MARTINEZ, E.T., LEYVA-VALLINA, M., SARKER, M.M.K., PUIG, D., PETKOV, N. and RADEVA, P. 2020. Hierarchical approach to classify food scenes in egocentric photo-streams. *IEEE journal of biomedical and health informatics* [online], 24(3), pages 866-877. Available from: https://doi.org/10.1109/JBHI.2019.2922390

Hierarchical approach to classify food scenes in egocentric photo-streams.

MARTINEZ, E.T., LEYVA-VALLINA, M., SARKER, M.M.K., PUIG, D., PETKOV, N. and RADEVA, P.

2020

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

OpenAIR @RGU

This document was downloaded from https://openair.rgu.ac.uk

SEE TERMS OF USE IN BOX ABOVE

DISTRIBUTED UNDER LICENCE

Hierarchical approach to classify food scenes in egocentric photo-streams

Estefanía Talavera, María Leyva-Vallina, Md. Mostafa Kamal Sarker, Domenec Puig, Nicolai Petkov and Petia Radeva

***Abstract*—Recent studies prove that the environment where people eat, can affect their nutritional behaviour [1]. In this work we provide automatic tools for personalized analysis of a person's health habits by the study of daily recorded egocentric photo-streams. In particular, we propose a new automatic approach for the classification of food-related environment, that is able to classify up to 15 such scenes. In this way, people could monitor the context of their food intake in order to get an objective insight into their daily eating routine. We propose a model that classifies food-related scenes organized in a semantic hierarchy. Also, we present and make available a new egocentric dataset composed of more than 33000 images recorded by a wearable camera, over which we test our proposed model. Our approach obtains an average and weighted average classification accuracy of 75.46% and 63.20%, respectively, outperforming clearly the baseline methods.**

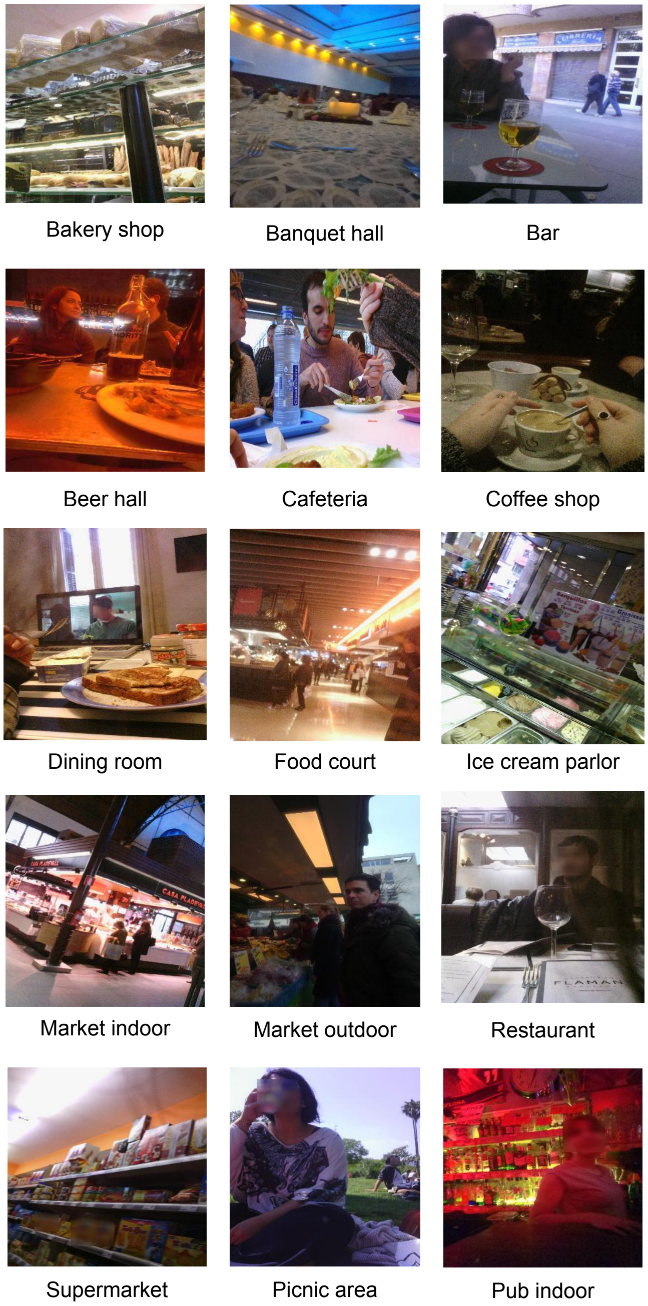
***Index Terms*—Egocentric vision, lifestyle, scenes classification, food scenes**

I. Introduction

NUTRITION is one of the main pillars of healthy habits. It is directly related to most chronic diseases like: obesity, diabetes, cardiovascular diseases, and also cancer and mental diseases [2], [3], [4]. Recent studies show that it is not only important *what people eat*, but also *how/where people eat* [1]. For example, it is well-known that who wants to lose weight is advised not to go to the supermarket while being hungry [5]. Social environment also matters; we eat more in certain situations, such as parties, than at home [6]. If we are exposed to food we feel the need or temptation to eat, same feeling of temptation that we experience at the supermarket [7]. Not only the sight plays its role, but also smell: everyone has walked in front of a bakery shop and felt tempted or hungry immediately [8]. The conclusion is that *where we are* can have direct impact on *what or how we eat* and, by extension, on our health [9]. However, there is a clear lack of automatic tools to monitor objectively the context of our food intake along time.

A. Our aim

Our aim is to propose an automatic tool based on robust deep learning techniques able to classify food-related scenes where a person spends time during the day. Our hypothesis is that if we can help people get insight into their daily eating routine, they can improve their habits and adopt a healthier lifestyle. By *eating routine*, we refer to activities related to the



Bakery shop
Banquet hall
Bar
Beer hall
Cafeteria
Coffee shop
Dining room
Food court
Ice cream parlor
Market indoor
Market outdoor
Restaurant
Supermarket
Picnic area
Pub indoor

Fig. 1: Examples of images of each of the proposed food-related categories present in the introduced EgoFoodPlaces dataset.

E. Talavera, M. Leyva-Vallina and N. Petkov are with the University of Groningen.
E-mail: estefania.tama@gmail.com
P. Radeva are with the University of Barcelona.
M.M.K. Sarker and D. Puig are with the University of Rovira i Virgili.

Manuscript received x, 2018; revised x, 2018.

acquisition, preparing and intake of food, that are commonly followed by a person. For instance, 'after work I go shopping and later I cook dinner and eat'. Or, 'I go after work directly to a restaurant to have dinner'. These two eating routines would affect us differently, having a direct impact on our health. The automatic classification of food-related scenes can represent a valuable tool for nutritionists and psychologists as well to monitor and understand better the behaviour of their patients or clients. This tool would allow them to infer how the detected eating routines affect the life of people and to develop personalized strategies for behaviour change related to food intake.

The closest approaches in computer vision to our aim focus either on scene classification, with a wide range of generic categories, or on food recognition from food-specific images, where the food typically occupies significant part of the image. However, food recognition from these pictures does not capture the context of food intake and thus does not represent a full picture of the routine of the person. It mainly exposes what the person is eating, at a certain moment, but not *where, in which environment*. These environmental aspects are important to analyze in order to keep track of the people behaviour.

B. Personalized Food-Related Environment Recognition

In this work, we propose a new tool for automatic analysis of food-related environments of a person. In order to be able to capture these environments along time we propose to use recorded egocentric photo-streams. These images provide visual information from a first-person perspective of the daily life of the camera wearer by taking pictures frequently: visual data about activities, events attended, environments visited, and social interactions of the user are stored. Additionally, we present a new labelled dataset that is composed of more than 33000 images, which were recorded in 15 different food-related locations.

The differentiation of food-related scenes that commonly appear in recorded egocentric photo-streams is a challenging task due to the need to recognize places that are semantically related. In particular, images from two different categories can look very similar, although being semantically different. Thus, there exists a high intra-class variance in addition to not high inter-class similarity. In order to face this problem, we consider a taxonomy taking into account the relation of the studied classes. The proposed model for food-related scene classification is a hierarchical classifier that embeds convolutional neural networks emulating the defined taxonomy.

Hence, the contributions of the paper are three-fold:

- A deep hierarchical network for classification of food-related scenes from egocentric images.
- A taxonomy of food-related environments organized in a fine-grained way that takes into account the main food-related activities (eating, cooking, buying, etc.). Our classifier is able to classify the different categories and subcategories of the taxonomy within the same model.
- An egocentric dataset of 33000 images and 15 food-related environments. We call it FoodEgoPlaces and, together with its ground-truth, is publicly available in http://www.ub.edu/cvub/dataset/.

As an example of application, we illustrate the utility of the proposed method for the detection of food-related scenes.

The paper is organized as follows: in Section II, we highlight some relevant works related to our topic, in Section III we describe the approach proposed for food scene recognition. In Section IV, we introduce our FoodEgoPlaces dataset and outline the experiments performed and obtained results. In Section V, we discuss the results achieved. Finally, in Section VI, we present our conclusions.

II. Previous Works

Scene recognition has been extensively explored in different fields, namely: robotics, security, environmental monitoring or egocentric videos. In this section, we describe previous works addressing this topic.

The recognition and monitoring of food-intake has been previously addressed in the literature [10], [11], [12]. For instance, in [10], the authors proposed the use of a microphone and a camera worn on the ear to get insight of the subject's food intake. On one side, the sound allows the classification of chewing activities, and on the other side, the selection of key-frames create overview of the food intake that otherwise would be difficult to quantify. An food-intake log supported by visual information allows to infer the food-related environment where a person spends time. However, no work has focus on this challenge so far.

A. Scene classification

The problem of scene classification was originally addressed in the literature by applying traditional techniques ([13], [14], just to mention a few), over handcrafted features. Nowadays, deep learning is the state-of-the-art.

As for the former case, one of the latest works on scene recognition using traditional techniques is [13], which aim was to recognize 15 different scenes categories of outdoor and indoor scenes. The proposed model was based on the analysis of image sub-region geometric correspondences by computing histograms of local features. In [14], the proposed approach focused on indoor scenes recognition, extending the number of recognized scenes to 67, where 10 of them are food-related. Having the hypothesis that similar scenes contain specific objects, their approach combines local and global image features for the definition of prototypes for the studied scenes. Very soon scene recognition was outperformed using deep learning.

Convolutional Neural Networks (CNNs) are a type of feed-forward artificial neural network, which connectivity patterns were inspired by the animal's visual cortex neurons connections [15]. Since Yann LeCun's LeNet [16] was introduced, many other deep architectures have been developed and applied to different computer vision known problems, achieving better results than the state-of-art techniques: MNIST [16] (images), Reuters [17](documents) and TIMIT [18] (recordings in English), ImageNET [19] (Data Sets classification), etc. Within the wide range of recently proposed architectures, some of the most popular are: GoogleNet [20], AlexNet [21], ResNet

[22], or VGGNet [23]. The use of CNNs for learning high-level features has shown a huge progress in scene recognition outperforming traditional techniques like [14]. This is mostly due to the availability of large datasets, those presented in [14], [24] or the ones derived from the MIT Indoor dataset ([25], [26]). However, the performance at *scene recognition* level has not reached the same level of success as *object recognition*. Probably, this is a result of the difficulty presented when generalizing the classification problem, due to the huge range of different environments surrounding us (e.g. 400 in the Places2 dataset [25]).

In [27], CNN activation features were extracted and concatenated following a spatial pyramid structure, and used to train one-vs-all linear classifiers for each scene category. In contrast, in [25] the authors evaluate the performance of the responses from the trained Places-CNN as generic features, over several scene and object benchmarks. Also, a probabilistic deep embedding framework, which analyses regional and global features extracted by a neural network, is proposed in [28]. In [29], two different networks called Object-Scene CNNs, are combined by late fusion; the 'object net' aggregates information for event recognition from the perspective of objects, and the 'scene net' performs the recognition with help from the scene context. The nets are pre-trained on the ImageNet dataset [19] and Places dataset [25] respectively. Recently, in [30] the authors combine object-centric and scene-centric architectures. They propose a parallel model where the network operates over different scale patches extracted from the input image. None of these methods have been tested on egocentric images that themselves represent a challenge for image analysis.

B. Classification of egocentric scenes

In order to obtain personalized scene classification we need to analyze egocentric images acquired by a wearable camera. Egocentric image analysis is a relatively recent field within computer vision concerning the design and development of Computer Vision algorithms to analyze and understand photo-streams captured by a wearable camera. In [31], several classifiers were proposed to recognize 8 different scenes (not all of them food-related). First, they discriminate between food/no-food and later, they train One-vs-all classifiers to discriminate among classes. Later, in [32] a multi-class classifier was proposed, with a negative-rejection method applied. In [31], [32] they only consider 8 scene categories, just 2 of them are food-related (*kitchen* and *coffee machine*) and without visual or semantic relation.

C. Food-related scene recognition in egocentric photo-streams

In our preliminary work presented in [33], we proposed a MACNet neural architecture for the classification of food-related scenes. This network input image is scaled into five different resolutions(the original image, with a scale value of 0.5). The five scaled images are fed to five blocks of atrous convolutional networks [34] with three different rates (1, 2, and 3) to extract the key features of the input image in multi-scale. In addition, four blocks of pre-trained ResNet are used to extract 256, 512, 1024 and 2048 feature maps, respectively. Each feature maps extracted by an atrous convolutional block is concatenated with the corresponding ResNet block to feed the subsequent block. Finally, the features obtained from the fourth ResNet layer is the final features are used to classify the food places images using two fully connected (FC) layers.

However, the challenge still remains due to the high variance that environments take in real-world places, and the wide range of possibilities of how a scene can be captured. In this work, we propose an organization of the different studied classes into semantic groups following the logic that relates them. We define a taxonomy, i.e. a semantic hierarchy relating the food-related classes. We organize environments according to the actions related to them: cooking, eating, acquiring food products. We demonstrate that by creating different levels of classification, and classifying scenes by the person action, it can serve as a natural prior for more specific environments and thus can further improve the performance of the model. The proposed classification model, implemented following this taxonomy, allows in an end-to-end process to analyze at different semantic levels of where the camera wearer spends time.

To the best of our knowledge, no previous work has focused on the problem of food-related scenes recognition at different semantic levels, either from conventional or egocentric images. Our work aims to classify food-related scenes from egocentric images recorded by a wearable camera. We believe that these images highly describe our daily routine and can contribute to the improvement of healthy habits of people.

III. Hierarchical approach for food-related scenes recognition in egocentric photo-streams

We propose a new model to address the classification of food-related scenes in egocentric images. It follows a hierarchical semantic structure, which adapts to the taxonomy that describes the relation among classes. The classes are hierarchically implemented from less to more specific ones. Therefore, the model is scalable and can be adapted depending on the classification problem, i.e. if the taxonomy changes.

For the purposes of food-related scene classification, we define a semantic tree which is depicted in Figure 2. We redefine the problem inspired by how humans hierarchically organize concepts into semantic groups. This organization groups classes that relate to the action *eating*, *preparing*, or *acquiring*. Later it splits eating into eating outdoor or indoor. Some of the subcategories group several classes, such as the subcategory *eating indoor* that encapsulates seven food-related scenes classes: bar, beer hall, cafeteria, coffee shop, dining room, restaurant, and pub indoor. In contrast, *preparing* and *eating outdoor* are represented uniquely by *kitchen* and *picnic area*, respectively. The semantic hierarchy was defined following the collected food-related classes and their intrinsic relation.

The differentiation among classes at the different levels of the hierarchy needs to be performed by a classifier. In this work, we propose to use CNNs for the different levels of classification of our food-related scenes hierarchy. The aggregation of CNNs layers mimics the food-related scenes structure presented in Fig. 2. Due to the good quality of the scene classification results over the Places2 dataset [26], we use the pre-trained VGG365 network [23] on which we build our hierarchical model. Note that this approach resembles the DECOC classifier [35] that proves the efficiency of decomposing a multi-class classification problem in several binary classification problems organized in a hierarchical way. The difference with the food-related scene classification is that in the latter case the classes are organized semantically in meta-classes corresponding to nutrition-related activities instead of constructing meta-classes without explicit meaning, but according to the entropy of training data [35].

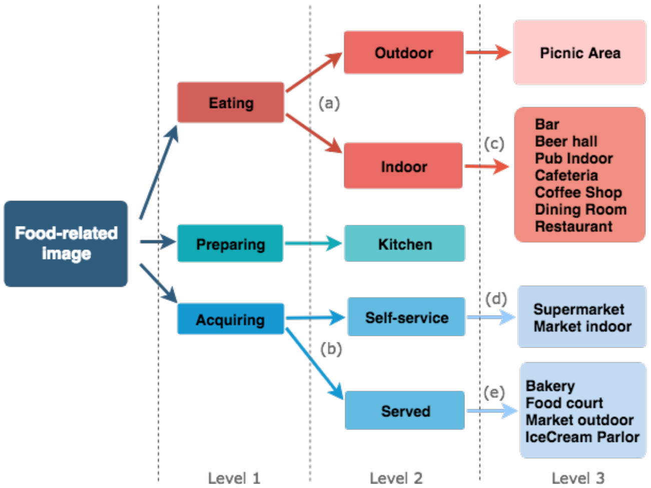


Fig. 2: Proposed semantic tree for food-related scenes categorization. For their later reference, we mark with dashed lines the different depth levels, and with letters the sub-classification groups.

Given an image, the final classification label is based on the aggregation of estimated intermediate probabilities obtained for the different levels of the hierarchical model, since a direct dependency exists between levels of the classification tree. The model aggregates the chain of probabilities by following the statistical inference method. The probability of an event is based on its prior estimated probabilities.

Let us consider classes C^i and C^{i-1} so that superscript shows the level of the class in the hierarchy and C^{i-1} is parent of C^i in the hierarchical organization of the tree. Thus, we can write:

$$P(C^i, x) = P(C^i, x|C^{i-1}, x) * P(C^{i-1}|x) \qquad (1)$$

where $P()$ relates to probabilities. $P(C^{i-1}, x|C^i, x)$ represents the likelihood of C^{i-1}, given image x, occurring given that C^i, given image x, is happening, while $P(C^i, x)$ and $P(C^{i-1}, x)$ are marginal probabilities given image x, i.e. the probabilities of independently observing C^i and C^{i-1}, respectively.

Note that we can estimate $P(C^i, x|C^{i-1}, x)$ from the classifier of the network trained to classify the classes children of class C^i, $P(C^{i-1}, x|C^i, x)$ is 1 since C^i is a subclass of C^{i-1}. $P(C^{i-1}, x)$ can be recursively estimated by considering the estimated probability on C^{i-1} and its class parent. Hence, we obtain that for each node C^i in the hierarchy (in particular, for the leaves), we get:

$$P(C^i, x) = \Pi_{j=1}^{i} P(C^j, x|C^{j-1}, x) * P(C^{j-1}, x) \qquad (2)$$

Without loss of generality, we consider that the probability of the class in the root is the probability to have food-related image, $(P(C^0))$, obtained by a binary classifier.

Let us illustrate the process with an example. Following the semantic tree in Fig. 2, our goal is to classify an egocentric image belonging to the class *dining room*. We observe that as *dining room* is a subclass of *indoor* and *indoor* is of *eating*, etc. Thus, the probability of *dining room* occurring giving image x is computed as:

$$\begin{aligned} &P(diningroom, x) \\ &\quad = P(diningroom, x|indoor, x) * P(indoor, x|eating, x) \\ &\qquad * P(eating, x|foodrelated, x) * P(foodrelated, x) \end{aligned} \qquad (3)$$

To summarize, given an image our proposed model computes the final classification as a product of the estimated intermediate probabilities at the different levels of the defined hierarchical tree.

IV. Experiments and Results

In this section, we describe a new home-made dataset that we make it public, the experimental setup, the metrics used to evaluate the analysis, and the results obtained.

A. Dataset

In this work, we present *EgoFoodPlaces*, a dataset composed of more than 33000 egocentric images from 11 users organized in 15 food-related scene classes. The images were recorded by a Narrative Clip camera[1]. This device is able to generate a huge number of images due to its continuous image collection. It has a frame rate of 2-3 images per minute. Thus, users regularly record an amount of approximately 1500 images per day. The camera movements and the wide range of different situations that the user experiences during his/her day, lead to new challenges such as background scene variation, changes in lighting conditions, and handled objects appearing and disappearing throughout the photo sequence.

Food-related scene images tend to have an intrinsic high inter-class similarity, see Fig. 1. To determine the food-related categories, we selected a subset of the ones proposed for the Places365 challenge [36]. We focus on the categories with a higher number of samples in our collected egocentric dataset, disregarding very unlikely food-related scenes, such as *beer garden* and *ice-cream parlor*. Also, we found that 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.

[1] http://getnarrative.com/

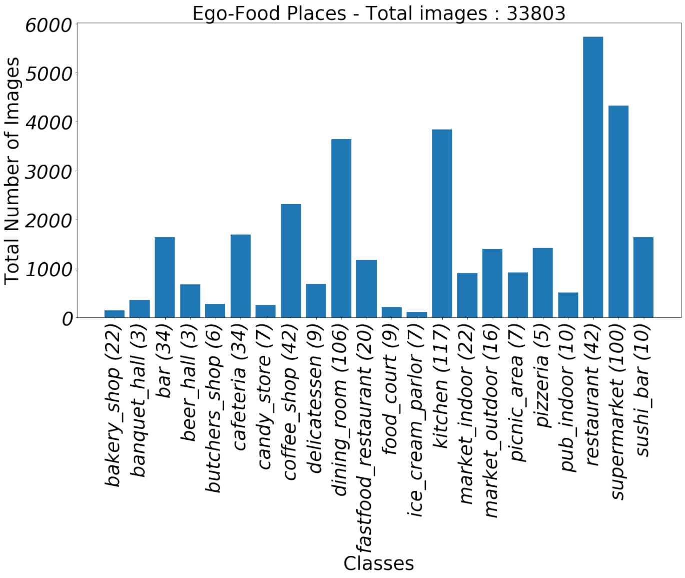


Fig. 3: Total number of images per food-related scene class. We give the number of collected events per class between parenthesis.

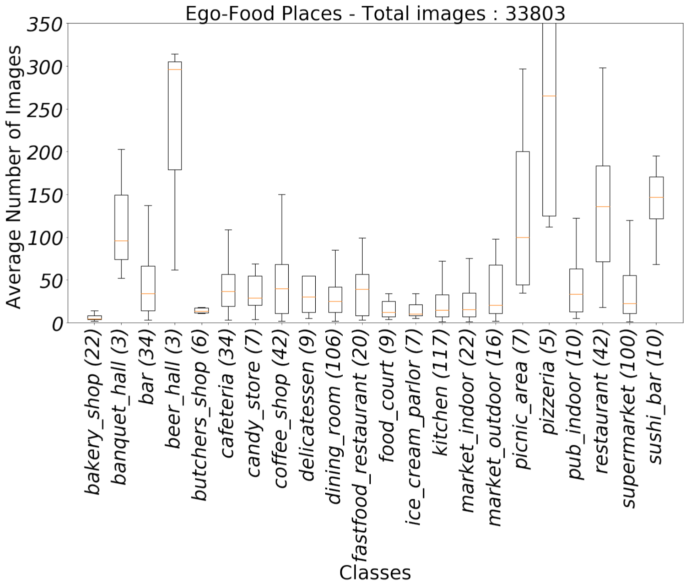


Fig. 4: Illustration of the variability of the size of the events for the different food-related scene classes. The data is presented by making the width of the box proportional to the size of the group. We give the number of collected events per class between parenthesis. The range of the data of a class is shown by the whiskers extend from its data box.

EgoFoodPlaces was collected during daily activities of the users. To build the dataset, we select the subset of images from the EDUB-Seg dataset that described food-related scenes, introduced in [37], [38], and later extended it with new collected frames. The dataset was gathered by 11 different subjects, during a total of 107 days, while spending time in scenes related to the *acquisition, preparing* or *consumption* of food. The dataset has been manually labelled into a total of 15 different food-related scenes classes: *bakery, bar, beer hall, cafeteria, coffee shop, dining room, food court, ice cream parlour, kitchen, market indoor, market outdoor, picnic area, pub indoor, restaurant, and supermarket.* In Fig. 3, we show the number of images per different classes. This figure shows the unbalanced nature of the classes in our dataset, reflecting the different prolongation of time that a person spends on different food-related scenes.

Since the images were collected by a wearable camera when performing any of the above mentioned activities, the dataset is composed by groups of images close in time. This leads to two possible situations. On one hand, images recorded 'sitting in front of a table while having dinner' will most likely be similar. On the contrary, in scenes such as 'walking at the supermarket' the images vary since they follow walking movement of the user in a very changeable environment.

In Fig. 4, we present the dataset by classes and events. This graph shows how the average, maximum and minimum spent time for the given classes differ. Note that this time can be studied since it is directly related to the amount of recorded images in the different food-related scenes. As we previously assumed, classes with a small amount of images correspond to not usual environments or environments where people do not spend a lot of time in (e.g. *bakery*). In contrast, the most populated classes refer to everyday environments (e.g. *kitchen, supermarket*), or to environments where more time is needed to complete performing activity (e.g. *restaurant*).

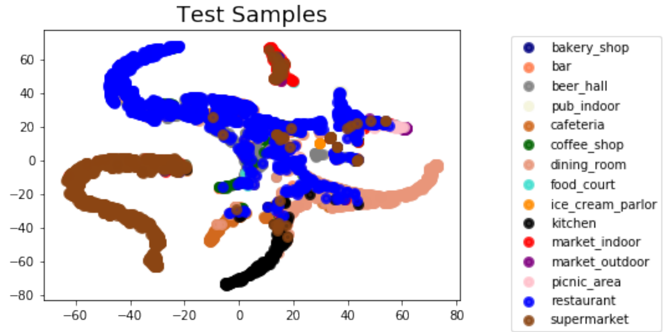


Fig. 5: Visualization of the distribution of the classes using the t-SNE algorithm.

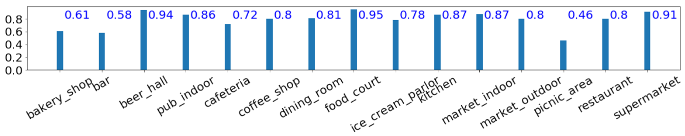


Fig. 6: Mean Silhouette Score for the samples within the studied food-related classes. The score is shown with bars and in blue text on top of them.

1) Class-variability of the EgoFoodPlaces dataset: To quantify the degree of semantic similarity among the classes in our proposed dataset, we compute the intra- and inter-class correlation. We use the classification probabilities output of the proposed baseline VGG365 network in order to find suitable descriptors for our images for this comparison. This network was trained for the classification of the proposed 15 food-related scenes. These descriptors encapsulate the semantic similarities of the studied classes.

To study the intra-class variability, we compute the mean silhouette coefficient for all samples, that is defined as,

$$Silhouette\_score = (b - a)/max(a, b) \tag{4}$$

where (a) corresponds to the intra-class distance per sample, and (b) corresponds to the distance between a sample and the closest class to which the sample is part of. Note that the silhouette takes values from 1 to -1; the highest value represents high density and separated clusters. The value 0 represents overlapping of clusters. Negative values indicate that there are samples with more similar clusters than the one they have been assigned to. The mean Silhouette score is 0.81 for the samples of our dataset depicted per class in Fig. 6. This value indicates that the classes are consistent and meaningful.

Furthermore, we visually illustrate the inter-class variability of the classes by embedding the 15 dimensional descriptor vector to 2 dimensions using the t-SNE algorithm [39]. The results are shown in Fig. 5. This visualization allows us to better explore the variability among the samples in the test-set. For instance, classes such as *restaurant* and *supermarket* are clearly distinguishable as a cluster. In contrast, we can recognize the classes with lower recognition rate, as the ones overlapping with *supermarket* and *restaurant*. For instance, *market indoor* is merged in its majority with *supermarket*. At the same time, the class *restaurant* clearly overlaps with *coffee shop* and *picnic area*.

B. Experimental setup

In this work, we propose to build the model on top of the VGG365 network [36] since it outperformed state-of-the-art CNNs when classifying conventional images into scenes. We selected this network, because it was already pre-trained with images describing scenes, and after evaluating and comparing its performance to the state-of-the-art CNNs. The classification accuracy obtained by the VGG16[23], InceptionV3[40], and ResNet50[41], were 65.09%, 71.62%, and 70.83%, respectively, lower than the 74.12% accuracy achieved by the VGG365 network.

We build our semantic hierarchical classification model by aggregating VGG365 nets over different subgroups of images/classes, emulating the semantic hierarchy proposed for food-related scenes recognition in Fig. 2. The final probability of a class is computed by the model, as described in Section III.

The model has an explicit semantic hierarchy that does not just aim to classify a given sample of food-related scenes, but also to get understanding of the semantic tree proposed. Therefore, we focus on the comparison of performances of existent methodologies with similar semantic classification approaches.

We compare the performance of the proposed model with the following baseline experiments:

1) FV: Fine-tuning of the VGG365 network with *EgoFoodPlaces*.
2) FV-RF: We use this categorical distribution obtained by the fine-tuned VGG365 in (1) as image descriptors. Later, we train the Random Forest classifier with 200 trees [42].
3) FV-SVM: Fine-tuned VGG365 to obtain image descriptors and Support Vector Machines [43].
4) FV-KNN: Fine-tuned VGG365 to obtain image descriptors and k-Nearest Neighbors [44] (n=3).
5) SVM-tree: We use the categorical distribution obtained by the fine-tuned VGG365 as images descriptors of the subsets of images that represents the nodes of the tree. Later, we train SVM as nodes of the proposed taxonomy.
6) MACNet [33]: We fine-tuned the MACNet network introduced in [33] to fit our proposed dataset.

We make use of the Scikit-learn machine learning library available for Python for the training of the traditional classifiers (SVM. RF, and KNN). For all the experiments, the images are re-sized at size 256x256. For the CNNs, we fine-tuned the baseline CNNs with a training batch size of 8, and run the validation set each 1000 iterations. The training of the CNNs was implemented using Caffe [45] and its Python interface. The code for the implementation of our proposed model is publicly available in https://github.com/estefaniatalavera/Foodscenes_hierarchicalmodel.

C. Dataset Split

In order to robustly generalize the proposed model and fairly test it, we assure that there are no images from the same scenes/events in both training and test sets. To this aim, we divide the dataset into events for the training and evaluation phases. Events are captured by sequentially recorded images that describe the same environment, and we obtain them by applying the SR-Clustering temporal segmentation method introduced in [38]. The division of the dataset into training, validation and test, aims to maintain a 70%, 10% and 20% distribution, respectively (see Table I). As it can be observed in Fig. 3, *EgoFoodPlaces* presents highly unbalanced classes. In order to face this problem, we could either subsample classes with high representation, or add new samples to the ones with low representation. We decide not to discard any image due to the relatively small number of images within the dataset. Thus, we balanced the classes for the training phase by over-sampling the classes with less elements. For all the experiments performed, the images used for the training phase are shuffled in order to give robustness to the network. Together with the EgoFoodPlaces dataset, the given labels, and the training, validation and test files are publicly available for further experimentation (http://www.ub.edu/cvub/dataset/).

D. Evaluation

We measure the performance of the proposed method by computing the normal and weighted accuracy. The use of weighted accuracy aims to face the unbalanced of the dataset, and intuitively expresses the strength of our classifier. This metric normalizes based on the number of samples per class.

TABLE I: Split of the proposed EgoFoodPlaces dataset for the performed experiments.

Categories	bakery shop	bar	beer hall	pub indoor	cafeteria	coffee shop	dining room	food court	ice-cream shop	kitchen	market indoor	market outdoor	picnic area	restaurant	supermarket	Num Images
Images Class	342	1632	926	1383	666	2920	3639	1057	107	3835	1731	1245	1608	7105	5607	33803
Training	301	1126	568	1198	175	2219	2636	1032	70	2660	1421	661	1348	4001	4195	23611
Validation	14	181	296	45	178	212	295	18	12	415	89	488	87	1105	498	3933
Test	27	325	62	140	313	489	708	7	25	760	221	96	173	1999	914	6259

Thus, one missed label will weigh more if we have less instances of the evaluated class, and vice-versa:

$$w_i = \frac{1}{n_i * C}.$$

For a class i with n_i instances, a hit or a miss will weigh w_i, if we have C different classes.

Moreover, we qualitatively compare the given labels by our method and the best of proposed baseline to sample images from the test set.

E. Results

We present the obtained classification accuracy at image level for the performed experiments in Table II. As it can be observed, our proposed model achieves the highest accuracy and weighted average accuracy, with 75.46% and 63.20%, respectively, followed by the SVM and Random Forest for the accuracy and SVM and KNN for the weighted accuracy.

Our proposed hierarchical model has the capability of recognizing not only the 15 classes corresponding to the leaves of the tree in the semantic tree (see Fig.2), but also the meta-classes at the different semantic levels of depth. Thus, specialists can analyze the personal data and generate strategy for improvement of life-style of people by studying their food-related behaviour either from a broad perspective, such as when the person *eats* or *shops*, or into a more detailed one, like *if the person used to eat in a fast-food restaurant or at home*.

But a logical question is if the model provides robust classification of meta-classes as well. To this aim, we evaluate the classification performance at the different levels of the defined semantic tree. Note that since each class is related to a meta-class on a higher level, an alternative to our model would be to obtain the meta-classes accuracy from their sub-classes classification. We compare the accuracy of meta-classes from their classification by the proposed model vs inferring the accuracy from the classification of the subclasses samples using a FV-SVM model (see Table III - on accuracy of L1, and L2 meta-classes). As one can observe in Table **??**, our model achieves higher accuracy classifying meta-classes in all cases. This proves that it is a robust tool for the classification of food-related scenes classes and meta-classes.

TABLE II: Food-related scene percentage of accuracy classification performance. We present the accuracy for each category, the average accuracy and the weighted accuracy for all models. We rename the fine-tuning of the VGG365 as 'FV', and the later use of its output probabilities for the training of the State-of-the-Art models.

Categories	BSN	FV	FV+RF	FV+KNN	FV+SVM	MACNet[33]
bakery shop	**37.04**	29.63	18.51	29.63	29.62	25.93
bar	33.54	25.85	22.77	25.54	26.77	**42.77**
beer hall	91.94	88.71	85.48	87.10	87.09	**96.77**
pub indoor	**82.86**	**82.86**	82.14	**82.86**	82.85	78.57
cafeteria	**73.80**	58.47	54.95	56.87	59.10	59.10
coffee shop	65.64	68.10	65.03	66.05	67.89	**69.73**
dining room	69.21	**70.90**	62.28	64.69	70.33	68.93
food court	**57.14**	42.86	28.57	42.86	57.14	42.85
ice cream parlor	48.00	**72.00**	44.00	68.00	**72.00**	20.00
kitchen	94.74	94.21	94.47	94.60	94.34	**95.53**
market indoor	77.38	**78.28**	70.14	72.40	76.92	61.09
market outdoor	**44.79**	30.21	21.88	28.13	30.20	35.42
picnic area	0.00	0.00	0.00	0.00	0.00	0.00
restaurant	79.99	78.74	87.49	**81.94**	78.84	77.99
supermarket	91.90	92.45	**93.44**	92.78	92.56	93.22
Avg Acc.	**75.46**	74.12	74.90	74.71	74.76	74.19
Weigh Avg Acc.	**63.20**	60.88	59.25	61.25	62.47	58.88

If we observe the confusion matrix in Fig. 7, we can get insight about the miss-classified classes. We can see how our algorithm tends to confuse the classes belonging to the semantic level of *self-service* (acquiring) and *eating indoor* (eating). We believe that this is due to the unbalanced aspect of our data and the intrinsic similarity within the sub-categories of some of the branches of the semantic tree.

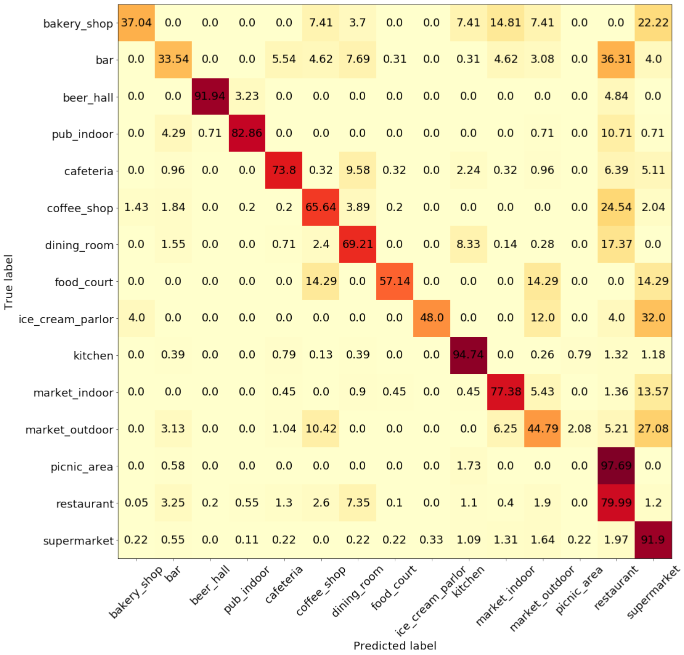


Fig. 7: Confusion matrix with the classification performance of the proposed hierarchical classification model.

TABLE III: Classification performance at different levels of the proposed semantic tree for food-related scenes categorization. We compute the achieved accuracy (Acc) per level, and the weighted accuracy (W-Acc) where we consider the number of samples per class. The different semantic levels (L) are introduced in Fig. 2.

		Our Method		FV		SVMTree		FV+RF		FV+SVM		FV+KNN	
		Acc	WAcc	Acc	WAcc	Acc	WAcc	Acc	WAcc	Acc	WAcc	Acc	WAcc
Fold 1	Level 1 (L1)	0.945	0.957	0.916	0.921	0.923	0.930	0.922	0.927	0.916	0.920	0.917	0.922
	Level 2a (L2a)	0.920	0.625	0.892	0.606	0.899	0.610	0.896	0.609	0.887	0.602	0.892	0.606
	Level 2b (L2b)	0.833	0.891	0.846	0.901	0.846	0.904	0.846	0.911	0.845	0.903	0.845	0.906
Fold 2	Level 1 (L1)	0.928	0.923	0.897	0.881	0.907	0.896	0.896	0.880	0.897	0.881	0.897	0.882
	Level 2a (L2a)	0.880	0.637	0.840	0.608	0.855	0.619	0.838	0.606	0.838	0.607	0.842	0.609
	Level 2b (L2b)	0.844	0.949	0.847	0.943	0.847	0.947	0.843	0.950	0.844	0.945	0.836	0.939
Fold 3	Level 1 (L1)	0.938	0.950	0.924	0.931	0.931	0.942	0.922	0.928	0.927	0.935	0.923	0.930
	Level 2a (L2a)	0.916	0.652	0.896	0.637	0.909	0.646	0.892	0.634	0.900	0.640	0.895	0.636
	Level 2b (L2b)	0.920	0.944	0.915	0.938	0.917	0.940	0.916	0.947	0.920	0.943	0.920	0.947
Fold 4	Level 1 (L1)	0.929	0.951	0.910	0.915	0.917	0.927	0.903	0.907	0.910	0.915	0.901	0.904
	Level 2a (L2a)	0.920	0.639	0.874	0.608	0.887	0.616	0.864	0.600	0.873	0.607	0.861	0.599
	Level 2b (L2b)	0.869	0.925	0.894	0.929	0.894	0.932	0.888	0.944	0.893	0.934	0.891	0.931
Fold 5	Level 1 (L1)	0.943	0.943	0.930	0.922	0.934	0.930	0.930	0.926	0.928	0.933	0.928	0.923
	Level 2a (L2a)	0.903	0.633	0.882	0.619	0.890	0.625	0.887	0.622	0.622	0.887	0.886	0.622
	Level 2b (L2b)	0.912	0.936	0.908	0.930	0.909	0.933	0.902	0.934	0.932	0.909	0.902	0.929
fold70	Level 1 (L1)	0.941	0.940	0.895	0.890	0.901	0.904	0.894	0.890	0.892	0.887	0.897	0.891
	Level 2a (L2a)	0.930	0.800	0.868	0.747	0.758	0.881	0.866	0.745	0.863	0.743	0.872	0.750
	Level 2b (L2b)	0.869	0.925	0.867	0.933	0.940	0.873	0.877	0.950	0.871	0.944	0.861	0.933
fold70-1	Level 1 (L1)	0.932	0.937	0.908	0.913	0.923	0.927	0.899	0.909	0.910	0.914	0.902	0.906
	Level 2a (L2a)	0.909	0.785	0.878	0.767	0.899	0.760	0.867	0.741	0.880	0.756	0.875	0.756
	Level 2b (L2b)	0.907	0.929	0.909	0.928	0.909	0.928	0.909	0.932	0.913	0.932	0.906	0.923
fold80	Level 1 (L1)	0.963	0.974	0.954	0.954	0.959	0.965	0.957	0.961	0.956	0.957	0.955	0.926
	Level 2a (L2a)	0.963	0.785	0.937	0.764	0.950	0.775	0.945	0.771	0.940	0.766	0.943	0.712
	Level 2b (L2b)	0.924	0.956	0.916	0.948	0.915	0.951	0.910	0.952	0.915	0.952	0.916	0.913
fold80-1	Level 1 (L1)	0.948	0.943	0.922	0.905	0.931	0.921	0.939	0.931	0.927	0.914	0.933	0.926
	Level 2a (L2a)	0.924	0.721	0.890	0.699	0.907	0.708	0.916	0.714	0.898	0.701	0.913	0.712
	Level 2b (L2b)	0.814	0.935	0.785	0.916	0.782	0.914	0.781	0.917	0.787	0.920	0.779	0.913

The classes with higher classification accuracy are *kitchen* and *supermarket*. We deduce that this is due to the very characteristic appearance of the environment that they involve and the amount of different images of such classes in the dataset. On the contrary, *picnic area* is not recognized by any of the methods. The confusion matrix indicates that the class is embedded by the model into the class *restaurant*. This can be inferred by visually checking the images, since in both classes a table and another person usually appear in front of the camera wearer. Moreover, from the obtained results, we can observe a relation between the previously computed Silhouette Score per class and the classification accuracy achieved by the classifiers. Classes with high consistency are better classified, while classes such as *bar, bakery shop, picnic area*, or *market outdoor* have lower classification performance.

The achieved results are rather similar. Therefore, we calculate the *t-test* to evaluate the statistical significance of the differences. Our proposed model outperforms FV, FV+RF, FK+KNN, FV+SVM, and MacNet with statistical significance (p=0.X, p=0.X, p=0.X, p=0.X, p=0.X, for paired t-test). The smaller the *p* value, the higher the statistical significance.

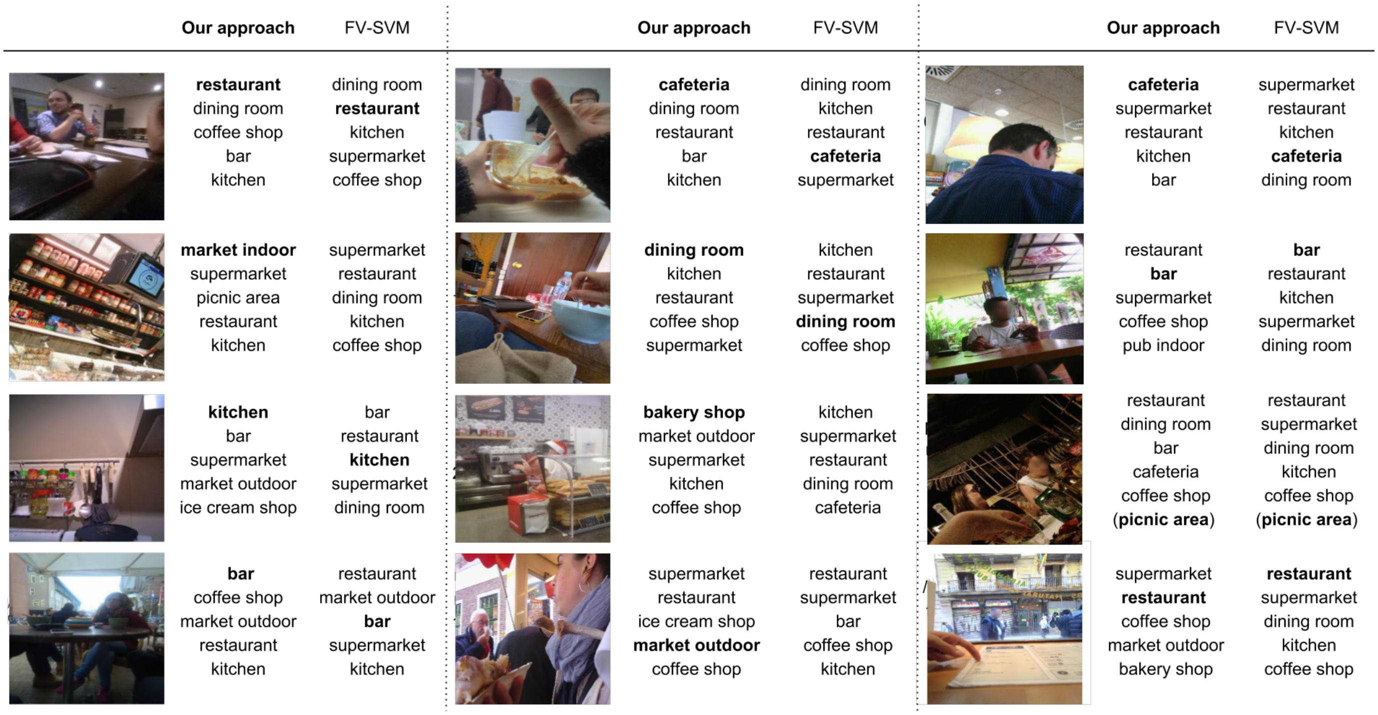
Qualitatively, in Fig. 8 we illustrate some correct and wrong classifications by our proposed model and the trained SVM (FV-SVM). We highlight the groundtruth class of the images in boldface. Even though the performance of the different tested models does not differ much, the proposed model has the ability to better generalize, as its weighted average accuracy indicates.

V. Discussions

The proposed dataset is composed by manually selected images from recorded day photo-streams. These extracted images belong to food-related events, described as groups of sequential images representing the same scene. It is to be highlighted that for the performed experiments, images belonging to the same event stayed together for either training or testing phase. Even though the classification of such scenes could have been events rather than images, we do not dispose of a higher number of events for the training phase in the case of event-based scene classification. The creation of a bigger egocentric dataset is a recurrent ongoing work. Next lines of work will address the analysis of events in order to study if they are connected and time-dependant.

Recorded egocentric images can be highly informative about the lifestyle, behaviour and habits of a person. In this work, we focus on the implementation of computer vision algorithms for data extraction from images. More specifically, on characterizing food-related scenes related to an individual for future assistance in controlling obesity and other eating disorders being of high importance for the society.

Next steps could involve the analysis other information e.g. the duration and regularity of nutritional activities. Based on extracted information regarding individuals, their daily habits can be extracted and characterized. The daily habits of people can be correlated to their personality, since people's routine affect them differently. Moreover, within this context social relations and their relevance can be studied: the number of people a person sees per day, the length and frequency of their meetings and activities, etc and how social context influence



Our approach	FV-SVM	Our approach	FV-SVM	Our approach	FV-SVM
restaurant dining room coffee shop bar kitchen	dining room **restaurant** kitchen supermarket coffee shop	**cafeteria** dining room restaurant bar kitchen	dining room kitchen restaurant **cafeteria** supermarket	**cafeteria** supermarket restaurant kitchen bar	supermarket restaurant kitchen **cafeteria** dining room
market indoor supermarket picnic area restaurant kitchen	supermarket restaurant dining room kitchen coffee shop	**dining room** kitchen restaurant coffee shop supermarket	kitchen restaurant supermarket **dining room** coffee shop	restaurant **bar** supermarket coffee shop pub indoor	**bar** restaurant kitchen supermarket dining room
kitchen bar supermarket market outdoor ice cream shop	bar restaurant **kitchen** supermarket dining room	**bakery shop** market outdoor supermarket kitchen coffee shop	kitchen supermarket restaurant dining room cafeteria	restaurant dining room bar cafeteria coffee shop **(picnic area)**	restaurant supermarket dining room kitchen coffee shop **(picnic area)**
bar coffee shop market outdoor restaurant kitchen	restaurant market outdoor **bar** supermarket kitchen	supermarket restaurant ice cream shop **market outdoor** coffee shop	restaurant supermarket bar coffee shop kitchen	supermarket **restaurant** coffee shop market outdoor bakery shop	**restaurant** supermarket dining room kitchen coffee shop

Fig. 8: Examples of top 5 classes for the images in the test set. We show the results obtained by the proposed model, and compare them with the obtained ones by the trained SVM classifier. The class in bold corresponds to the true label of the given image.

people. All this information extracted from egocentric images is still to be studied in depth leading to powerful tools for objective, long-term monitoring and characterization of behaviour of people for better and longer life.

The introduced model can be easily extrapolated and implemented to other classification problems with semantically correlated classes. Organizing classes in a semantic hierarchy and embedding a classifier to each node of the hierarchy allow to consider the estimated intermediate probabilities for the final classification.

The proposed model computes the final classification probability based on the aggregation of the probabilities of the different classification levels. The random probability of a given class is $1/|C|$, where $|C|$ is the number of children the parent class of that node has. Hence, having a high number of sub-classes (children nodes) for a specific node would tend to lower probability. There is risk that a 'wrong class node' gets higher final classification probability if it has few brother-sin the tree compared to the 'correct class node'.

1) Application to recorded days characterization:: Food-related scenes recognition is very useful to get understanding about the patterns of behaviour of people. The presence of people at certain food-related places is of importance when describing their lifestyle and nutrition. While in this work we focus on the classification of such places, we use the labels given to the photo-streams to characterize the camera wearer's 'lived experiences' related with food. The characterization given by the proposed model allow us to address the scene detection at different semantic levels. Thus, by using high level information we increase the robustness and the level of the output information of the model.

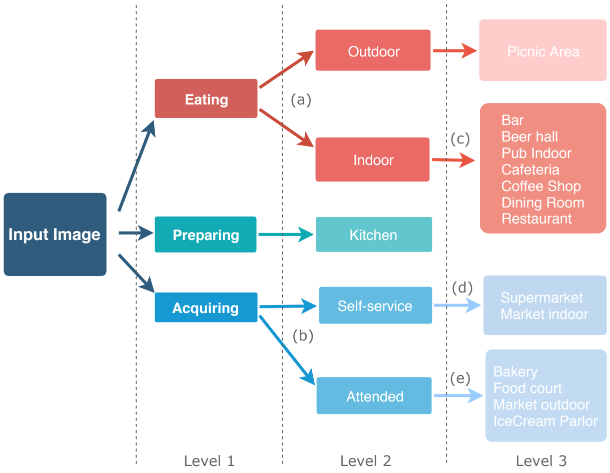


Fig. 9: Illustration of detecting food-related events in egocentric photo-streams of a camera wearer.

VI. Conclusions

In this paper, we introduced a multi-class hierarchical classification approach, for the classification of food-related scenes in egocentric photo-streams. The contributions of our presented work are three-fold:

- We propose a hierarchical model based on the combination of different layers of deep neural network, mirroring the given taxonomy for food-related scenes classification.

This model be easily adapted to other classification problems and implemented on top of other different CNNs and traditional classifiers.

- A taxonomy of food-related environments that considers the main activities related to food (eating, cooking, buying, etc.).
- We make publicly available the FoodEgoPlaces dataset and the model code. FoodEgoPlaces is composed by more than 3300 egocentric images describing 15 categories of food-related scenes of 11 camera wearers.

The performance of the proposed architecture is compared with several built baseline methods. We demonstrated that the proposed end-to-end semantic model based on a hierarchical network outperforms such methods. As an incentive, the proposed model has the ability of end-to-end automatically classifying different semantic levels of depth. Thus, specialists can analyze the nutritional habits of a person and generate recommendations for improvement of the life-style of people by studying their food-related behaviour either from a broad perspective, such as when the person *eats* or *shops*, or into a more detailed one, like *when the person is eating in a fast-food restaurant*.

As future work, we plan to explore how to enrich our data using domain adaptation techniques. Domain adaptation allows the adaptation of the distribution of data to other target data distribution. Egocentric datasets tend to be relatively small due to the low frequency rate of the recording cameras. We believe that by combining techniques of transfer learning, we will be able to explore how the collected dataset can be extrapolated to already available data, sets such as Places2. We expect that the combination of data distributions will improve the achieved classification performance. Therefore, further analysis on this line will allow us to get better understanding of people's lifestyle, which will give insight into their health and daily habits.

Acknowledgment

This work was partially founded by projects TIN2015-66951-C2, SGR 1742, CERCA, Nestore Horizon2020 SC1-PM-15-2017 (n 769643), *ICREA Academia 2014* and Grant 20141510 (Marató TV3). The founders had no role in the study design, data collection, analysis, and preparation of the manuscript. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of several Titan Xp GPU used for this research. The collected data as part of the study and given labels is publicly available from the research group's website: http://www.ub.edu/cvub/dataset/

References

[1] M. N. Laska, M. O. Hearst, K. Lust, L. A. Lytle, and M. Story, "How we eat what we eat: identifying meal routines and practices most strongly associated with healthy and unhealthy dietary factors among young adults," *Public health nutrition*, vol. 18, no. 12, pp. 2135–2145, 2015.

[2] P. M. Stalonas and D. S. Kirschenbaum, "Behavioral treatments for obesity: Eating habits revisited," *Behavior Therapy*, vol. 16, no. 1, pp. 1–14, 1985.

[3] J. B. Hopkinson, D. N. Wright, J. W. McDonald, and J. L. Corner, "The prevalence of concern about weight loss and change in eating habits in people with advanced cancer," *Journal of pain and symptom management*, vol. 32, no. 4, pp. 322–331, 2006.

[4] L. M. Donini, C. Savina, and C. Cannella, "Eating habits and appetite control in the elderly: the anorexia of aging," *International psychogeriatrics*, vol. 15, no. 1, pp. 73–87, 2003.

[5] A. Tal and B. Wansink, "Fattening Fasting: Hungry Grocery Shoppers Buy More Calories, Not More Food," *JAMA Intern Med.*, vol. 173, no. 12, pp. 1146–1148, 2013.

[6] S. Higgs and J. Thomas, "Social influences on eating," *Current Opinion in Behavioral Sciences*, vol. 9, pp. 1–6, 2016.

[7] E. Kemps, M. Tiggemann, and S. Hollitt, "Exposure to television food advertising primes food-related cognitions and triggers motivation to eat," *Psychology & Health*, vol. 29, no. 10, p. 1192, 2014.

[8] W. B. S. C. Ren A de Wijk, Ilse A Polet and J. H. Bult, "Food aroma affects bite size," *BioMed Central*, pp. 1–3, 2012.

[9] N. Larson, M. Story, and M. J, "A review of environmental influences on food choices," *Annals of Behavioural Medicine*, vol. 38, pp. 56–73, 2009.

[10] J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: a novel wearable device for monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1772–1779, 2014.

[11] D. Ravì, B. Lo, and G.-Z. Yang, "Real-time food intake classification and energy expenditure estimation on a mobile device," *Wearable and Implantable Body Sensor Networks (BSN), 2015 IEEE 12th International Conference on*, pp. 1–6, 2015.

[12] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G.-Z. Yang, "An intelligent food-intake monitoring system using wearable sensors," *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*, pp. 154–160, 2012.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[14] A. Quattoni and A. Torralba, "Recognizing indoor scenes." *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 413–420, 2009.

[15] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] D. D. Lewis, "Reuters-21578," *Test Collections 1*, 1987.

[18] J. Garofolo and et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Philadelphia: Linguistic Data Consortium*, 1993.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICRL)*, pp. 1–14, 2015.

[24] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[25] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," *Advances in Neural Information Processing Systems 27*, pp. 487–495, 2014.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An Image Database for Deep Scene Understanding," *ArXiv*, pp. 1–12, 2016.

[27] M. Koskela and J. Laaksonen, "Convolutional network features for scene recognition," *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1169–1172, 2014.

[28] L. Zheng, S. Wang, F. He, and Q. Tian, "Seeing the big picture: Deep embedding with contextual evidences," *CoRR*, vol. abs/1406.0132, 2014.

[29] L. Wang, Z. Wang, and W. Du, "Object-Scene Convolutional Neural Networks for Event Recognition in Images," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6, 2015.

[30] L. Herranz, S. Jiang, and X. Li, “Scene Recognition With CNNs: Objects, Scales and Dataset Bias,” *Conference on Computer Vision and Pattern Recognition*, pp. 571–579, 2016.

[31] A. Furnari, G. M. Farinella, and S. Battiato, “Temporal segmentation of egocentric videos to highlight personal locations of interest,” *European Conference on Computer Vision*, pp. 474–489, 2016.

[32] A. Furnari, G. Farinella, and S. Battiato, “Recognizing Personal Locations From Egocentric Videos,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 1–13, 2017.

[33] 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.

[34] 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*, vol. 40, no. 4, pp. 834–848, 2018.

[35] O. Pujol, P. Radeva, and J. Vitrià, “Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1007–1012, 2006.

[36] 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*, vol. 40, pp. 1452–1464, 2018.

[37] E. Talavera, M. Dimiccoli, M. Bolanos, M. Aghaei, and P. Radeva, “R-clustering for egocentric video segmentation,” *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 327–336, 2015.

[38] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva, “Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation,” *Computer Vision and Image Understanding*, vol. 155, pp. 55–69, 2017.

[39] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[42] T. K. Ho, “Random decision forests,” *Proceedings of the Third International Conference on Document Analysis and Recognition Vol.1*, pp. 278–282, 1995.

[43] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[44] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” pp. 675–678, 2014.

Estefania Talavera received her BSc degree in electronic engineering from Balearic Islands University in 2012 and her MSc degree in biomedical engineering from Polytechnic University of Catalonia in 2014. She is currently a PhD student at the University of Barcelona and University of Groningen. Her research interests are lifelogging and health applications.

Mara Leyva received her BSc degree in Software Engineering from the University of Oviedo in 2011. She continued her studies with the MSc in Artificial Intelligence in the Polytechnic University of Catalonia, graduating in 2017. She is currently a PhD Student at the Johann Bernoulli Institute for Mathematics and Computer Science in the University of Groningen.

Md. Mostafa Kamal Sarker Md, Mostafa Kamal Sarker is a PhD student and pre-doctoral researcher at the Intelligent Robotics and Computer Vision group, Department of Computer Engineering and Mathematics Security, Rovira i Virgili Univeirsity, since, September 2016. He received his B.S. degree from Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 2009, and his M.S. degree from Chonbuk National University, Jeonju, South Korea, in 2013 supported by the Korean government Brain Korea21 (BK21) scholarship program. From September 2013 to August 2016, he worked as a researcher on a project from National Research Foundation of Korea (NRF) that is funded by the Ministry of Education of South Korea. His research interests include the areas of image processing, pattern recognition, computer vision, machine learning, deep learning, egocentric vision and visual lifelogging.

Domenec Puig received the M.S. and Ph.D. degrees in computer science from Polytechnic University of Catalonia, Barcelona, Spain, in 1992 and 2004 respectively. In 1992, he joined the Department of Computer Science and Mathematics at Rovira i Virgili University, Tarragona, Spain, where he is currently Associate Professor. Since July 2006, he is the Head of the Intelligent Robotics and Computer Vision group at the same university. His research interests include image processing, texture analysis, perceptual models for image analysis, scene analysis, and mobile robotics.

Prof. Nicolai Petkov received the Dr. sc.techn. degree in Computer Engineering (Informationstechnik) from the Dresden University of Technology, Dresden, Germany. He is the Professor of Computer Science and Head of the Intelligent Systems group of the Johann Bernoulli Institute of Mathematics and Computer Science of the University of Groningen, the Netherlands. He is the author of two monographs and coauthor of another book on parallel computing, holds four patents, and has authored over 100 scientific papers. His current research is in image processing, computer vision and pattern recognition, and includes computer simulations of the visual system of the brain, brain-inspired computing, computer applications in health care and life sciences, and creating computer programs for artistic expression. Prof. Dr. Petkov is a member of the editorial boards of several journals.

Dr. Petia Radeva is a senior researcher and associate professor at the University of Barcelona. She is Head of Computer Vision at University of Barcelona group and the MiLab of Computer Vision Center. Her present research interests are on development of learning-based approaches for computer vision, egocentric vision and medical imaging.