Brightness Adaptive Food Recognition Using CNN

Duan-Yu Chen and Sheng-Chieh Chang

***Abstract*-- With the advance of technology, people's quality of life is getting better and better, but the number of death due to illness gradually increased. At the top ten causes of death, diabetes and kidney diseases account for the fifth and ninth respectively. The two diseases in the following treatment are strict in the diet control. At present, although there are many wearable devices able to calculate how much calories consumed, the wearing device estimating the calories and sodium content of food for the user is not yet popular since food recognition is still challenging due to the variety of their color, shape, and texture. Among them, brightness is one of the most critical problems. In view of this, we employ the brightness to calculate the reflective region, and adjust its hue and saturation without changing the texture. Hence, the region can be consistent with other non-reflective region of color consistency. In the food recognition, we use convolution neural network to extract the simple to complex features, and convert it to fisher vector through fisher kernel. Eventually, we utilize support vector machine for classification.**

I. INTRODUCTION

Food recognition is still a challenging task even the deep learning-based approaches have been proposed. Bossard *et al.* [1] introduced a novel method to mine discriminative parts using random forests (RF) and used a new, publicly dataset available dataset for real-world food recognition with 101'000 images, they coined this dataset Food-101, as it contained 101 categories, the result had an average accuracy 50.76%. They also made use of AlexNet [2] to achieve top-1 classification accuracy of 56.40%. Meyers *et al.* [3] applied Visual Geometry Group (VGG) and got the top-1 classification accuracy of 79%. Liu *et al.* [4] utilized two methods which were Grey World method and Histogram equalization on inception model based CNN approach to two real-world food image data sets (UEC-256 and Food-101) and achieved impressive results. For reflective region on food in images, we can replace the reflective region by inpainting. In [34], the original image is decomposed into two components, one of which is processed by inpainting and the other by texture synthesis. The output image was the sum of the two processed components. This approach remained limited to the removal of small image gaps. Criminisi *et al.* [5] introduced this method by exemplar-based synthesis suffices to remove the object. Dong *et al.* [6] had presented a technique for filling image regions based on a texture-segmentation step and a tensor-voting algorithm for the smooth linking of structures across holes. However, these approaches might change the texture of food images when the reflective region is too large. Hence regional brightness adjustment is better than inpainting and entire image brightness adjustment. Chondro and Ruan [7] proposed a method to solve the overexposed problem, and thus we will introduce this method as our preprocessing.

In summary, we integrate the brightness preprocessing and food recognition in Chapter II. Chapter III will present the experiment results. Some concluding remarks will be shown in Chapter IV.

II. PROPOSED METHOD

Brightness adjustment in [8] is used and then our inpainting approach is applied for image preprocessing. The result is demonstrated in Fig. 1. The training framework is shown in Fig.2 and its operation of convolution and maxpooling are shown in Fig.3. The convolutional layers contain their respective weights, and rectified linear units (ReLU) is introduced in layers (convolution 1-4) as activity function to solve the nonlinear distribution of feature map. In addition, we applied fisher vector to represent these features in each image. In general, the image is described using Gaussian mixture model (GMM), and its descriptors are global, hence fisher vector is employed as global features.

Fig. 1 Original image(left); brightness adjustment (right)

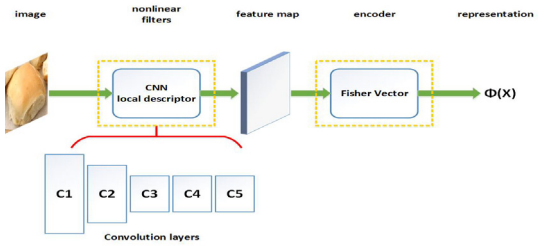

Fig. 2 Proposed training framework

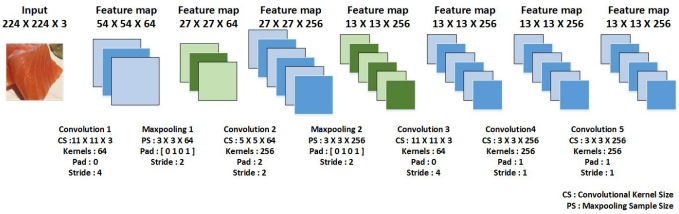

Fig. 3 the architecture of VGG model (imagenet-f)

III. Experiment Results

The dataset consists of 48 categories totally, and contains about 200-300 images in each category; we evaluate the performance and compare our method against others. The proposed system we designed is running on the hardware environment of Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz, 8GB memory and the developing environment of Matlab R2014a under Window 7.

We evaluate the system performance with about 15000 food images, which contain red pepper, yellow pepper, herring, salmon, waffle, chickpeas, etc. shown in **Fig**.4.

Fig. 4 Demonstration of food training samples

We select the top-x accuracy to evaluate the performance of our method and compare the input images with the brightness adjustment scheme against original input image. In addition, in terms of nutrients such as calories, fat, and cholesterol in the vegetable accounted for only a small proportion or not even at all, wherefore we divide all kinds of food into rough categories, such as vegetable, fish, meat, bread, etc. The results are presented in **Table** 1.

Table 1 Performance of w/ and w/o brightness adjustment

categories	input	Top-1 accuracy	Top-3 accuracy	Top-5 accuracy
Rough	Original	94.31%	99.30%	99.72%
	Adjusted	**96.94%**	**99.67%**	**99.85%**
Detail	Original	91.07%	98.43%	99.15%
	Adjusted	**94.44%**	**98.94%**	**99.43%**

For performance evaluation, CIFAR [9] and bilinear CNN (BCNN) [10] are selected for comparison. DCNN consists of 3 convolutional layers, 2 pooling layers, 1 fully connected layer and softmax layer. The final features of fully connected layer are 64 dimensions. In addition, two CNN models, which can be the same, combine BCNN or not. Each image will obtain two CNN features, and then combine them into bilinear vector of which dimensions will increase. For example, if the kernels of model A and B are 200 and 100 respectively, the dimension of bilinear vector is 20,000. **Table** 2 shows the evaluated performance. Although the top-3 and top-5 accuracy at rough and detail categories of FVCNN slightly lose BCNN, the top-1 accuracy wins BCNN up to 2.42%.

Table 2 The evaluated performance of different methods

categories	methods	Top-1 accuracy	Top-3 accuracy	Top-5 accuracy
Rough	DCNN [52]	96.29%	99.61%	99.54%
	BCNN [53]	96.12%	**99.73%**	**99.94%**
	FVCNN	**96.94%**	99.67%	99.85%
Detail	DCNN [52]	91.43%	98.79%	99.54%
	BCNN [53]	92.02%	**99.11%**	**99.74%**
	FVCNN	**94.44%**	98.94%	99.43%

IV. Conclusion

In this work, an adaptive brightness adjustment has been used for local region preprocessing. The performance gain using the adjusted images in terms of precision is about 3%. In addition, a novel CNN framework has been used to extract image features. Meanwhile, we have used GMM to encode the features into Fisher vectors, which is effective to represent the distribution of the texture features in each image. Our method outperforms the state-of-the-art works about 2.5% for top-1 recognition in terms of precision.

Reference

[1] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 mining discriminative components with random forests," *in Proc. European Conference on Computer Vision*, pp. 446–461, Sept. 2014.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton."Imagenet classification with deep convolutional neural networks," *in Proc. Advances in Neural Information Processing System*, pp. 1106–1114, Dec. 2012.

[3] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an Automated Mobile Vision Food Diary," *in Proc. IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp. 1233-1241, Dec. 2015.

[4] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma. "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," *in Proc. International Conference on Smart Homes and Health Telematics*, pp 37–48. *Springer*, Jan. 2017.

[5] A. Criminisi, P. Perez and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *in Proc. IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200-1212, Sept. 2004.

[6] B. Dong, H. Ji, J. Li, Z. Shen, and Y. Xu, "Wavelet frame based blind image inpainting," *in Proc. Applied and Computational Harmonic Analysis,* vol. 32, no. 2 pp. 268–279, May 2011.

[7] P. Chondro and S. J. Ruan, "Perceptually Hue-Oriented Power-Saving Scheme with Overexposure Corrector for AMOLED Displays," in *Journal of Display Technology*, vol. 12, no. 8, pp. 791-800, Aug. 2016.

[8] M. Cimpoi, S. Maji and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3828-3836, June 2015.

[9] A. Krizhevsky, and G. Hinton, "Convolutional deep belief networks on cifar-10." *Unpublished manuscript* (2010).

[10] T. Y. Lin, A. RoyChowdhury, and S. Maji. "Bilinear CNN models for fine-grained visual recognition." *in Proc. the IEEE International Conference on Computer Vision*, Dec. 2015.