

Learning to Rank Food Images

Dario Allegra¹, Daniela Erba², Giovanni Maria Farinella¹, Giovanni Grazioso²,
Paolo Danilo Maci¹, Filippo Stanco¹, and Valeria Tomaselli³

¹ University of Catania, Italy

² University of Milan, Italy

³ STMicroelectronics, Italy

{allegra, gfarinella, fstanco}@dmi.unict.it

{daniela.erba, giovanni.grazioso}@unimi.it

valeria.tomaselli@st.com

Abstract. In the last decade food understanding has become a very attractive topic. This has implied the growing demand of Computer Vision algorithms for automatic diet assessment to treat or prevent food related diseases. However, the intrinsic variability of food, makes the research in this field incredibly challenging. Although many papers about classification or recognition of food images have been published in recent years, the literature lacks of works which address volume and calories estimation problem. Since an ideal food understanding engine should be able to provide information about nutritional values, the knowledge of the volume is essential. Differently from the state-of-art works, in this paper we address the problem of volume estimation through Learning to Rank algorithms. Our idea is to work with a predefined set of possible portion size and exploit a ranking approach based on Support Vector Machine (SVM) to sort food images according to the volume. At the best of our knowledge, this is the first work where food volume analysis is treated as a ranking problem. To validate the proposed methodology we introduce a new dataset of 99 food images related to 11 food plates. Each food image belongs to one over three possible portion size (i.e., small, medium, large). Then, we provide a baseline experiment to assess the problem of learning to rank food images by using three different image descriptors based on Bag of Visual Words, GoogleNet and MobileNet. Experimental results, confirm that the exploited paradigm obtain good performances and that a ranking function for food volume analysis can be successfully learnt.

Keywords: Learning to Rank · Ranking SVM · Food volume · Diet Monitoring

1 Introduction

Nowadays, people tend to ignore the impact that the food may have in their life. Unfortunately, an inadequate nutrition is one of the main causes of many chronic diseases such as obesity, diabetes, cancer, osteoporosis, dental diseases

and cardiovascular problems [1, 26]. Also, people ignoring healthy diet can incur in malnutrition problems.

Malnutrition can be defined as “any nutritional imbalance” that comprises over and under-nutrition and mainly involves elderly people, even in developed countries. The inadequate nutrition has been documented either in institutionalised as well as in free living elderly and exerts negative effects on health outcomes. The prevalence of malnutrition was estimated as 14% in older population. Moreover, the 28 – 45% of older people, recently moved to care homes or hospitals, were malnourished on admission [10]. This situation has serious consequences for individuals as well as for society, including the increasing risk for morbidity, mortality and consequently for social cost. However, nutrition screening of older adults is extremely difficult: some of the screening methods can be self-reported, with possible misreports, others can be only administered by trained clinicians, while the biochemical markers are time consuming and expensive to use.

In recent years, these facts led Computer Vision researchers to develop new solutions for automatically collected information during food intake in the context of people diet monitoring [3, 20, 23, 30]. Nevertheless, the intrinsic food variability in colour and shapes, as well as the great assortment of ingredients, makes very challenging the development of an efficient and effective food understanding engine. Ideally, a comprehensive system should be able to detect food dish, recognise the ingredients, estimate the volume and finally provide nutritional values. In the last decade, the spread of annotated datasets of food images [19, 16, 22, 5, 29, 13, 12, 11, 24] coupled with the massive use of learned based feature, have led promising performance for food detection and recognition tasks. On the contrary, quantity estimation studies have suffered the lack of proper datasets. Since 3D information have to be inferred for a correct volume estimation, in a totally unsupervised environment with no spatial references, it results in a extremely challenging problem [23, 8, 6, 25, 7, 2, 18].

However, in order to simplify patients diet monitoring, the current practice in healthcare facilities is to serve a set of standard food portions (e.g., small, medium, large). In this context, it is not necessary to estimate the exact amount of food in the dish, and the problem can be addressed using ranking strategies. In other words, given two images I_1 and I_2 , one looks for a function $rank(\cdot)$ such that $rank(I_1) < rank(I_2)$ if the food amount depicted in I_1 is lower then the food amount depicted in I_2 . With this idea in mind, and inspired by [21], we investigate the use of Learning to Rank algorithms [17] to sort food images according to the portion size.

To this aim, we introduce a new annotated dataset of 99 images and provide a baseline by using Ranking SVM [14] with three different kinds of visual features, i.e. Bag of words, GoogleNet features [27, 28] and MobileNet features[15]. Although Learning to Rank have been successfully used in Information Retrieval, Natural Language Processing and Data Mining [17], at the best of our knowledge this is the first time it is employed for food amount ranking. Hence, the novel contributions of this work is two-fold:

- a new annotated dataset to face food ranking problem with respect to the portion size;
- the assessment of Ranking SVM [21] to learn a ranking function with different visual features.

The paper is organized as following. In Section 2 we report the related works in the field. In Section 3 we focus on the contribution of this work by detailing the proposed dataset and the considering a Learning to Rank approach. Section 4 describes the experimental settings and reports results. Section 5 summarise conclusions.

2 Related Works

Several works related to food understanding are available in literature. Most of them focus on food detection, recognition and classification and are motivated by the increasing demand of assistive technologies for diet monitoring [23, 30, 3, 20].

In 2009, Puri et al. [23] proposed a system for food recognition and volume estimation. The classifier was obtained as a linear combination of multiple weak SVM classifiers trained on texture features. The volume was inferred by using RANSAC and dense stereo matching for depth map construction. In [8], Dehais et al. employed stereo pairs to compute disparity map and build a dense points cloud which is aligned to the table plane. This framework was designed to work by using a reference card placed on the table. By assuming the different food items in the plate are already segmented, each segment is projected on the 3D space for volume computation. In 2013, Chen et al. [6] presented a method for volume estimation from single view image. The approach requires a specific shape model for each food category and a calibrated reference marker. In 2017, Dehais et al. [7] proposed to estimate volume of multi-food meals with unconstrained 3D shape using stereo vision. The approach required two meals image of food placed inside elliptical plate, a credit card sized reference next to the plate and a segmentation map. Allegra et al. [2], proposed to use RGB-D images and supervised learning to perform depth estimation. The authors performed semantic segmentation by using U-Net and then they used a modified version of the CNN in [9] for depth inference from single RGB input. In 2018, Lu et al. [18] presented a Multi-Tasking Learning approach to estimate volume of food items from single RGB image. The proposed CNN architecture is composed by multiple modules. The first one is a feature extraction module and it is composed by ResNet50 and Feature Pyramid Network (FPN). The second module is the depth prediction net, which is mainly based on an encoder-decoder design architecture with skip connections and multi-scale side predictions. Semantic segmentation is performed by a Region Proposal Network (RPN) and a recognition net. The output of RPN is used to provide a food candidate mask and then the final volume was obtained through a CNN regressor.

Unlike the previous works, in this paper we propose to address food amount estimation as a ranking problem. We neither perform depth estimation nor other

3D reconstruction steps since we operate in a context with a set of fixed food portions (small, medium and large portions served at the canteen of the University). Ranking is one of the fundamental problems in information retrievals. Given a query q and a collection of documents D that match the query, ranking consists of sorting the documents according to some criterion. Learning to Rank (LTR) refers to machine learning strategies for training a model for a ranking task [17]. Specifically, we use LTR to train a model for a ranking task which lies in sorting RGB food images according to the relative visual attribute “portion size”. The concept of relative attribute was introduced in [21] by Parikh and Grauman. Their aim was to provide a way to estimate the degree of an attribute give an image. Hence, differently than predicting the presence of an attribute, a relative attribute indicates the strength of an attribute in an image with respect to other images.

3 Material and Methods

The main contribution of this paper lies in using Ranking SVM to train a model for ranking food images according to the portion size. Additionally, for validation purpose, we introduce a new dataset whose details are described in the following section. The dataset is publicly available for research purposes⁴. In order to provide a baseline, we perform a comparison between three different kinds of image descriptors which we exploit with Ranking SVM: Bag of Words [4], GoogleNet features [28] and MobileNet features [15].

3.1 Proposed dataset

The proposed dataset consists of 99 RGB images belonging to 11 different classes (es. “insalata orzo e verdura”, “cordon bleu”, “fusilli alla crudaiola”, etc.). Each image is associated to one portion size among three possible portion sizes: small, medium and large. For each class we collected 9 images corresponding to 3 images for each portion size. Moreover, in order to introduce variability, during acquisition we have used plates with two different diameters: 18cm (small plate) and 22.8cm (large plate). All the acquisition have been performed by a standard RGB camera fixed in a support and a centered top view with respect to the plate. In Fig. 1 are shown some examples of images belonging to the proposed dataset. At the best of our knowledge, state-of-art datasets are not suitable to test LTR approaches. Hence, differently from them, the proposed dataset includes multiple portion sizes for each dish to properly test LTR methods. To promote new task of ranking food images and to make repeatable our experiments, the proposed dataset is publicly available.

3.2 Ranking SVM

Ranking SVM is a popular rank method proposed by Herbrich et al. [14]. The idea behind Ranking SVM is to transform ranking into pairwise classification

⁴ Dataset Page: <http://iplab.dmi.unict.it/foodLTR/>

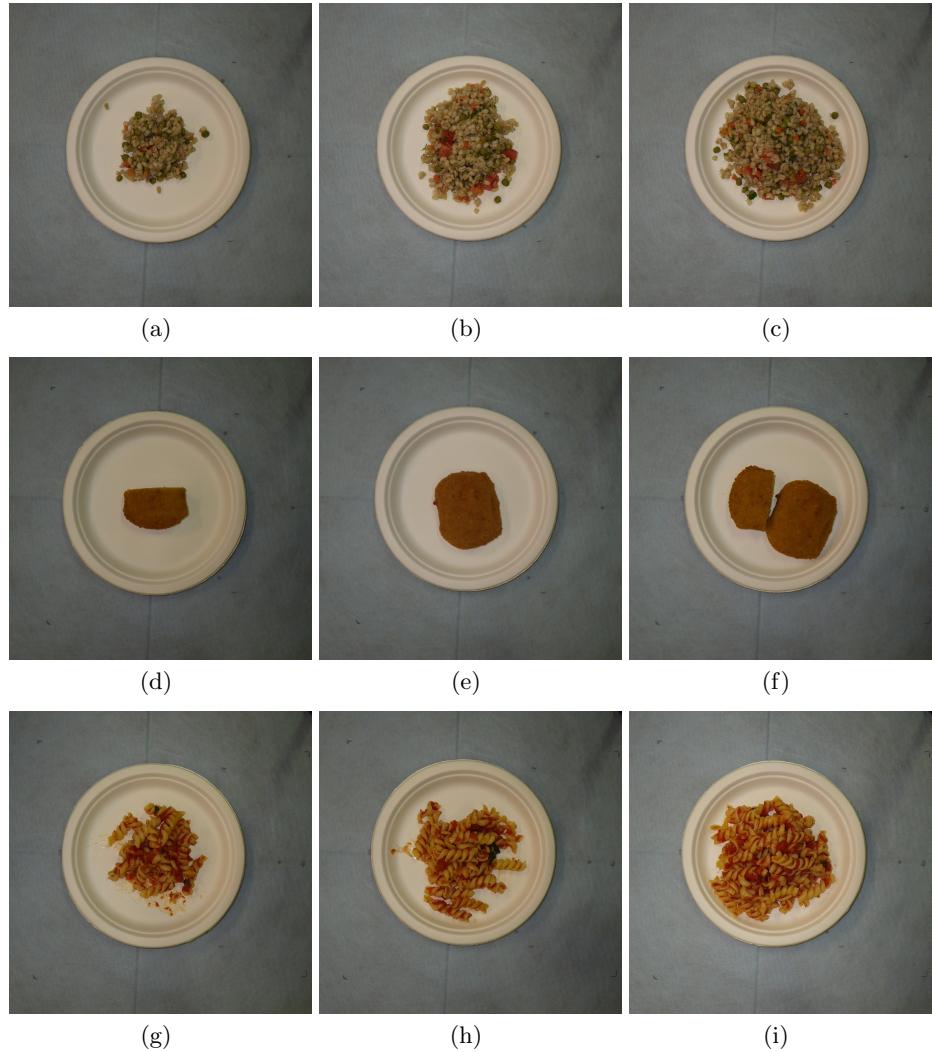


Fig. 1. Examples of three different food classes within the dataset (rows) and related portion sizes: small (first column) medium (second column) large (third column).

and employ the standard SVM strategy to perform the learning task according to one specific attribute. Given the image descriptor $\mathbf{x}_i \in \mathbb{R}^n$ of the image I_i , the final aim is to find a ranking function $r_m(\cdot)$:

$$r_m(\mathbf{x}_i) = \mathbf{w}_m^T \mathbf{x}_i \quad (1)$$

such that $r_m(\mathbf{x}_i) \geq r_m(\mathbf{x}_j)$ if and only if the strength of the attribute in I_i is higher than the strength of the attribute in the image I_j .

Hence, more formally, Ranking SVM is defined as the following constrained optimization problem:

$$\text{minimize} \quad \frac{1}{2} \|\mathbf{w}_m^T\|_2^2 + C \left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2 \right) \quad (2)$$

subject to:

$$|\mathbf{w}_m^T(\mathbf{x}_i - \mathbf{x}_j)| \geq 1 - \xi_{ij} \quad \forall (i, j) \in O_m \quad (3)$$

$$|\mathbf{w}_m^T(\mathbf{x}_i - \mathbf{x}_j)| \leq \gamma_{ij} \quad \forall (i, j) \in S_m \quad (4)$$

$$\xi_{ij} \geq 0 \quad \gamma_{ij} \geq 0 \quad (5)$$

where ξ_{ij} and γ_{ij} are slack variables to relax the constraints and control SVM margins; $C > 0$ is a regularization parameter to limit the growth of slack variables; O_m is the set of the ordered pair (i, j) such that image I_i has a higher presence of the attribute than the image I_j ; S_m is the set of the ordered pair (i, j) such that the images I_i and I_j have about the same presence of the attribute.

Hence, during the training phase, one has to provide the descriptors of the training images and the set O_m and S_m .

3.3 Image representation

In order to apply Ranking SVM, each RGB image I_i has to be described by a feature vector \mathbf{x}_i . In this study we employ three different strategies to build the descriptor \mathbf{x}_i , namely the Bag of Words paradigm with SURF, pre-trained GoogLeNet Inception v3 and pre-trained MobileNet. In the following, we report some details about the employed representation models.

Bag of visual Words All BoW approaches are based on frequency statistics on primitive unit (pixel, lexical unit, etc.) of a finite set. It means that a vocabulary of “words” must be built.

The general idea behind this method is to represent an image as a histogram of visual words frequency. As first step, key-point descriptors are extracted from the images. Subsequently, a clustering algorithm is employed to quantize the key-points feature space by identifying K centroids to be used as a vocabulary V composed by K visual words. Then, the final representation of an image consists of a normalised histogram H , where the bin H_i is related to the frequency of

the visual word $v_i \in V$. For experimental purpose, we have used Speeded Up Robust Features (SURF) [4] for key-points extraction, and K -Means algorithm to create a vocabulary with a specific size K . In our experiments we adopt a 64-dimensional key-point descriptor and a vocabulary size of $K = 500$.

GoogLeNet Inception V3 GoogLeNet Inception *v3* [28] is a Convolutional Neural Network (CNN) pre-trained on ImageNet. Its architecture consists of 42 layers. Differently from the previous version [27], in *v3* the computational burden is limited, but the effectiveness in term of accuracy is unaltered. This is realized by replacing 5×5 convolutions with 3×3 convolutions and by employing factorization methods. In this work, we used pre-trained GoogLeNet Inception *v3* to extract visual features from food images, so the last layer is removed and a Global Average Pooling layer is placed to get a 2048-dimensions feature vector.

MobileNet MobileNet [15] is based on a streamlined architecture that employs depthwise separable convolution to build light and efficient deep neural networks. Depthwise separable convolution is a form of factorized convolutions which replaces a standard convolution with both, depthwise convolution and a 1×1 convolution called pointwise. This makes MobileNet suitable for low-powerful devices. To extract visual features for our experiments, we used a MobileNet pre-trained on ImageNet. The architecture consists of 28 layers, but we remove the last one and use Global Average Pooling to get a 256-dimensions descriptor.

4 Experiments and Results

To assess the proposed approach, we employ three-fold cross validation method. Hence, we run three times the experiments by using 66 images for training purpose and 33 for testing. Finally, we average the results. Note that test set always includes 1 small, 1 medium and 1 large portion image for each of the 11 classes. Moreover, all the images are resized to 250×250 pixels. In training phase we consider all the possible pairs in order to define O_m and S_m (see Section 3.2), whereas in test phase we validate the method by considering the ordering inferred for images triplets. For a proper quantitative evaluation we use two different evaluation measures, i.e. accuracy and Hamming distance, that are detailed in the next sections.

4.1 Training phase

Training Ranking SVM model requires to build the set O_m and S_m described in Section 3.2. The number of all possible pairs (i, j) can be easily computed as $\binom{66}{2}$, hence $|O_m| + |S_m| = 2145$. Moreover, since S_m includes all the pairs related to the images with the same degree of the considered attribute (i.e, portion size), we have that $|S_m| = 3 \times \binom{22}{2} = 693$. The rest of 1452 pairs belong to O_m . We would like to highlight that although 66 images could be considered a limited number of data for training purpose, the actual training data for Ranking SVM lie in the pairs in $O_m \cup S_m$, i.e. 2145.

4.2 Testing phase

Since in this work we want to distinguish between three different portion size, we decide to evaluate the proposed method by considering images triplets. Hence, we use the $\binom{33}{3} = 5456$ possible triplets as testing samples. For each images triplet $\{I_s, I_p, I_q\}$ and the related descriptors $\{\mathbf{x}_s, \mathbf{x}_p, \mathbf{x}_q\}$, we compared the ground truth ranking $r^*(\cdot)$ (which we can compute using labels: small, medium, large) with respect to the one inferred through Ranking SVM approach, i.e. $r_m(\cdot)$.

4.3 Evaluation measures

To quantitatively evaluate the performances of a learned ranking function $r_m(\cdot)$, we used two different measures: accuracy and Hamming distance. In this context, the accuracy is related to the number of correctly sorted triplets over the total number of tested triplets. More formally, we define the accuracy of the ranking function $r_m(\cdot)$:

$$acc = \sum_{k=1}^M \left(\frac{\delta_k}{M} \right) \quad (6)$$

where M is the number of testing triplets. Given the ground truth ranking values $\{r^*(\mathbf{x}_s^k), r^*(\mathbf{x}_p^k), r^*(\mathbf{x}_q^k)\}$ related to the k -th triplet $\{I_s^k, I_p^k, I_q^k\}$, and assuming a ranking order $r^*(\mathbf{x}_s) \leq r^*(\mathbf{x}_p) \leq r^*(\mathbf{x}_q)$, the value of δ_k is defined as following:

$$\delta_k = \begin{cases} 1 & \text{if } r_m(\mathbf{x}_s) \leq r_m(\mathbf{x}_p) \leq r_m(\mathbf{x}_q) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

In a nutshell, we can say that $\delta_k = 1$ if the ground truth ranking order and the inferred ranking order agrees on a triplet.

Differently, Hamming distance evaluates the specific mismatching between the elements of the ground truth sorted triplet and the elements of the predicted one. Moreover, we assign a higher penalty if the mismatching occurs between a small portion and a large portion. For the sake of formalism, we define an ordered triplet as $\mathbf{T} = (I_s, I_p, I_q)$ and a function $label : \mathbb{I} \rightarrow \{1, 2, 3\}$:

$$label(I) = \begin{cases} 1 & \text{if image } I \text{ depicts a small portion} \\ 2 & \text{if image } I \text{ depicts a medium portion} \\ 3 & \text{if image } I \text{ depicts a large portion} \end{cases} \quad (8)$$

Then, if \mathbf{T}^* is a correctly ordered triplet and \mathbf{T}_m is the predicted one, the hamming distance between \mathbf{T}^* and \mathbf{T}_m is defined as:

$$hd(\mathbf{T}^*, \mathbf{T}_m) = \sum_{k=1}^3 | label(T_k^*) - label(T_{mk}) | \quad (9)$$

Table 1. Experimental results

	BoW	GoogLeNet	MobileNet
Accuracy	83.45%	87.84%	94.05%
Hamming dist.	0.2963	0.2440	0.1193

where T_k is the k -th images in the ordered triplet. In our experimental settings, Hamming distance is computed for all the testing triplets and then the average is considered.

4.4 Results

To benchmark the dataset and validate the proposed approach we perform test with BoW, GoogLeNet and MobileNet features on 5456 triplets. Experimental results in terms of both accuracy and Hamming distance are reported in Table 1.

Considering the evaluation in terms of accuracy, Ranking SVM on food images achieves best results when deep learning based features are used to represent images. This result is not surprising, since deep learning features tend to be more descriptive than the classic ones, even without fine-tuning. Specifically, MobileNet representation performed better than GoogLeNet in our experiments (94.05% vs 87.84%).

The performance in terms of Hamming distance follow the same trend. Since it measures the error degree, MobileNet presents the lowest average Hamming distance. For a better understanding of the values report in Table 1, we want to remark that, according the description given in Section 4.3, the higher value (i.e., the maximum error) for Hamming distance is 4. For example, it can be obtained if we compare to ordered triplet with labels (1, 2, 3) and (3, 2, 1) respectively.

5 Conclusion

The work presented in this paper is motivated by the massive demand for automatic systems for food intake monitoring. In a context where a predefined set of portion size are available, we proposed a novel approach for food portion sorting based on Learning to Rank algorithms. This allows to face the problem by avoiding depth and 3D estimation as in literature works which operate in less-constrained environments. To validate the proposed method, we introduce a new dataset of 99 images, which includes 11 categories and three distinct food portions, namely small, medium and large. Experiments have been performed with three different image descriptors based on Bags of Words, GoogleNet and MobileNet and with the LTR algorithm Ranking SVM. Results show that the evaluated method gives the opportunity for future extensions and industrial applications. Currently, we are planning to extend the dataset by performing new acquisitions in real healthcare facilities under the supervision of nutritionists.

6 Acknowledgments

We acknowledge Laser Consortium (Monza) for supporting the acquisition of food images used in this work.

References

1. Diet, nutrition and the prevention of chronic diseases. Tech. Rep. WHO Technical Report Series - 916, Report of a Joint WHO/FAO Expert Consultation (1 2002)
2. Allegra, D., Anthimopoulos, M., Dehais, J., Lu, Y., Stanco, F., Farinella, G.M., Mougiakakou, S.: A multimedia database for automatic meal assessment systems. In: Lecture Notes in Computer Science. vol. 10590 (2017)
3. Arab, L., Estrin, D., Kim, D.H., Burke, J., Goldman, J.: Feasibility testing of an automated image-capture method to aid dietary recall. European Journal of Clinical Nutrition **65**, 1156–62 (5 2011)
4. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European Conference on Computer Vision. pp. 404–417 (2006)
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 mining discriminative components with random forests. In: Lecture Notes in Computer Science. vol. 8694, pp. 446–461 (9 2014)
6. Chen, H.C., Jia, W., Yue, Y., Li, Z., Sun, Y.N., Fernstrom, J.D., Sun, M.: Model-based measurement of food portion size for image-based dietary assessment using 3d/2d registration. Measurement Science and Technology **24**(10) (2013)
7. Dehais, J., Anthimopoulos, M., Shevchik, S., Mougiakakou, S.: Two-view 3d reconstruction for food volume estimation. IEEE Transactions on Multimedia **19**, 1090–1099 (5 2017)
8. Dehais, J., Shevchik, S., Diem, P., Mougiakakou, S.G.: Food volume computation for self dietary assessment applications. In: International Conference on Bioinformatics and Bioengineering (11 2013)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Neural Information Processing Systems. vol. 3, pp. 2366–2374 (1 2014)
10. Elia, M., Stratton, R.J.: Geographical inequalities in nutrient status and risk of malnutrition among english people aged 65 y and older. Nutrition **21**(11), 1100–1106 (2005)
11. Farinella, G.M., Allegra, D., Moltisanti, M., Stanco, F., Battiato, S.: Retrieval and classification of food images. Computers in Biology and Medicine **77**, 23–39 (2016)
12. Farinella, G.M., Allegra, D., Stanco, F.: A benchmark dataset to study the representation of food images. In: Lecture Notes in Computer Science. vol. 8927, pp. 584–599 (3 2015)
13. Farinella, G.M., Allegra, D., Stanco, F., Battiato, S.: On the exploitation of one class classification to distinguish food vs non-food images. In: Lecture Notes in Computer Science. vol. 9281, pp. 375–383 (2015)
14. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Bartlett, P.J., Schölkopf, B., Schuurmans, D., Smola, A.J. (eds.) Advances in Large Margin Classifiers, vol. 88, pp. 115–132. MIT Press (2000)
15. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)

16. Kawano, Y., Yanai, K.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Lecture Notes in Computer Science. vol. 8927, pp. 3–17 (3 2014)
17. Li, H.: A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems* **94-D**, 1854–1862 (10 2011)
18. Lu, Y., Allegra, D., Anthimopoulos, M., Stanco, F., Farinella, G.M., Mougiaakakou, S.: A multi-task learning approach for meal assessment. In: Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. pp. 46–52 (2018)
19. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: International Conference on Multimedia and Expo. pp. 25–30 (7 2012)
20. O’Loughlin, G., Cullen, S.J., McGoldrick, A., O’Connor, S., Blain, R., O’Malley, S., Warrington, G.D.: Using a wearable camera to increase the accuracy of dietary analysis. *American Journal of Preventive Medicine* **44**, 297–301 (3 2013)
21. Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision. pp. 503–510 (2011)
22. Pouladzadeh, P., Yassine, A., Shirmohammadi, S.: Foodd: Food detection dataset for calorie measurement using food images. In: Lecture Notes in Computer Science. vol. 9281, pp. 441–448 (8 2015)
23. Puri, M., Zhu, Z., Yu, Q., Divakaran, A., Sawhney, H.: Recognition and volume estimation of food intake using a mobile device. In: Workshop on Applications of Computer Vision (12 2009)
24. Ragusa, F., Furnari, A., Farinella, G.M.: Understanding food images to recommend utensils during meals. In: Lecture Notes in Computer Science. vol. 10590, pp. 419–425 (2017)
25. Rhyner, D., Loher, H., Dehais, J., Anthimopoulos, M., Shevchik, S., Botwey, H.R., Duke, D., Stettler, C., Diem, P., Mougiaakakou, S.: Carbohydrate estimation by a mobile phone-based system versus self-estimations of individuals with type 1 diabetes mellitus: A comparative study. *Journal of Medical Internet Research* **18** (5 2016)
26. Suthumchai, N., Thongsukh, S., Yusuksataporn, P., Tangsripairoj, S.: Foodforcare: An android application for self-care with healthy food. In: International Student Project Conference (ICT-ISPC). pp. 89–92 (5 2016)
27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition. pp. 1–9 (6 2015)
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Z., W.: Rethinking the inception architecture for computer vision. In: Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
29. Xin, W., Kumar, D., Thome, N., Cord, M., Precioso, F.: Recipe recognition with large multimodal food dataset. In: International Conference on Multimedia Expo Workshops. pp. 1–6 (7 2015)
30. Zhu, F., Bosch, M., Woo, I., Kim, S., Boushey, C.J., Ebert, D.S., Delp, E.J.: The use of mobile devices in aiding dietary assessment and evaluation. *IEEE Journal of Selected Topics in Signal Processing* **4**, 756–766 (8 2010)