OF FOOD READINESS USING DEEP CONVOLUTIONAL NEURAL NETWORKS ON THERMAL VIDEOS

Yijun Jiang, Miao Luo, Sean Banerjee, Natasha Kholgade Banerjee

Clarkson University, Potsdam, NY
{jiangy, luom, sbanerje, nbanerje}@clarkson.edu

ABSTRACT

In this paper, we perform prediction of food readiness during cooking by using deep convolutional neural networks on thermal video data. Our work treats readiness prediction as ultra-fine recognition of progression in cooking at a per-frame level. We analyze the performance of readiness prediction for eggs, pancakes, and bacon strips using two types of neural networks: a classifier network that bins a frame into one of five classes depending on how far cooking has progressed at that frame, and a regressor network that predicts percentage of cooking time spent at each frame. Our work provides classification accuracies of 98% and higher within one step of the ground truth class using the classifier, and provides an average error of within 20 seconds for the elapsed time predicted using the regressor when compared to ground truth.

Index Terms— deep convolutional neural networks, cook time prediction, fine-grained, activity recognition

1. INTRODUCTION

The rapid spread of consumer capture devices such as RGB, depth, and thermal cameras has increased the potential for the use of these devices in understanding activities such as cooking in everyday environments. There exists a significant body of work in automating food recognition [1–8] from RGB cameras motivated by applications in recommendation of healthy food choices by ubiquitous devices, and robotic automation of cooking activities. However, these approaches focus largely on recognition from static images.

In this work, we perform temporal prediction of food readiness on video data from thermal cameras by using deep convolutional neural networks (CNNs). While there exist approaches to perform understanding of food related activities such as cooking and preparation [9–12] from RGB videos, these approaches either focus on coarse-grained activity recognition where all frames in a single video are given the same label, or fine-grained activity recognition, where a single video comprises multiple unit actions such as 'chopping tomato', 'beating egg', and 'making omelet', and sets of frames corresponding to a single unit action are given the same label. Our work treats temporal readiness prediction as *ultra-fine* activity recognition, where individual frames of a unit action receive different labels, with each label representing the amount of progression in cooking that has occurred at that frame. Additionally, most existing approaches perform recognition of the cooking activity after it is done, while our work allows online prediction of future food readiness during the cooking process.

We use thermal images captured by an FLIR Vue Pro 640 thermal camera in our work to leverage the rise in temperature of the food item over individual frames during cooking for readiness prediction. Thermal cameras are seeing a growing application in non-destructive food quality analysis and assessment of internal temperature of meats for food safety [13–18]. However, while the potential for thermal imaging in food readiness understanding has been discussed [19], our approach is one of the first academic contributions to perform frame-level ultra-fine prediction of food readiness using thermal cameras. We perform per-frame prediction of progression in cooking for actions uninterrupted by human interaction, such as a pancake cooking on one side, an egg fried sunny side up, and a bacon strip crisping in a pan.

We provide two approaches to perform temporal prediction of readiness during cooking of pancakes, eggs, and bacon strips—the components of a traditional American breakfast— using videos from the FLIR Vue Pro 640 thermal camera. Our first approach treats prediction as a classification problem, and uses a deep CNN classifier to provide discrete labels to individual frames of the action. Frames in an early cooking stage are classified as belonging to a lower class and frames at a later stage as belonging to a higher class. We obtain minute-level or half-minute level prediction using this approach. We train five-class classifiers for each food item, and receive an average classification accuracy of 78.21%, 74.99%, and 76.48% against ground truth class labels on datasets of 24 pancakes, 21 eggs, and 48 bacon strips cooked for 120 seconds, 300 seconds, and 300 seconds respectively. When considering classification within one step of the ground truth labels, we achieve accuracy of 98% and above.

Our second approach treats prediction as a regression problem, and uses a deep CNN regressor to provide continuous labels to individual frames. Each label represents the percentage of total cooking time spent at that frame, and provides

978-1-5386-1737-3/18/$31.00

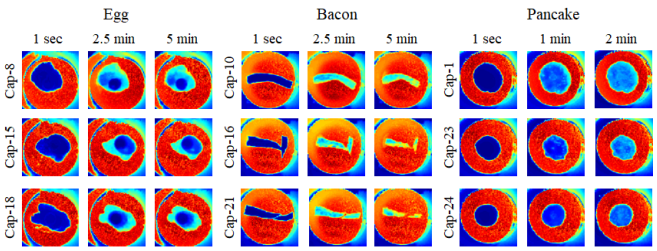

Fig. 1. Three different captures of eggs, bacon, and pancakes at various time intervals using the FLIR Vue Pro 640 thermal camera. As shown in Cap-16 and Cap-21 for bacon, we retain flawed samples as they represent real world shape deviations.

sub-minute or sub-half-minute prediction. We train separate regressors for each food item, and receive an average regression error of 6.12 seconds, 17.4 seconds, and 18.9 seconds for pancakes, eggs, and bacon strips. Using CNNs running on GPU-equipped computers, we achieve per-frame classification or regression times well within real time rate, at 8 milliseconds per frame for 75×75 images. Using either approach, our work has the potential to enable real-time status update to users on time left for cooking completion.

2. RELATED WORK

A large body of work exists on performing recognition of food type from images, where the principal challenge is the variety in appearance for the same category of food. Yang et al. [7] provide an approach that extracts statistics of pairwise local features from soft labels for image pixels. Matsuda et al. [4] provide a two-step method to for multiple-food recognition. The first step of their method detects candidate regions by combining results of several region detectors. In the second step, their method uses a feature-fusion approach to perform recognition on bounding boxes of the candidate regions. Yanai et al. [8] achieve 78.77% and 67.57% for the UEC-FOOD100 and UEC-FOOD256 datasets by applying pre-trained and fine-tuned deep CNNs. They also provide a real-time food recognition system [3] for mobile devices. Kagaya et al. [2] use CNNs to classify ten food categories and find that color features dominate the process of food recognition. Bossard et al. [1] compare random forests and CNNs on 101 food categories. Unlike these approaches that perform food type recognition on static images, our work performs temporal prediction of food readiness using video.

Several approaches exist to use video from color (RGB) cameras to perform fine-grained recognition of food cooking and preparation activities. Rohrbach et al. [9] provide a novel dataset containing 65 scripted cooking activities composited of unit actions such as 'move pan', 'cut onion', 'open egg cup', and 'stir'. They evaluate the performance of articulated pose tracks and holistic video features in detecting the unit actions and in combining them to detect composite activities.

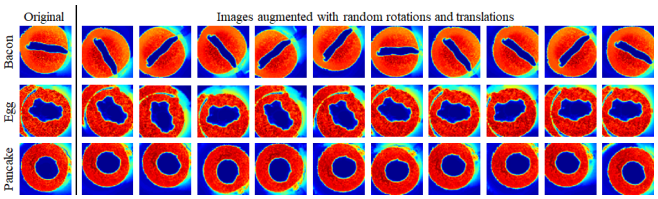

Fig. 2. Original images for bacon, egg, and pancake along with the 10 random rotations and translations used to augment the dataset and prevent the neural network from overfitting.

Ni et al. [10] use an iterative approach to infer the locations of multiple objects involved in fine-grained activities using a long-short term memory (LSTM) network. Lea et al. [11] provide segmental spatiotemporal CNNs for fine-grained recognition of food preparation actions such as making a salad, preparing a sandwich, and making coffee. They contribute an algorithm to perform accelerated segmental inference. Lea et al. [12] provide an improvement to [11] using temporal convolutional networks to capture long-range patterns with a hierarchy of temporal convolutional filters. While these approaches label groups of frames in a single unit action with the same label, our work provides increased discrimination within each unit action, by labeling individual frames within the action as belonging to various stages during cooking using the CNN classifier or as having a certain progression in percentage of cooking time using the CNN regressor. To obtain ultra-fine recognition, our work uses a thermal camera instead of the RGB cameras used in these approaches.

There exists a significant body of work on using thermal imaging to understand food quality; a review of this work may be found in [13] and [14]. Stajnko et al. [15] estimate the number and diameter of apples in an orchard with thermal images of apple trees. Varith et al. [17] detect apple bruises during the warming up process of fruits by using thermal images, and obtain high accuracy due to differences in the thermal conductivity between bruised and sound tissues. Chelladurai et al. [16] use thermal images combined with pair-wise classification models developed by linear and quadratic discriminant analyses to detect fungal infection in stored wheat. Ibarra et al. [18] combine thermal images of chicken with neural networks to predict the internal temperature of chicken during cooking and for 3 minutes after cooking. While their work is the closest to ours, they use a fully connected network where the input nodes represent thermal intensities converted to temperature for a single pixel. The task of elapsed time prediction addressed by our work requires a simultaneous analysis of all pixels in a frame, which we accomplish using CNNs.

3. DATA PREPARATION

Data Collection. We use an FLIR Vue Pro 640 thermal camera to capture thermal images of bacon strips, eggs, and pan-

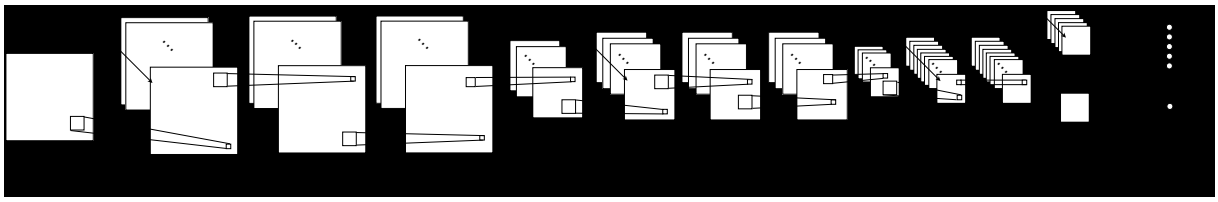

Fig. 3. CNN architecture used in our work. Here, 'Conv' stands for the convolutional layer, 'BN' for batch normalization, and 'GAP' for global average pooling. The last 128-filter convolutional layer branches into either a 5-filter layer for the classifier or a 1-filter layer for the regressor. We create and train separate CNNs for the classifier and the regressor.

cakes at 1 fps cooked on a Faberware 1500W double burner. We record 16-bit images instead of 8-bit white-balanced images to accurately capture low and high thermal intensities. To ensure consistency within each food item, we use prepackaged strips of bacon, cartons of grade A medium sized eggs, and Betty Crocker Bisquick ready pancake mix poured at the same level in 30mL plastic containers. Each sample of a food item is cooked independently of the other. In Figure 1, we show three different captures for egg, bacon and pancake at various time intervals. We do not exclude flawed samples, such as Cap-16 and Cap-21 in Figure 1 for bacon, as they represent the shape deviations found in the real world.

We empirically determine the cooking time for each food item by conducting an initial food readiness experiment. Our food readiness experiment shows that it takes 300 seconds for a single strip of bacon to crisp on one side at a temperature setting of 156°C, 300 seconds for a single egg to cook sunny side up at a temperature setting of 127°C, and 120 seconds for a single 30mL container of pancake mix to cook on one side at a temperature setting of 127°C. Similar to several home-use electrical burners, the Faberware burner automatically turns on and off the heating element to maintain an even temperature. To ensure consistency across captures, we start our capture process immediately after the heating indicator switches off. Our final dataset consists of 48 strips of bacon, 21 eggs, and 24 pancakes, with a total of 14,400 frames for bacon, 6,300 frames for eggs, and 2,880 frames for pancakes.

Training and Test Data Generation. We generate training and testing datasets by performing n-fold cross validation. We create 12 folds for bacon with each fold consisting of 4 random bacon samples, 7 folds for eggs with each fold consisting of 3 random egg samples, and 6 folds for pancakes with each fold consisting of 4 random pancake samples. To remove background pixels, we manually select a region of interest around the burner for the bacon, eggs, and pancake data and automatically crop all images. We resize each image to 75×75 pixels to keep the training tractable and prevent overfitting. Figure 1 shows original thermal images from the sensor, and cropped and resized images.

Labels for CNNs. We provide two types of prediction approaches in this work, one that uses a CNN classifier to categorize frames into 5 stages, and one that use a CNN regressor to predict the progression in cooking. For the CNN classifier, we generate the label of each image by equally dividing the total elapsed cooking time of each capture into 5 classes. For the ultra-fine CNN regressor, the label for the i^{th} frame with frame number f_i is generated as f_i/N, where N is the total number of frames in the capture.

Augmenting of Training Data. To improve the performance of the CNNs and to prevent overfitting, we synthetically augment the training dataset to 10 times the original size by applying random affine transformations as suggested in [20]. We perform random rotations between 0 and 60 degrees and random translations within an offset of ±9 pixels horizontally and vertically to the original images. Example training images after augmenting are shown in Figure 2.

4. CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Network architecture. Figure 3 shows the architecture of the CNNs used in our work. The network consists of two repeated blocks, each of which has three layers that perform convolution, batch normalization [21], and application of the rectified linear unit (ReLU) activation function, followed by a 2×2 max pooling layer. The first block uses a bank of 32 3×3 filters for convolution while the second block uses double the number of 3×3 filters, i.e., 64 filters as recommended in [20], [22], and [23]. The second block is followed by two layers that perform convolution with 128 filters, batch normalization, and ReLU application, with the first layer using 3×3 filters, and the second using 1×1 filters (i.e., a direct weighting of the original input node). The classifier CNN contains a penultimate layer with 5 1×1 convolutional filters for the 5 classes, while the regressor CNN contains a penultimate layer with a single 1×1 filter. At the output layer, instead of using a fully connected layer after the last convolutional layer, we use global average pooling [24] to minimize overfitting by reducing the number of parameters. The classifier CNN uses the softmax function to convert the pooling results into classification probabilities, while the regressor CNN provides the result of pooling as the final output.

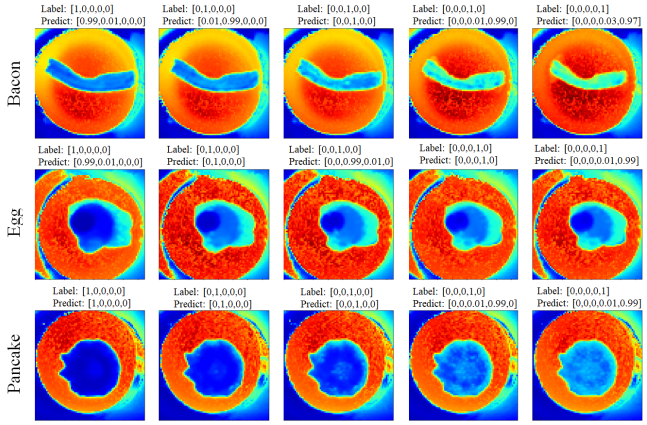

Fig. 4. Ground truth classes and class probabilities predicted using classifier CNN for various frames in bacon, egg, and pancake sequences.

Training. We train both the classifier and the regressor CNNs using back-propagation with Adam [25] as the adaptive gradient optimizer. We choose a batch size of 32 and we train for 100 epochs. After each max pooling layer, we include dropout with probability of 25% to prevent overfitting. For the classifier CNN, we use cross entropy as the loss function, while for the ultra-fine CNN regressor, we optimize the mean square error during back-propagation. We implement the architecture of both CNNs using Keras wrapped around the TensorFlow [26] library with GPU support.

We perform training and testing using four GPU-equipped computers, each containing 32GB of RAM and 500GB of NVMe SSD storage. Two computers have NVIDIA GeForce GTX 1080 Ti graphics processing units (GPUs) and AMD Ryzen 1700X eight-core processor. The remaining two computers have an Intel Core i7-4790K 4GHz four-core processor, with one having an NVIDIA GeForce GTX 980 GPU and the other having an NVIDIA GeForce GTX 980 Ti GPU. For each fold, the training phase takes on average around 2 hours. During the testing phase, each image takes on average 8 milliseconds for classification or regression enabling real-time prediction of progression in cooking.

5. RESULTS

As shown in the second and third columns of Table 1, we evaluate the performance of the classifier CNN using two metrics: top-1 accuracy, and accuracy in a sliding window of 3 elements. We define the sliding window of 3 as follows: if the actual label is i, then prediction labels $i-1$, i, and $i+1$ are considered correct. As shown by the confusion matrices in Figure 6, the intuition of our sliding window of 3 metric is based on the observation that a large number of misclassified samples are located in neighboring classes. For example, for bacon when the actual class is 2 it is primarily confounded

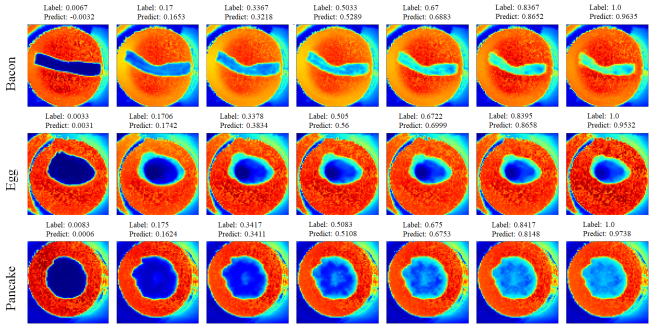

Fig. 5. Ground truth and elapsed time percentages predicted using regressor CNN for various frames in bacon, egg, and pancake sequences.

with class 1 and class 3. Our best top 1 accuracy is achieved for pancake at 78.2%, followed by bacon at 76.5%, and egg at 74.9%. As shown in Table 1, the sliding window of 3 approach improves the top 1 accuracy by 20.82% for pancake, 23.11% for egg, and 22.44% for bacon, to average accuracies of 98% and above. Figure 4 shows ground truth classes and class probabilities for frames from a bacon, an egg, and a pancake sequence. The class with the highest probability is assigned the label for the corresponding input. The supplementary material provides class probabilities for frames from several video sequences used in our work.

The fourth column of Table 1 shows the mean absolute error in using the regressor CNN to perform prediction of the percentage of total cooking time that has elapsed at each frame. In the fifth column, we report mean absolute error in prediction of elapsed cooking time in seconds. For food items with cook times within two minutes such as pancakes, we demonstrate prediction of cooking times within an error of 6.12 seconds on average. For food items with five minutes of cooking time, such as eggs and bacon, we demonstrate prediction of cooking times within an error of less than 20 seconds on average. Figure 5 shows the ground truth and predicted elapsed time as percentages of the total cooking time for frames from a bacon, an egg, and a pancake sequence. The supplementary material provides predicted elapsed time percentages for frames from several video sequences.

Figure 7 shows plots of the predicted time on a per-frame basis for bacon strips (left), eggs (center), and pancakes (right). The horizontal axis in each graph represents the number of frames elapsed (which at 1fps also corresponds to the actual elapsed time). The vertical axis provides the prediction. We show a graph of the ground truth elapsed time in red (at the 45° line), the mean predicted time in black, plots one standard deviation from the mean in dark gray, and plots of the maximum and minimum times in light gray. The mean plot follows the ground truth closely for pancakes, while for eggs and bacon strips, the mean plots show a deviation in the direction of under-predicting the time after 250 seconds. An

Datasets	Classifier Accuracy		Regressor Avg. Error in Elapsed Time	
	Top 1	Sliding Win. 3	as Percentage	in Seconds
Pancake	78.21%	99.03%	5.1%±2.19%	6.12±2.63
Egg	74.89%	98.00%	5.8%±1.18%	17.4±3.53
Bacon	76.48%	98.92%	6.3%±1.84%	18.9±5.53

Table 1. Accuracy in labeling of cooking progression (Sliding Win. 3: accuracy in sliding window of 3 classes).

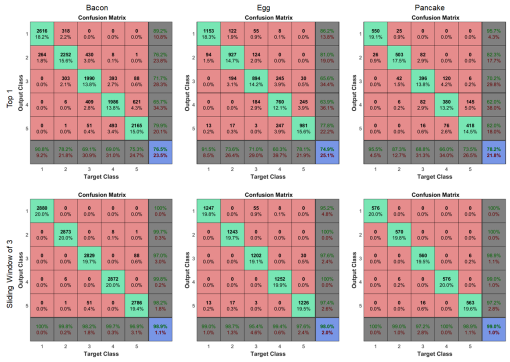

Fig. 6. Confusion matrices for top 1 classification (top row) and sliding window of 3 classification (bottom row) for bacon, eggs and pancakes.

explanation for this phenomenon is that after around 4 minutes of cooking, the bacon strips and eggs do not show a significant change in thermal intensity. For typical consumers, sunny side up eggs fried or bacon strips are considered done within 4 minutes. While the maximum plots for bacon and pancake, and the minimum plot for bacon show a significant deviation from the mean, our data indicates that these are outliers. For bacon strips, eggs, and pancakes respectively, 72.1%, 69.5%, and 73.1% of the predicted results lie within one standard deviation of the mean, and 94.9%, 96.1%, and 94.8% of the predicted results lie within two standard deviations of the mean.

6. DISCUSSION

We have presented two approaches in this paper to perform prediction of food readiness using thermal video data by performing ultra-fine recognition on a per-frame basis. Our first approach treats prediction as a classification of frames into five stages, and uses CNN classifiers to provide minute-level prediction of status for eggs and bacons, and nearly half-minute level predictions for pancakes. To achieve sub-minute prediction, our second approach provides CNN-based regressors to predict percentage of cooking time elapsed at each frame. The classifier approach provides a classification accuracy of 76.5% on average against the ground truth class, and accurately predicts the correct class within one step of the ground truth upwards of 98% of the time. The regressor approach provides average errors within 20 seconds.

One of the main limitations of our approach is that it is not sensitive to the deviation in underlying burner behavior. Figure 8 shows examples of results with low and high accuracies, where images with high deviation from ground truth show high deflection in burner intensities from expected values. The effect is most pronounced for the first two bacon images in Figure 8(b). For the pancakes, the deviation in burner temperature induces them to be overcooked or undercooked (first and second pancakes in Figure 8(a)), showing that the predicted results do reflect real-world behavior. As part of future work, we are interested in developing approaches that explicitly model the burner behavior. Another reason for error is deviation in food shape from the average. For instance, the first bacon strip in Figure 8(a) is torn, and the first two eggs in Figures 8(a) and 8(b) have the yolk offset to the edge instead of being in the center. In future work, we will investigate techniques for spatially aware prediction of food readiness.

One area of future work is to leverage the strengths of the classifier in providing high accuracy within 1 step of the ground truth label to improve the frame-level prediction of the regressor by having the regressor depend on the results of the classifier, leading to a merged architecture for regression. While we use image sizes of 75×75 to keep the training tractable, we are interested in performing empirical evaluation of change in classification and regression accuracy and performance by altering the image sizes between ultra low resolution to very high resolution, with proportional synthetic augmenting of the data. In future work, we will provide an expanded dataset to include foods with a variety of textures, viscosities, and cooking times. We will also investigate the effect of differences in taste preferences for food readiness, and impact of image noise due to real-world effects such as fumes or vapor from cooking, food splatter on the camera, and built-in sensor noise.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*, 2014.

[2] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *ACMMM*, 2014.

[3] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, 2015.

[4] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *IEEE ICME*, 2012.

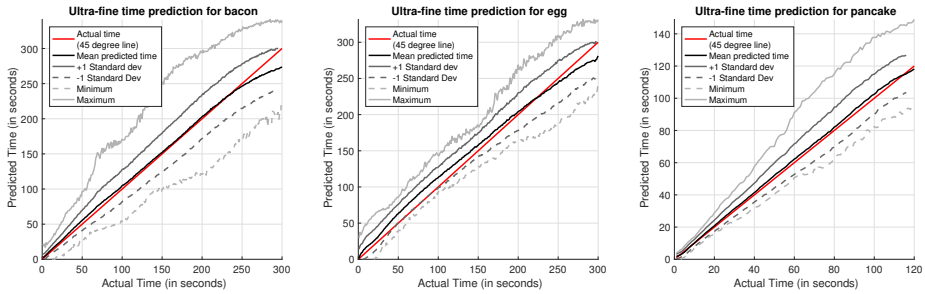

Fig. 7. Average predicted cooking time completed at each frame (black) plotted against actual time (red) for bacon strips, eggs, and pancakes. Plots at ± 1 standard deviation, and at the maximum and minimum values of predicted time are also shown.

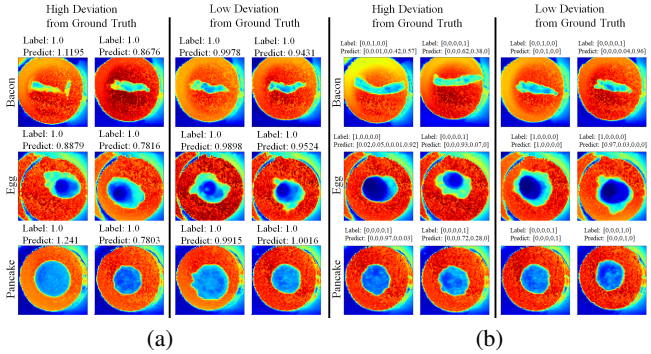

Fig. 8. Results with high and low deviations from ground truth for (a) elapsed time percentages from the regressor CNNs and (b) class probabilities from the classifier CNNs.

[5] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung, "Automatic Chinese food identification and quantity estimation," in *SIGGRAPH Asia 2012 Technical Briefs*, 2012.

[6] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *IEEE ICIP*, 2011.

[7] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *IEEE CVPR*, 2010.

[8] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *IEEE ICME Workshops*, 2015.

[9] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *IEEE CVPR*, 2012.

[10] B. Ni, X. Yang, and S. Gao, "Progressively parsing interactional objects for fine grained action detection," in *IEEE CVPR*, 2016.

[11] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *ECCV*, 2016.

[12] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," *CoRR*.

[13] A. A. Gowen, B. K. Tiwari, P. J. Cullen, K. McDonnell, and C. P. O'Donnell, "Applications of thermal imaging in food quality and safety assessment," *Trends in food science & technology*, 2010.

[14] M. Teena and A. Manickavasagan, "Thermal infrared imaging," in *Imaging with Electromagnetic Spectrum*. 2014.

[15] D. Stajnko, M. Lakota, and M. Hočevar, "Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging," *Computers and Electronics in Agriculture*, 2004.

[16] V. Chelladurai, D. S. Jayas, and N. D. G. White, "Thermal imaging for detecting fungal infection in stored wheat," *Journal of Stored Products Research*, 2010.

[17] J. Varith, G. M. Hyde, A. L. Baritelle, J. K. Fellman, and T. Sattabongkot, "Non-contact bruise detection in apples by thermal imaging," *Innovative Food Science & Emerging Technologies*, 2003.

[18] J. G. Ibarra, Y. Tao, and H. Xin, "Combined ir imaging-neural network method for the estimation of internal temperature in cooked chicken meat," *Optical Engineering*, 2000.

[19] "Thermal imaging cameras in the food industry," http://www.flir.co.uk/cs/display/?id=41781.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[24] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "TensorFlow: A system for large-scale machine learning.," in *OSDI*, 2016.