Global Journal of Computer Science and Technology: F
Graphics & Vision
Volume 18 Issue 1 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

# Recognition and Classification of Fast Food Images

By Amatul Bushra Akhi, Farzana Akter, Tania Khatun
& Mohammad Shorif Uddin
*Jahangirnagar University*

*Abstract-* Image processing is widely used for food recognition. A lot of different algorithms regarding food identification and classification has been proposed in recent research works. In this paper, we have use an easy and one of the most powerful machine learning technique from the field of deep learning to recognize and classify different categories of fast food images. We have used a pre trained Convolutional Neural Network (CNN) as a feature extractor to train an image category classifier. CNN's can learn rich feature representations which often perform much better than other handcrafted features such as histogram of oriented gradients (HOG), Local binary patterns (LBP), or speeded up robust features (SURF). A multiclass linear Support Vector Machine (SVM) classifier trained with extracted CNN features is used to classify fast food images to ten different classes. After working on two different benchmark databases, we got the success rate of 99.5% which is higher than the accuracy achieved using bag of features (BoF) and SURF.

*GJCST-F Classification:* *I.4.0*


*Strictly as per the compliance and regulations of:*

# Recognition and Classification of Fast Food Images

Amatul Bushra Akhi [α], Farzana Akter [σ], Tania Khatun [ρ] & Mohammad Shorif Uddin [ω]

*Abstract-* Image processing is widely used for food recognition. A lot of different algorithms regarding food identification and classification has been proposed in recent research works. In this paper, we have use an easy and one of the most powerful machine learning technique from the field of deep learning to recognize and classify different categories of fast food images. We have used a pre trained Convolutional Neural Network (CNN) as a feature extractor to train an image category classifier. CNN's can learn rich feature representations which often perform much better than other handcrafted features such as histogram of oriented gradients (HOG), Local binary patterns (LBP), or speeded up robust features (SURF). A multiclass linear Support Vector Machine (SVM) classifier trained with extracted CNN features is used to classify fast food images to ten different classes. After working on two different benchmark databases, we got the success rate of 99.5% which is higher than the accuracy achieved using bag of features (BoF) and SURF.

## I. Introduction

Automatic food identification and calorie estimation become an important issue in last few years because of the negative impact of obesity in our health. Obesity may cause cardiovascular diseases, diabetes mellitus type 2, obstructive sleep apnea, cancer, osteoarthritis, asthma, etc. [1] Researchers said that junk foods and processed foods are responsible for increasing the childhood obesity[2]. Eating extra calories can harm the healthy production and functioning of the synapses of our brain. Fried chicken, pizza, burger, etc. are favorite fast food for both child and adults. People often buy these high-calorie foods to control their appetite especially when they are busy and unable to take their meal in time. Today's people are more conscious about their health issues and try to maintain a healthy diet. Due to the availability of smart phone and computer-aided object recognition techniques become more popular for dietary assessment. Although the identification of food and estimation of its calorie is a very challenging task but many effective steps already have taken in this regards. We also propose an easy but more effective calorie measurement technique that helps people to identify the amount of junk food and snacks they can intake as well as to decide whether the food is harmful or not good for their health. We use both PFID datasets and our own data sets and apply deep neural network with SVM classifier. Deep learning neural networks have multi-layer structure which can easily extract complicated features from input images and supervised learning classifier SVM can efficiently perform a non-linear classification [3]. Our experimental result shows the better performance of CNN with a higher accuracy rate.

*Author α σ ρ ω: Department of Computer Science and Engineering, Jahangirnagar University, Savar, Bangladesh.*
*e-mails: akhi2010.cse@gmail.com, farzanajoti @gmail.com, shondha90@gmail.com, shorifuddin@gmail.com*

## II. Literature Review

Obesity is conceding a great problem in today's life. The preeminent reason of obesity is consuming more calories than we burn which can seriously undermine the quality of life. Researchers says, accurately assessing dietary intake is an important factor to reduce this risk. To meet this exigency, researches have taken some approaches to measure the calorie of a food. In 2009, an extensive food image and video dataset was built named the Pittsburgh Fast-Food Image Dataset (PFID), containing 4545 still images of 101 different food items, such as "chicken nuggets" and "cheese pizza" etc. [1]. The researcher applied Support Vector Machine (SVM) classifier on this dataset and achieved a classification accuracy of 11% with the color histogram method and 24% with the bag-of-SIFT (Scale-Invariant Feature Transform)-features method [2]. Chen et al. (2012) focused on this major issue and proposed a method of food identification and quantity estimation for dietary assessment. They use Gabor and color features to represent food items. A multi-label SVM classifier combined with multi-class Adaboost algorithm is used to show that the new technique can successfully improve the performance of original SIFT and LBP feature descriptors. Around 50 categories (100 sample images of each) of food such as soup, dumplings etc. are used and achieved 68.3% accuracy [3]. Probst et al. (2015) is motivated to introduce another prototype for dietary assessment with the help of smart phone as well as the features of image processing and pattern recognition. Scale invariant feature transformation (SIFT), local binary patterns (LBP), color etc. common visual features are used for espying food images. The bag-of-words (BoW) model is used to perceive the images taken by the phone [4].

Deep learning gradually becomes a very powerful image recognition technique, and CNN is the most popular deep learning architecture. In 2015, Yanai et al. applied deep convolutional neural network (DCNN) technique on ImageNet dataset and achieved accuracy

78.77% for UEC-FOOD100 and 67.57% for UEC-FOOD256 dataset [5]. Kagaya et al., applied CNN on their own dataset for the identification and recognition of the food item. CNN provide higher accuracy than traditional support-vector-machine-based methods where the accuracy rate for recognition was 73.70% and for detection was 93.80% [6]. In 2016, Hassannejad et al. proposed a deep convolutional neural network (DCNN) technique having a depth of 54 layers on UEC FOOD 100, ETH Food-101 and UEC FOOD 256 dataset and achieved 88.28%, 76.17% and 81.45% as top-1 accuracy and 97.27%, 96.88% and 92.58% as top-5 accuracy for dietary assessment [7].Christodoulidis et al. applied a 6-layer deep convolutional neural network on their own dataset containing 573 food items to classify food and the accuracy rate was 84.9% [8]. In 2016, Singla et al., proposed a new method of identifying food/non-food items and recognizing food category successfully using a GoogLeNet model based on deep convolutional neural network. According to their experimental results they achieved a high accuracy rate of 99.2% in food/non-food item classification and 83.6% in food item recognition [9]. Liu et al. [10], propose a new Convolutional Neural Network (CNN)-based food image recognition algorithm and applied it on UEC-256 and Food-101 data sets and achieved 87.2% and 94.8% accuracy respectively. In [11], a five-layer CNN with bag-of-features (BoF) and support vector machine was applied on a dataset containing 5822 images of ten categories and the overall accuracy of 56%. After that researcher applied Data expansion techniques to increase the size of training images for which the accuracy was increased by 90%.

Due to the complexity of food images, many of the previously-proposed methods for food recognition achieved low classification accuracy. In our proposed system we used two training data sets one is publicly available PFID data set another is manually created by us with images captured by smart phone or camera. We use Support Vector Machine (SVM) classifier with a trained CNN to extract and to classify fast food images of ten different classes and achieved accuracy 99.5

## III. Dataset

Two benchmark datasets such as Pittsburg Fast-food Image Dataset (PFID) and Food-101 Dataset images are used in this paper to evaluate the accuracy of food recognition. The PFID collection is proposed by Chen et al. is used to measure the accuracy of recognition algorithms consists of 4,545 still images is divided into 101 categories of standard computer vision approach. This dataset of foods each of which is categorized into three instances. For each categories of foods both images and videos are captured in both restaurant conditions and a controlled lab setting. Each instance of each food has four still images in restaurant environment, six still images in the laboratory setting. In Food-101, a challenging data set of 101 food categories, with 101000 real world images in total are introduced. It includes very diverse but also visually and semantically similar food classes where each class consists of 1000 of image among which 250 are manually reviewed test images and 750 are training images.

## IV. Methodology

At the very beginning of our experimental method, it is very important to do several preprocessing to make the images ready for work properly. Fig 1 shows the complete methodology of our proposed system.

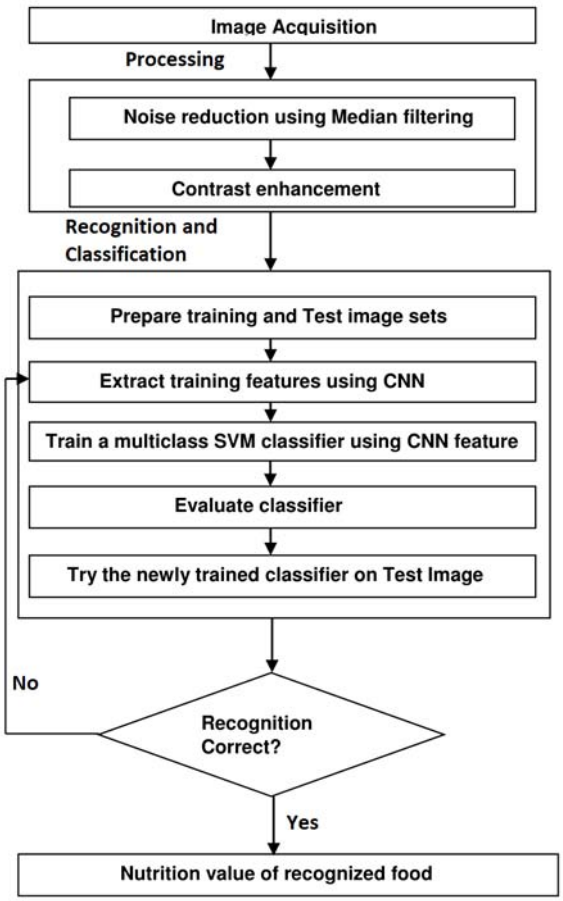


*Fig.1:* Block diagram of the proposed methodology

*a) Preprocessing*

A raw image contains of certain factors such as noise, climatic conditions, poor resolution and unwanted background for which it is not suitable enough to classification and identification. So it is important to improve image quality and prepare the image for further processing to detect the object as accurately as possible. In this paper the pre-processing process consists of noise reduction and contrast enhancement.

i. *Noise reduction using median filter*

The contamination of digital image by salt-and-pepper noise is largely caused by error in image acquisition. Thus, noise reduction is essential for the accuracy of further processing. In salt-and-pepper noise a certain percentage of individual pixels in digital image are randomly digitized into two extreme intensities. To remove this kind of noise effectively we use a non-linear median filter which can remove salt and pepper noise without significantly reducing the sharpness of an image.

*Fig. 2:* Example from Food 101 dataset. Left: Image with 'Salt & Pepper' noise Right: Image after reducing noise with median filter

ii. *Contrast enhancement*

Image contrast is an important factor which is used to evaluate image quality in addition to distinguish one object from another as well as background. In image processing, contrast enhancement is used to improve the appearance of an image for human visual analysis or subsequent machine analysis. It is created by the difference in luminance reflected from two adjacent surfaces as well as the difference in the color and brightness of the object. In this paper, to contrast the test image we use histogram equalization technique.

*Fig. 3:* Contrast enhancement using histogram equalization

*b) Recognition and classification*

This section describes the way to recognize the test image classify the image category by using well established deep learning approach called Convolutional Neural Network (CNN). Object recognition using deep learning is one of the most successful object classification techniques and our target is to classify a given image into one of the pre-determined training objects.

i. *Prepare Training and Test image sets*

We have split up the entire dataset into two subsets namely the training set and validation or testing dataset. 30% images were randomly selected for training dataset and the remainder 70% images for test datasets. Our data set is contrived by ten different types of fast food such as chicken wings, chocolate cake, ice-cream, French fries, pizza, hamburger etc. To perform this experiment, we use 1000 images for each categories of food. The training set contains 750 images and testing set contains 250 images for each of the food category. We have trained our classifier engine by using a pertrained CNN as a feature extractor. Some sample images from training dataset are given below:


*Fig. 4:* Sample images from training data set

Some sample images from our test image set has been given below:

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

*Fig. 5:* Sample images from test data set

ii. *Extract training features using a pretrained CNN*

Convolution neural networks (CNN), a widely used deep learning tools are inspired from the biological structure of a visual cortex. Along with input and output layer CNN consists of multiple hidden layers, such as convolutional layers followed by max-pooling layers, and fully-connected layers. The architecture of a convolution neural networks can vary depending on the types and numbers of layers included.

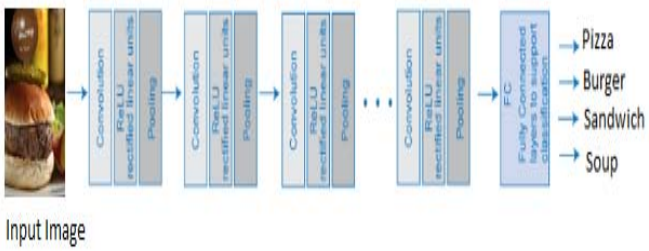


*Fig. 6:* CNN

In each layer, convolution networks are arranged in 3-D manner to produce a 3-D output from 3-D input. In the first convolutional layer, the neurons are connected to the regions of 3-D input images to transform them into a 3-D output. The hidden units (neurons) in each layer learn nonlinear combinations of the original inputs, which is called feature extraction. The focus is to capture basic image features such as edges and blobs from the beginning layer of the network to extract the feature of the image. These basic image features are preprocessed by deeper network layers to form higher level image features which are better suited for recognition tasks as they are the combination of all the primitive features into a richer image representation. In this work, "ResNet-50" has been used as a pretrained network model loaded using resnet50 function from Neural Network Toolbox in Matlab, trained on the ImageNet dataset, which has 1000 object categories and 1.2 million training images.

We have manipulate this pretrained CNN by changing the initial learning rate lower than the default and the maximum number of stages to 20 for preventing it from over fitting our data. The following figure represent the performance of this fine-tuned network on our data:

| Epoch | Iteration | Time Elapsed (seconds) | Mini-batch Loss | Mini-batch Accuracy | Base Learning Rate |
|---|---|---|---|---|---|
| 3 | 50 | 145.73 | 0.2541 | 91.41% | 0.000100 |
| 5 | 100 | 289.74 | 0.1266 | 98.44% | 0.000100 |
| 7 | 150 | 433.96 | 0.0398 | 100.00% | 0.000100 |
| 9 | 200 | 578.37 | 0.0274 | 100.00% | 0.000100 |
| 11 | 250 | 722.87 | 0.0586 | 99.22% | 0.000100 |
| 14 | 300 | 867.10 | 0.0300 | 100.00% | 0.000100 |
| 16 | 350 | 1011.27 | 0.0304 | 100.00% | 0.000100 |
| 18 | 400 | 1155.47 | 0.0170 | 100.00% | 0.000100 |
| 20 | 450 | 1299.65 | 0.0081 | 100.00% | 0.000100 |

*Fig. 7:* Performance of network on few Epochs

We have resized all the images because net can only work on RGB images which are 224-by-224-by-3. Each and every layer of a CNN produces a response to input images but only few of them are capable for image feature extraction. At the very beginning the layer extract specially the blobs and edges features. Fig 5 displays the network filter weights from the first convolutional network. It gives an illustration about the reason behind well performance of the features extracted using CNN.

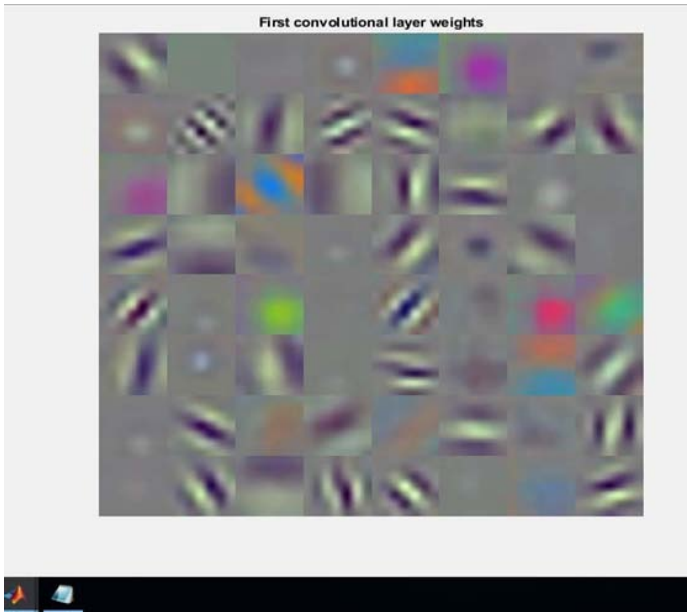


*Fig. 8:* First convolutional layer weights

The deeper network layer then further process these unrefined features extracted by the first layer and create a richer imager feature representation. These higher level features are more suitable for a recognition task than the first one [15].The easiest way to extract deeper layer features using the activations method in matlab.

iii. *Train Multi class SVM classifier using CNN features*

In this step extracted CNN features are used to train a multiclass SVM classifier. At the very beginning SVM were designed for binary classification which separates the binary classes (k = 2) with a maximized margin criterion [16]. But real life problems sometimes require the classification for more than two categories. These type of problems can be solved by the construction of multiclass SVMs, where we create a two-class classifier over a feature vector $\emptyset(\vec{x},y)$ obtained from the pair consisting of the input features and the class of the data. During the test, the classifier chooses the class,

$$y=\text{argmax}_{y'} \vec{w}^{\top} \emptyset(\vec{x},y') \quad (1)$$

The margin during training is the gap between this value for the correct class and for the nearest other class, and so the quadratic program formulation will require that,

$$\forall_i \forall_y \neq y_i \ \vec{w}^{\top} \emptyset(\vec{x}_i, y_i) - \vec{w}^{\top} \emptyset(\vec{x}_i, y) \geq 1 - \vec{\zeta}_i \quad (2)$$

This general method can be extended to give a multiclass formulation of various kinds of linear classifiers [17].In this work a fast Stochastic Gradient Descent solver is used for training by setting the fitcecoc function's 'Learners' parameter to 'Linear' because this algorithm is specially suitable when training data size is huge. This helps speed-up the training when working with high-dimensional CNN feature vectors [18]. When training deep learning models, the objective function is considered as a sum of a finite number of functions:

$$f(x) = \frac{1}{n} + \sum_{i=1}^{n} f_i \text{ (x)} \quad (3)$$

Where *fi*(**x**) is a loss function depending on the training data instance indexed by *i*. It is important to highlight that the per-iteration computational cost in gradient descent scales linearly with the training data set size *n*. Hence, when *n* is huge, the per-iteration computational cost of gradient descent is very high. [19].

iv. *Evaluate the classifier*

To evaluate the trained classifier, first of all we extract the CNN features from the images of our test set. These test features are then passed to the classifier to calculate the accuracy of the trained classifier.

## V. Experimental Result

Our proposed system creates a classifier depending on the extracted features of CNN for identification of the object. The obtained success rate of recognition and classification has been represented using a confusion matrix. A confusion matrix also called an error matrix is a contingency table that comprise of the information about actual and predicted classifications done by a classification system. Fig 10 and Fig 11 shows the confusion matrix that appraises the Accuracy rate of the classification using our algorithm.

Convolution Neural Network Confusion Matrix over BAR Image Set (Accuracy 99.13%)

| KNOWN | Burger | Chicken_ Breasts | Chicken_ Nuggets | Chicken_salad | Cake | Pizza | Sandwich | Soup |
|---|---|---|---|---|---|---|---|---|
| Burger | 1.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chicken_ Breasts | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chicken_ Nuggets | 0.00 | 0.10 | 0.70 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 |
| Chicken_salad | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cake | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| Pizza | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| Sandwich | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | | 0.90 | 0.00 |
| Soup | 0.00 | 0.10 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 | 0.90 |

PREDICTED CLASS

*Fig. 9:* Confusion Martix for Barfood 101 image dataset

Convolution Neural Network Confusion Matrix over PFID Image Set (Accuracy 95.78%)

| KNOWN | Burger | Chicken_ Breasts | Chicken_ Nuggets | Chicken_salad | Cake | Pizza | Sandwich | Soup |
|---|---|---|---|---|---|---|---|---|
| Burger | 1.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chicken_ Breasts | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chicken_ Nuggets | 0.26 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chicken_salad | 0.00 | 0.00 | 0.03 | 1.00 | 0.00 | 0.00 | 0.00 | 0.10 |
| Cake | 0.00 | 0.00 | 0.00 | 0.10 | 0.93 | 0.00 | 0.00 | 0.10 |
| Pizza | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| Sandwich | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.89 | 0.00 |
| Soup | 0.05 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.90 |

PREDICTED CLASS

*Fig. 10:* Confusion Martix for PFID image dataset

The entries in the matrix are True Positive (TP) rate, True Negative (TN) rate, False Positive (FP) rate, False Negative (FN) rate for each type of dataset. The accuracy (*AC*) is the ratio of the total number of predictions that were correct. It is derived by the equation:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

The confusion matrix shows that we get different accuracy but very closer via the same algorithm.We got 99.13% accuracy for Barfood 101 dataset whereas we achieved around 95.79% accuracy over PFID dataset which is higher than the accuracy obtained with Bag of SIFT or Bag of Surf (94%).Fig 12 shows the final output that truly identify a food item.

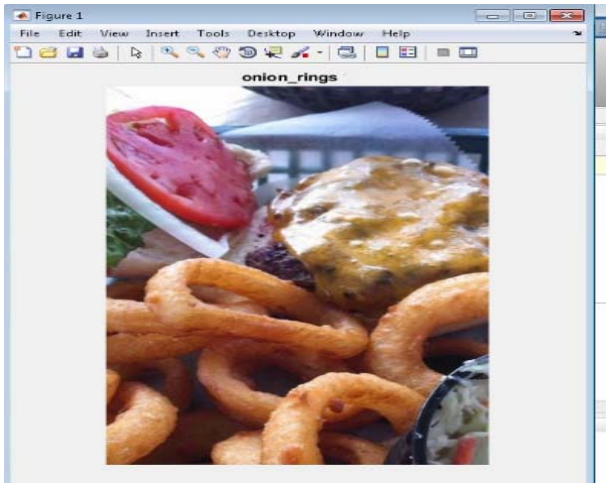


*Fig. 11:* True recognition of food item

Fig. 13 represents true detected result of sample image and shows an output image for False Negative predicted result. The output image shows this is an onion ring but the sample image contain a hamburger. This error is occurred due to different texture, shape or color feature and lighting effect.

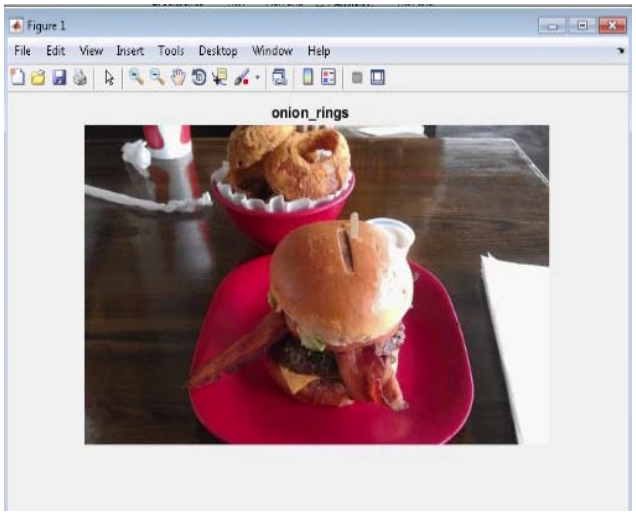


*Fig. 13:* False positive prediction

## VI. Conclusion

In this paper, we proposed a method to classify and to identify high calorie snacks (such as burger, pizza etc.) from the test image to measure the amount of calories has taken. In our experiment we apply CNN in PFID dataset that provides the accuracy 94% which is better than BOF. Also the false positive rate is not so high. People today are very conscious about their health. So, along with the patient, the health conscious person who has a major effect of food calories can be benefitted with this approach. In future, we will try to improve the accuracy by building a robust system which will identify all kinds of snacks more accurately.

## References Références Referencias

1. Dixon, J. B. (2010). The effect of obesity on health outcomes. Molecular and cellular endocrinology, 316(2), 104-108.
2. www.google.com
3. www.wikipedia.com
4. Chen, M.; Dhingra, K.; Wu, W.; Yang, L.; Sukthankar, R.; Yang, J. PFID: Pittsburgh Fast Food Image Dataset. In Proceedings of the ICIP 2009, Cairo, Egypt, 7–10 November 2009; pp. 289–292
5. Lowe, D.G. Object Recognition from Local ScaleInvariant Features. In Proceedings of the ICCV'99, Corfu, Greece, 20–21 September 1999; pp. 1150–1157.
6. Chen, M. Y., Yang, Y. H., Ho, C. J., Wang, S. H., Liu, S. M., Chang, E., & Ouhyoung, M. (2012, November). Automatic chinese food identification and quantity estimation. In SIGGRAPH Asia 2012 Technical Briefs (p. 29). ACM.
7. Probst, Y., Nguyen, D. T., Tran, M. K., & Li, W. (2015). Dietary assessment on a mobile phone using image processing and pattern recognition techniques: Algorithm design and system prototyping. Nutrients, 7(8), 6128-6138.
8. Yanai, K., & Kawano, Y. (2015, June). Food image recognition using deep convolutional network with pre-training and fine-tuning. In Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on (pp. 1-6). IEEE.
9. Kagaya, H.; Aizawa, K.; Ogawa, M. Food Detection and Recognition using Convolutional Neural Network. In Proceedings of the MM'14, Orlando, FL, USA, 3–7 November 2014; pp. 1055– 1088.
10. Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., & Cagnoni, S. (2016, October). Food image recognition using very deep convolutional networks. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management (pp. 41-49). ACM.

11. Christodoulidis, S., Anthimopoulos, M., & Mougiakakou, S. (2015, September). Food recognition for dietary assessment using deep convolutional neural networks. In International Conference on Image Analysis and Processing (pp. 458-465). Springer, Cham.
12. Singla, A.; Yuan, L.; Ebrahimi, T. Food/Non-Food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model. In Proceedings of the MADiMa'16, Amsterdam, The Netherlands, 15–19 October 2016; pp. 3–11.
13. Liu, C.; Cao, Y.; Luo, Y.; Chen, G.; Vokkarane, V.; Ma, Y. Deep Food: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. In Proceedings of the ICOST 2016,Wuhan, China,25–27 May 2016; pp. 37–48.
14. Lu, Y. (2016). Food Image Recognition by Using Convolutional Neural Networks (CNNs). arXiv preprint arXiv:1612.00983.
15. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
16. Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. 20(3), 273–297 (1995).
17. https://nlp.stanford.edu/IRbook/html/htmledition/multiclass-svms-1.html
18. https://www.mathworks.com/help/vision/examples/image-category-classification-using-deeplearning.html
19. http://gluon.mxnet.io/chapter06_optimization/gd-sgd-scratch.html#Stochastic-gradient-descent