Hand-held Food Localization and Food Recognition Using Convolutional Neural Network

Duan-Yu Chen and Hao-Syuan Wang

***Abstract*-- In modern society, calories and carbohydrate intake leads to the obesities and diabetes sharply increases. For this reason, food recognition and its application attracted more and more attention. However, a variety of problem such as deformation and color difference cause the difficulty in this task. Especially, localization problem of food item is the most difficult, because the background always colorful and messy. In view of this, optical flow algorithm, which commonly used for foreground separation, is employed in this paper. Based on the speed information, hand-held objects can be isolated from background according to the estimated optical flows. Then, gradient and RGB color value of each pixel in an image are used for recognition. With the advantage of convolutional neural network, high stability and high tolerance, we finally get the remarkable precision in the experiment results, which show the feasibility of our proposed approach for real-world environments.**

I. INTRODUCTION

According to the statistics, in Taiwan, 5% to 10% of the population suffers from diabetes. The considerable increase of patients per year leads to huge medical expenses. Furthermore, diabetes and other complications like kidney failure, cardiovascular diseases and hypertension, cause higher mortality rate than cancer. For this reason, everyone needs to monitor calorie and carbohydrate intake [1] for prevention of diabetes. In addition, using computer vision technology to achieve personal diet assistant has received much attention. Moreover, as the rapid development of mobile devices and networks, many methods are proposed to recognize and estimate the volume of food intake, which implemented by a mobile device such as cell phones [2]. However, food recognition is a difficult task since the appearance of food items are deformable in real world, even within the same category, not to mention the environmental change such as light, pose and resolution. Bossard *et al.* [3] introducing random forest on dataset of 101 food categories, with 101000 images in total in this dataset, they also found that convolutional neural networks (CNN) [4] performs well in this classification task. Therefore, CNN has become increasingly popular to this study [5-6]. The most distinctive characteristic is that better image features for recognition are automatically extracted via training, which is now a state-of-the-art technique for image recognition challenges such as the Large Scale Visual Recognition Challenge (LSVRC). Furthermore, many improved methods had been proposed to enhance the accuracy of CNN classifier, such as mean activity subtracting and data augmentation using principal components analysis (PCA) [7] had been introduced in Alex-Net [8]. However, food localization is still a serious problem before food recognition since the above methods recognize the foods which are placed on the table. To this end, we consider the objects that we holding in hands are candidates of food, not just the objects that placed on the table. Such properties not only solve the problem of food localization but also meet the situation in real world. Besides that, considering the distance between mouth and hands also helps in behavior analysis such as eating, drinking and smoking. For example, Hsieh *et al.* [9] detects cigarette by color histogram ratio and analysis the related event by the distance between mouth and hands. However, it is difficult for object segmentation using raw pixels. Optical flow [10] algorithm recently used to separate foreground from background also perform well on hand-held object localization, only requires general webcam to capture RGB image. Finally, CNN algorithm is used to build classifier and classify the food that holds in hands.

The remainder of this paper is organized as follows. In Section II we introduce the method of our proposed framework apply on face detection, hand-held object localization and food recognition. Then, detail of our experiment will be mention in Section III. Finally, conclusion is drawn in Section IV.

II. PROPOSED METHOD

In this section, we describe the detail of the proposed prototypes system. Figure 1 shows the flowchart of the system. In the first step, Haar-like features are used for face candidate detection and mouth detection. Then, optical flow between adjacent frames is used for object localization. Finally, recognition task is implemented via two cascaded stages CNN classifiers, which input are contour image and color image.

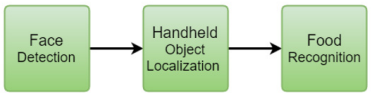

Fig. 1. Flowchart of the proposed system

In general, eating or drinking is completed by hands in front of the chest region. With such behavior, we set the region of interest around chest, convenient for hand-held object localization. We call this region as body region, which is under face region and proportional to the size of the face. Figure 2 shows the illustration. In this body region, we adopt optical flow to detect moving object. Furthermore, evaluating the magnitude of the speed often performs well on foreground separation, especially the moving objects between adjacent frames. The distance from hand-held object to mouth always related with eating, drinking and smoking. With such a correlation, we only recognize the hand-held object when

distance between mouth and hands is decreasing, and noise can easily be excluded and save a lot of computing costs.

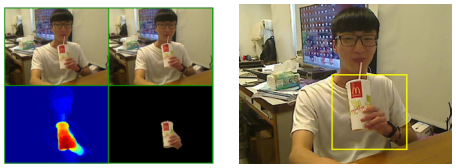

Fig. 2. Foreground segmentation using optical flow and Object region cropped around the object centroid

Then we apply the magnitude of optical flow O_{energy} as weights to enhance the foreground region and weaken the background region. The input of stage 1 CNN classifier $Img_{contour}$ can be calculate by

$$Img_{contour} = G_{obj} \times O_{energy}, \quad (1)$$

where G_{obj} represents the edge image of object region getting from gradient operator. The result of foreground enhancement is shown in Fig. 3. We can observe that almost all the background is removed by this method.

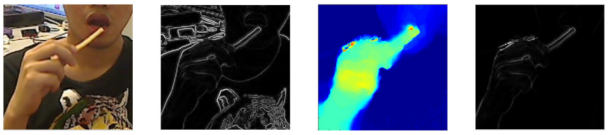

Fig. 3. Foreground enhancement using optical flow

As shown in Fig. 4, the architecture of our CNN classifier-1, contains 5 layers with weights, the first 4 layers are convolutional layers and the last layer is fully-connected layer. The output of the last fully-connected layer is fed to a 4-way softmax which produces a distribution over 4 class labels. For this stage, we rapidly verify the existence of hand-held object, so we construct this classifier with edge map input. we build our classifier 2 with the same architecture. The only difference is the input with pixels of object region since the RGB image obtains more pixels information than gray level image.

Fig. 4. An illustration of the architecture of our CNN classifier

III. Experiment Results

To evaluate the performance of our proposed approach, we evaluate the performance of our proposed approach on real-life videos and compare our proposed against two baseline methods. Our experiment focus on 3 specific food items acquired from fast food – hamburger, drink and french-fries. The proposed method was implemented in Matlab (R2014a) on a personal computer, which is equipped with Intel 4720HQ 2.6GHz processor and 8 GB memory. We evaluate our proposed with 10 videos, which contains motions of eating hamburger or french-fries and drinking soda. In this experiment, all the test datasets were captured in laboratory by using a webcam (Logitech C920, resolution: 640×480). Fig. 5 shows the performance evaluation in terms of ROC curve. It is clear that our proposed method outperforms other methods.

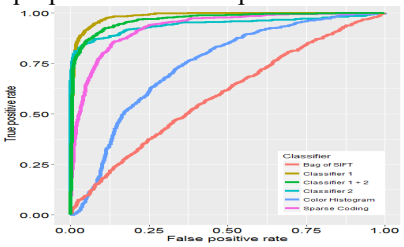

Fig. 5. ROC curve of 3 hand-held food categories

IV. Conclusion

In this paper, we present a real-time and robust diet monitoring system, especially for fast foods. The proposed method firstly locates the hand-held object by optical flow between adjacent frames, since the higher magnitude of the optical flow always represent moving object such as hands. Then, two cascaded stages of CNN classifiers are employed to recognize the hand-held object. Even a variety of poses and occlusions, we still achieve 92.63% average precision rate among four categories in only 0.25 second, which includes hamburger category, drink category, french-fries category and not food category.

Reference

[1] F. Zhu , A. Mariappan , D. Kerr , C. Boushey , K. Lutes , D. Ebert and E. Delp, "Technology-assisted dietary assessment," *In Proc. of IS&T/SPIE Conference on Computational Imaging VI,vol. 6814,* 2008.

[2] A. Mariappan, M. Bosch, F. Zhu, C. Boushey, D. Kerr, D. Ebert,, "Personal dietary assessment using mobile devices," *In Proc. of the IS&T/SPIE Conference on Computational Imaging VII, vol. 7246,* pp. 72460Z(1-12), 2009.

[3] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 – Mining Discriminative Components with Random Forests," *in Proc. of European Conference on Computer Vision,* 2014.

[4] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *in Proc. of IEEE, vol. 86, no. 11,* pp. 2278-2324, 1998.

[5] H. Kagaya, K. Aizawa, and M. Ogawa, "Food Detection and Recognition Using Convolutional Neural Network," *in Proc. of ACM International Conference Multimedia,* pp. 1085-1088, 2014.

[6] Y. Kawano and K. Yanai, "Food Image Recognition with Deep Convolutional Features," *in Proc. of ACM UbiComp Workshop on Smart Technology for Cooking and Eating Activities,* 2014.

[7] H. Jegou, O. Chum, "Negative evidences and co- occurrences in image retrieval: the benefit of PCA and whitening," In Proc. of European Conference on Computer Vision, 2012.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. of Advances in Neural Information Processing Systems, 2012.

[9] J. W. Hsieh, J. C. Cheng, L. C. Chen, C. H. Chuang, D. Y. Chen, "Handheld object detection and its related event analysis using ratio histogram and mixture of HMMs," *Journal of Visual Communication and Image Representation, vol. 25, no. 6,* pp. 1399-1415, 2014.

[10] C. Liu, "Beyond Pixels: Exploring New Representations and Applications for Motion Analysis," *PhD thesis, MIT,* 2009.