July 2019

Food Places Classification in Egocentric Images using Siamese Neural Networks

Md. Mostafa Kamal SARKER [a,1], Syeda Furruka BANU [b], Hatem A. RASHWAN [a], Mohamed ABDEL-NASSER [a], Vivek Kumar SINGH [a], Sylvie CHAMBON [c], Petia RADEVA [d] and Domenec PUIG [a]

[a] *DEIM,Rovira i Virgili University 43007 Tarragona, Spain*
[b] *ETSEQ, Rovira i Virgili University, 43007 Tarragona, Spain*
[c] *CNRS-IRIT, INP-ENSEEIHT, Universit de Toulouse, 31071 Toulouse, France*
[d] *Department of Mathematics, University of Barcelona, 08007 Barcelona, Spain*

Abstract. Wearable cameras are become more popular in recent years for capturing the unscripted moments of the first-person that help to analyze the users lifestyle. In this work, we aim to recognize the places related to food in egocentric images during a day to identify the daily food patterns of the first-person. Thus, this system can assist to improve their eating behavior to protect users against food-related diseases. In this paper, we use Siamese Neural Networks to learn the similarity between images from corresponding inputs for one-shot food places classification. We tested our proposed method with 'MiniEgoFoodPlaces' with 15 food related places. The proposed Siamese Neural Networks model with MobileNet achieved an overall classification accuracy of 76.74% and 77.53% on the validation and test sets of the "MiniEgoFoodPlaces" dataset, respectively outperforming with the base models, such as ResNet50, InceptionV3, and InceptionResNetV2.

Keywords. Egocentric vision, food pattern classification, siamese neural networks, one-shot learning, scene classification.

1. Introduction

Currently, overweight and obesity are major health issues in high-income countries. Many major risk factors for chronic diseases, including diabetes, cardiovascular diseases, and cancer are caused by overweight and obesity. According to the WHO[2] the obesity fact is increasing dramatically and become tripled since 1975. More than 1.9 billion adults of 18 years age having overweight and 650 million of them are obese counted in 2016 [1]. Therefore, the concern of avoiding obesity is extremely demanding in developed countries. Contrarily, the cost of health services caused by overweight and obesity are growing for the government each year to billions of dollars. For instance, the medical cost for obesity in Europe was approximated at about 81 billion in 2012. In keeping with the WHO estimates on obesity expenditure, this was 2%-8% of the total national expenditure in the 53 European countries [2].

[1] Corresponding Author: E-mail: mdmostafakamal.sarker@urv.cat.
[2] http://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Figure 1. Examples of images of food places from mini "MiniEgoFoodPlaces" dataset

Food places or environments, adverse interactions to food, food eating patterns are the main key features for the health care professional for understanding nutrition intakes and considering the possible risk of obesity. A recent study finds that thousands of cancer diagnoses tied to a poor diet[3]. What, where and how long we are eating affects our health. To combat overweight and obesity, we need to improve our lifestyle and eating patterns that can help to a healthy diet. Thus, food eating patterns are one of the major key aspects that have to be analyzed for preventing overweight and obesity. Capturing daily user information by the traditional camera is difficult. Recently, wearable cameras (lifelogging) can able to capture daily user information (see Fig. 1). Therefore, we prefer to use "Narrative clip V2' wearable camera[4] which capable of continuously capturing images that record visual information of first-person daily life known as "visual lifelogging". It can record lots of images by the continuous image capturing capacity which can develop a visual diary with activities of first-person daily life with unprecedented details [3]. Moreover, the information extracted from these images can considerably affect human behaviours, habits, and even health [4]. The food environment can badly affect peoples health, some people get hungrier if they continuously see and smell food, consequently they end up eating more [5,6]. Thus, tracking the period of food intake in food places will help to improve their food eating behaviors.

Traditional nutrition diaries are not suitable for monitoring the lifestyle and food patterns properly since it needs a large number of human interaction. Currently, mobile phones are commonly used for tracking the user diet by keeping the record of the food intake and its related calories. However, this is done by capturing the photos of the foods, which can make people uncomfortable. Therefore, we need an automatic system that can correctly record the user food patterns and help to analyze the lifestyle and nutrition as well. The main motivation of this research is using a wearable camera for capturing images from food places, where the users are engaged within foods (see Fig. 1) and analysis of everyday information (entering, exiting and period of stay time as shown in Fig. 2) of visited food places. This can develop a novel health care application that can prevent diseases related to food, like obesity, diabetes, heart diseases, and cancer. The main contributions of this work can be summarized as follows:

[3]https://edition.cnn.com/2019/05/22/health/diet-cancer-risk-study/index.html
[4]http://getnarrative.com/

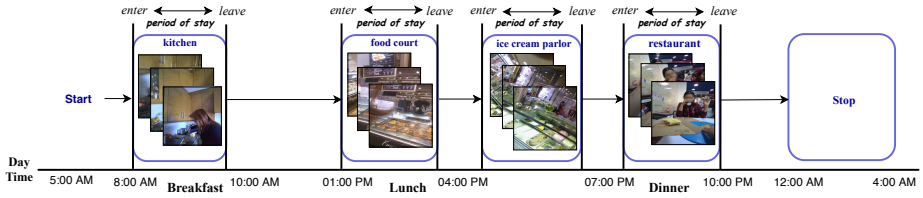

Figure 2. Examples of regularly spending time in food places.

- Introduce a small dataset called "MiniEgoFoodPlaces" with very few labeled images based on our previous "EgoFoodPlaces" dataset.
- Propose a Siamese Neural Networks (SNNs) for one-shot food places classification for the first time in places or scene classification domain. The outcomes show promising results for our classification task.

The paper is organized as follows. Section 2 discusses the related works of places or scene classification. The proposed SNNs based one-shot food places classification is described in Section 3. The experimental results and discussions are illustrated in Section 4. Finally, section 5 shows the conclusions and future work.

2. Related Works

Early work of places or scene recognition in conventional images has been discussed in the literature by applying classical approaches [7,8,9,10]. Recent breakthroughs of Convolutional Neural Networks (CNN), such as VGG16 [11], Inception [12] and ResNet50 [13] introduced new era of image classification. The era of places or scene classification turned into new dimensions after introducing two large-scale places datasets, Places2 [14] and SUN397 [15] with millions of labeled images. The combination of using deep CNNs models with large-scale datasets outperform the traditional places recognition methods. An overall state-of-the-art place or scene classification using deep neural networks has been summarized in a review article [16].

Recently, analyzing egocentric image is a highly promising field within computer vision and the era of this vision is called by "egocentric vision" that can help to develop algorithms for understanding the first person personalized scenes. A few research has been done for the scene classification in egocentric images. In [17], many classifiers were used to classify 10 different categories of scenes based on egocentric videos. They trained the classifiers by using One-vs-All cross-validation. Moreover, a multi-class classifier with a negative-rejection method was proposed in [18]. Both works [17,18] studied only 10 categories of scenes, only 1 of them are related to food places (i.e., *kitchen*).

Initially, we introduced two deep neural network named "MACNet" [19] based on atrous convolutional networks [20] and "MACNet+SA" [21] based on multi-scale atrous convolutional networks with a self-attention mechanism for the food places classification. Later on, we also introduced a semantic hierarchical approach for food scene classification [22]. However, food places classification is still a challenge due to the large diversity of food places environments in real-world, and the wide area of possibilities of how a scene can be captured from the person's point of view. Thus, we re-define our problem based on the similarity measure of the image of every place that can help to find

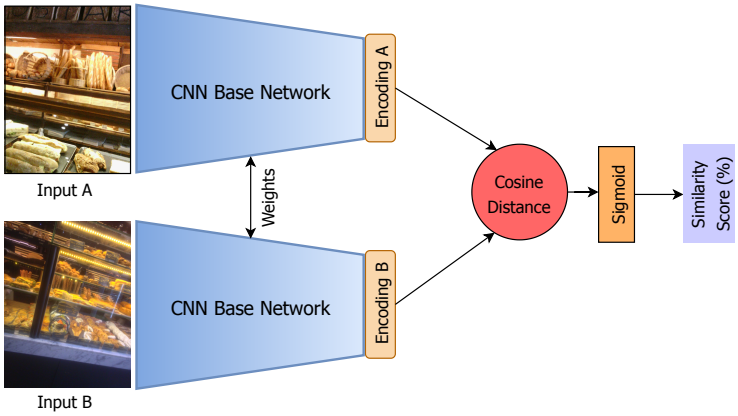

Figure 3. Architecture of our proposed Siamese Neural Networks model for food places classification.

the correlation with the images of the same places. The Siamese Neural Networks(SNNs) can help to learn the similarity to classify these places images using the one-shot approach. One-shot learning has been applied to different multimedia task including visual categorization [23,24,25]. Fei-Fei Li et al. [23] introduced a formal description of one-shot learning for object category tasks. In [24] introduced another kind of one-shot learning, in which the authors propose a SNNs to the rank similarity between input images. To the best of our knowledge, this is the first work on the food pattern classification based on one-shot learning using egocentric images to improve the healthy eating behavior of people.

3. Proposed Model

To train supervised deep learning models are needed large scale training data. which is highly time-consuming and costly. In one-shot learning, it is not feasible to get a huge amount of dataset for training. Therefore, one-shot learning with Siamese Neural Networks (SNNs) become more popular for image classification using small labeled data. Thus, we propose a SNNs for classifying food places. The details about the network described in next.

Traditional neural networks are learning to classify their inputs. Siamese networks are a special kind of neural network architecture that learns to differentiate between two input. The proposed SNNs is shown in Figure 3. The network gets a pair input A and B images, containing either 'similar' or 'dissimilar' food places images. Therefore, we used the binary label for every pair of images, 1 and 0 for the similar and dissimilar images respectively. Initially, the network uses a shared base network for generating encodings for both A and B input images. Consequently, these encodings are compared with one to another by using a cosine distance similarity metric.

$$cos(A,B) = \frac{A \cdot B}{||A|| \cdot ||B||}, \tag{1}$$

Finally, the resulted similarity is then classified using a sigmoid layer and obtained the final similarity score. For the base CNN networks, we use pre-trained models of Mo-

bileNet [26],InceptionV3 [27],InceptionResNetV2 [28],ResNet50 [13].We use transfer learning to train the base networks by removing the final fully connected (FC) layers and added a new FC and sigmoid layers on the top of the pre-trained base models.

4. Experimental Setup and Results

The SNNs is trained on a "MiniEgoFoodPlaces" dataset created from the "EgoFoodPlaces" dataset by reducing the number of images for balancing the class categories. We reconstruct the dataset from 22 to 15 classes because of some discriminating scenes like *pizzeria* and *fast-food restaurant* is artificial if the scene is recorded from a first-person view, and hence, we merged them to a *restaurant* class. The new dataset consists of 15 food places with 9750 images (7500, 750 and 1500 for training, validation and test images respectively). Initially, the pairs of similar and dissimilar categories are created from the train, test and validation sets. For creating the dissimilar pair, an image of class c is randomly paired with an image associated to one of the $n-1$ remaining classes, while pairs are created by categorizing the i^{th} and $(i+1)^{th}$ images of class c together.

The proposed model was implemented in Keras[29]: an open source deep learning library. The RMSprop [30] algorithm is used for model optimization. The "step" learning rate policy [31] is used with the base learning rate of 0.001 with 30 as a step value. The batch size is set to 32 for training with 100 epochs. The experiments are executed on NVIDIA GTX1080-Ti with 11 GB memory taking around 10 minutes to train the network. All the above parameters are used for testing the model as well.

Table 1. The average one-shot classifcation accuracy rates of the two validation and test sets of the MiniEgoFoodPlaces dataset with prpoposed SNNs using MobileNet [26], ResNet50 [13], InceptionV3 [27], Inception-ResNetV2 [28] CNN base networks.

CNN Base Network	Average Accuracy (%)	
	Validation	Test
MobileNet	**76.74**	**77.53**
ResNet50	73.96	74.10
InceptionV3	69.99	71.54
Inception-ResNet-v2	49.25	50.91

In this section, we have compared the proposed SNNs with the four base models, MobileNet [26], InceptionV3 [27], InceptionResNetV2 [28], ResNet50 [13] for both validation and test sets. Table 1 shows the average accuracy of the SNNs models. The MobileNet base SNNs yielded the highest average accuracy 76.74% and 77.53% in both validation and test dataset respectively. On the other hand, ResNet50, InceptionV3 and Inception-ResNet-V2 achieved overall accuracy of 73.96%, 69.99% and 49.25% on validation and 74.10%, 71.54% and 50.91% on test set. Currently, ResNet50, InceptionV3, and Inception-ResNet-V2 get better performance to compare with MobileNet in single image classification tasks. However, these models having a large number of training parameters and computational costly, While MobileNet having fewer parameters and less computational cost to train the model. Moreover, SNNs model with small CNN base network can achieve comparable results with a few training time, while all large models taking a long day or weeks.

5. Conclusion

In this paper, we proposed a Siamese Neural Networks for one-shot food places classification system. The main goal of this classification system is to analyze the first person egocentric images to create a dietary report of everyday food intake and help them control their unhealthy dietary habits. Instead of classifying one image, we use SNNs to classify the similarity between two images. Experiments show that the proposed approach can obtain a comparable result with a very small labeled dataset. Which means that it can provide an efficient solution for avoiding data labeling in a large scale image classification task.The proposed SNNs model with MobileNet CNN base network yields better performance the other state-of-the-art image classification models. Future work aims at developing a mobile application based on the SNNs model that integrates an egocentric camera with a personal mobile device to create a dietary report to keep a track on our eating behavior or routine for following a healthy diet.

Acknowledgement. This research is funded by the program Marti Franques under the agreement between Universitat Rovira Virgili and Fundació Catalunya La Pedrera. This work was partially funded by TIN2015-66951-C2-1-R, SGR 1742, and CERCA Programme / Generalitat de Catalunya. P. Radeva is partially supported by ICREA Academia 2014. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of boards Titan Xp GPU.

References

[1] C.M. Hales, C.D. Fryar, M.D. Carroll, D.S. Freedman and C.L. Ogden, Trends in obesity and severe obesity prevalence in US youth and adults by sex and age, 2007-2008 to 2015-2016, *Jama* **319**(16) (2018), 1723–1725.

[2] S. Cuschieri and J. Mamo, Getting to grips with the obesity epidemic in Europe, *SAGE open medicine* **4** (2016), 2050312116670406.

[3] M. Bolanos, M. Dimiccoli and P. Radeva, Toward storytelling from visual lifelogging: An overview, *IEEE Transactions on Human-Machine Systems* **47**(1) (2017), 77–90.

[4] E.R. Grimm and N.I. Steinle, Genetics of eating behavior: established and emerging concepts, *Nutrition reviews* **69**(1) (2011), 52–60.

[5] E. Kemps, M. Tiggemann and S. Hollitt, Exposure to television food advertising primes food-related cognitions and triggers motivation to eat, *Psychology & health* **29**(10) (2014), 1192–1205.

[6] R.A. de Wijk, I.A. Polet, W. Boek, S. Coenraad and J.H. Bult, Food aroma affects bite size, *Flavour* **1**(1) (2012), 3.

[7] A. Oliva and A. Torralba, Scene-centered description from spatial envelope properties, in: *International Workshop on Biologically Motivated Computer Vision*, Springer, 2002, pp. 263–272.

[8] J. Luo and M. Boutell, Natural scene classification using overcomplete ICA, *Pattern Recognition* **38**(10) (2005), 1507–1519.

[9] L. Cao and L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, 2007, pp. 1–8.

[10] J. Yu, D. Tao, Y. Rui and J. Cheng, Pairwise constraints based multiview features fusion for scene classification, *Pattern Recognition* **46**(2) (2013), 483–496.

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

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[13] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva, Learning deep features for scene recognition using places database, in: *Advances in neural information processing systems*, 2014, pp. 487–495.

[15] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva and A. Torralba, Sun database: Large-scale scene recognition from abbey to zoo, in: *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, IEEE, 2010, pp. 3485–3492.

[16] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, Places: A 10 million image database for scene recognition, *IEEE transactions on pattern analysis and machine intelligence* **40**(6) (2018), 1452–1464.

[17] A. Furnari, G.M. Farinella and S. Battiato, Temporal segmentation of egocentric videos to highlight personal locations of interest (2016), 474–489, Springer.

[18] A. Furnari, G.M. Farinella and S. Battiato, Recognizing Personal Locations From Egocentric Videos, *IEEE Transactions on Human-Machine Systems* **47**(1) (2017), 1–13, ISSN 21682291. doi:10.1109/THMS.2016.2612002.

[19] M. Sarker, M. Kamal, H.A. Rashwan, E. Talavera, S.F. Banu, P. Radeva and D. Puig, MACNet: Multi-scale Atrous Convolution Networks for Food Places Classification in Egocentric Photo-streams, *arXiv preprint arXiv:1808.09829* (2018).

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE transactions on pattern analysis and machine intelligence* **40**(4) (2018), 834–848.

[21] M.M.K. Sarker, H.A. Rashwan, F. Akram, E. Talavera, S.F. Banu, P. Radeva and D. Puig, Recognizing Food Places in Egocentric Photo-Streams Using Multi-Scale Atrous Convolutional Networks and Self-Attention Mechanism, *IEEE Access* **7** (2019), 39069–39082.

[22] E. Talavera, M. Leyva-Vallina, M. Sarker, M. Kamal, D. Puig, N. Petkov and P. Radeva, Hierarchical approach to classify food scenes in egocentric photo-streams, *arXiv preprint arXiv:1905.04097* (2019).

[23] L. Fei-Fei, R. Fergus and P. Perona, One-shot learning of object categories, *IEEE transactions on pattern analysis and machine intelligence* **28**(4) (2006), 594–611.

[24] G. Koch, R. Zemel and R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: *ICML deep learning workshop*, Vol. 2, 2015.

[25] R. Salakhutdinov, J. Tenenbaum and A. Torralba, One-shot learning with a hierarchical nonparametric bayesian model, in: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 195–206.

[26] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861* (2017).

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke and A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[29] F. Chollet et al., Keras: The python deep learning library, *Astrophysics Source Code Library* (2018).

[30] G. Hinton, N. Srivastava and K. Swersky, Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, *Cited on* **14** (2012).

[31] A. Sebag, M. Schoenauer and M. Sebag, Stochastic Gradient Descent: Going As Fast As Possible But Not Faster, in: *OPTML 2017: 10th NIPS Workshop on Optimization for Machine Learning*, 2017.