Abstract—Mining Web data to create AI-usable datasets, is still non-trivial. Unfortunately, despite the free data access, the formation of a dataset useful for machine learning applications cannot rely solely on a data mining phase. For any given query, the retrieved sample may include duplicated, misclassified or completely irrelevant content. The consequence of not “cleaning” those datasets is to end up with faulty, noisy and imbalanced datasets. Thus, curation is necessary, to tackle the variable degrees of inconsistency found on the retrieved samples. This paper suggests a pipeline consisting of state-of-the-art and off-the-shelf methods for curating an image dataset retrieved from the Web. As a case study, the pipeline is applied on expanding food datasets with currently un categorized Greek dishes, leveraging information found in a specialized ontology, aiming at increasing the accuracy in food recognition applications.

Keywords—Web mining, Image matching, Image classification

I. INTRODUCTION

Using the Web for data collection has, recently, become increasingly relevant. Recent advances in image-based Artificial Intelligence (AI) imply, amidst other things, the more the image data the better the performance. The computational ability, besides the practical need, to handle large datasets establishes a constant demand for more. The Web is readily accessible for such endeavors yet constructing a formal AI-useful dataset out of it, is still a daunting procedure. Attempting to manually gather data, such as images, is cost ineffective, time consuming and prone to errors. In spite of that, automatic data acquisition from the Web is not a trivial task, since the lack of a sophisticated methodology is expected to lead to a noisy, imbalanced and even error prone assortment of data.

Automatic data collection frameworks face a wide range of challenges, like (a) irrelevant content removal, (b) proper data class assignment, (c) composition of suitable queries and (d) rectification of corrupted data. In addition, common concerns exist in dataset formalization procedures, including (a) balancing of samples per class, (b) dealing with conflicting content, (c) constructing and leveraging inclusion or mutual exclusion strategies and (d) handling missing data. These are all issues to be considered during the creation of benchmark datasets [1][2], landmark recognition[3], as well as in more generic ones[4][5].

Respectively, food recognition received an increasing interest in the recent years [7], due to its wide span of applications in health care [6], habit-tracking [9], and tourism [8]. Food computing [10] refers to the automatic identification of food dishes or naming the exhibited ingredients, calculating the calorie and nutritional values, as well as producing warnings of potential health risk factors from images. In computer vision terminology these tasks correspond to food image retrieval, food segmentation, recipe generation, and dish classification. For this reason, there exist several food related datasets in the literature [11]. Some focus on specific assignments (e.g., segmentation [12]), while others apply to broader tasks (e.g., food retrieval [13]). A common topic among them is that they include categories of dishes from local cuisines or regional diets. As reported in TABLE I, a portion of the Mediterranean diet, such as dishes found in the traditional Greek cuisine, is currently absent from existing food datasets.

The importance of including Greek dishes to existing datasets is based partly on its cultural significance (and relation to the general Eastern Mediterranean cuisine) but also on its application significance since these new data challenge the state-of-the-art recognition methods. For instance, isomorphic food dishes (culinary or culturally assigned to different classes) are separable only by subtle differences. The challenges of constructing a dataset of food images for dish recognition are discussed later in detail. Thus, the contribution of this work is threefold: (a) it expands existing food datasets

| Table I. Cross-referencing labels of Greek dishes with the existing food datasets. |
|-----------------|-----------------|-----------------|
| Dataset         | Type            | Images/Classes  |
| ETH Food-101 [15] | Western         | 101K/101        |
| UPMC Food-101 [20] | Western         | 90.8K/101       |
| Food524DB [36] | Misc.           | 3.9K/50         |
| FoodX-251 [23] | Misc.           | 158.8K/251      |
| AHA Food-200 [32] | Misc.           | 197.3K/200      |
| AHA Food-500 [11] | Misc.           | 400K/500        |
| GREFood         | Greek           | 90.7K/165       |
with 165 Mediterranean diet Greek dishes, (b) it proposes a sophisticated pipeline for curating images acquired automatically from the Web and (c) it introduces a challenging dataset. This paper proceeds by reviewing the related literature, then by presenting the novel curation pipeline, and, subsequently, by showcasing experimental results and overall conclusions.

II. RELATED WORK

The work most closely related to ours, is the one which introduced the ISIA Food-500 [11]. That dataset is publicly available and contains around 400K images not equally distributed in 500 categories of diverse cuisine origins. The authors followed a similar approach to ours to enrich their dataset i.e., they included images from the Web and removed duplicate and non-food images. Further data inspection was applied with crowdsourcing. However, the authors weren’t specific about the details of the data acquisition procedure. It is not clear how strict the rule about removing duplicates was (i.e., allowing the inclusion of geometrically transformed images, as similar but not duplicates), and how they dealt with the assignment of labels to conflicting dish classes (food co-occurrence).

Non-automatic image acquisition procedures for food datasets are divided in in-situ acquisition [14] and browsing for them on the Web [15]. Typically, in-situ procedures involve an expensive overhead prior to analyzing the data and constructing the actual dataset (i.e., setting-up a camera, removing obstacles from the scene, avoiding capturing humans and human parts to preserve their anonymity, and finally composing the frame) [16][17][19]. An advantage of those manual methods is that the samples have preassigned labels. On the other hand, Web browsing manual procedures involve: (a) composing a list of food classes; (b) browsing the Web for representative samples; (c) downloading the images; and (d) cleaning-up errors. A disadvantage related to this type of dataset construction, is that humans have limited capability in processing a large number of queries simultaneously (e.g., limited attention span) and their effort is time consuming and prone to mistakes.

Automatic frameworks for food dataset construction are discussed in [18][20]. This approach includes the following: (a) an acquisition protocol; (b) querying search engines; (c) downloading the retrieved content; (d) identifying non-related images for removal; (e) de-duplicating of images; and (f) correcting mistakenly assigned samples to wrong classes. Finding content in such uncontrolled environment might be faster than other approaches, but introduces additional challenges i.e., the sample per class distribution might not be the desired one and food co-occurrence class assignment is often disregarded. Food co-occurrence is discussed in [21], where various image descriptors were used to extract local and global features to recognize multiple-food photos considering co-occurrence statistics. Later, the same authors employed a manifold learning approach to improve their results [22].

Finally, other frameworks (not explicitly related to food) for image acquisition from the Web are discussed in [4][5], all reported similar challenges to those previously described.

III. PROPOSED PIPELINE

The Web is a valuable resource, which can be exploited in order to get large collections of data. Data repositories, content providers and search engines can be utilized for such endeavors. However, collecting data with no treatment at all, can easily become futile, especially for practical AI applications. To meet this challenge, we propose a pipeline for the task of creating an AI-usable dataset with data acquired from the Web. This pipeline consists of several modules, which collaborate to construct a balanced thematic dataset.

The proposed pipeline is graphically shown in Fig. 1. An ontology is used to query the web for data. The images collected from the web are processed via a module to filter out duplicate and irrelevant content. The valid ones form candidate pools in respect to their class. In case a candidate pool doesn’t have the desired minimum number of samples, it is merged with the semantically closest class according to the information provided by the ontology. Additionally, not all of these images will fit into their respective final class, since misclassified samples are expected to have been collected in the first place. To assign images into classes we assume intra-class visual coherence, thus a ranking based on visual similarity is used. Thus, to assign a candidate image to its respective class an aggregated ranking of its intra-class visual similarity and its position in the retrieved results is calculated. Finally, the top candidate images within the range of the minimum number of desired samples are harnessed.

More analytically, the framework upon the dataset construction has to be defined, namely the aim of the dataset, its requirements, and the expectations. In this case study a food centric dataset of real-world images with Greek food dishes had to be constructed. We chose this topic because food images are particularly intriguing in terms of image analysis, as there is an expected large intra-class variations in global appearance, there is the challenge of food co-occurrence, there can be subtle discriminative details amongst classes, not to mention that food configuration is not standard.

The source of the dishes’ labels, basically, has to meet two criteria: (a) to be a credible source; and (b) to be described in a formal ontological representation. The latter is useful since it provides information about the relations among the dishes. As stated, the source of the images was selected to be the Web, and more than one popular search engines can be used as image resources. It is reminded that search engines provide indexed content according to a query, in this case a food label. Again, as stated in the introductory material, it is expected that undesired content will appear in the retrieved results, including duplicate images and irrelevant (non-food) images.
These images if included in a dataset designed for the purpose of recognition might have a negative effect, so they have to be excluded (rejection modules in Fig. 1). Furthermore, as duplicate images we define: (a) exact copies; and (b) the images that are a geometrically transformed copy of another. This definition and the related policy is followed throughout the data collection procedure and no duplicate images are allowed within the class and across the dataset. As far as it concerns the dataset balancing, the pipeline has to be able to adjust the number of images per class without sacrificing the label assignment accuracy. This is influenced, mainly, by two factors: (a) the number of retrieved results from the Web, in which some labels might be too specific to enable the discovery of rich content, and (b) the no-duplicates policy. If some images presented ambiguous (or mixed) content and were retrieved for more than one label, they were assigned to the class that acquired them first. For example, an image presenting *fava puree with grilled octopus*, could fit either in the *grilled octopus* or in the *fava puree* labels; by default, it is assigned to the first label it has been retrieved for. Both of these factors result in a decreased number of candidate samples for some classes and ultimately in an imbalanced dataset.

Since the relations between dishes is revealed by the employed ontology, a pruning or class-merging procedure can ameliorate this, without giving up the accuracy in label assignment (pruning module in Fig. 1). Notwithstanding, when seeking for a large number of images on the Web, some retrieved samples can be misclassified from the start. To deal with this we assume within class resemblance, and, hence, a ranking system based on visual similarity, or other informative criteria, to amend for excluding misclassified retrieved samples from the candidate pool (ranking aggregation module in Fig. 1). Finally, the desired number of samples per class can be adjustable; setting the minimum to 550 samples per class followed by a maximin approach to maintain as many images as possible.

A. Knowledge base

Recently, a trilingual thesaurus of food served in restaurants in Eastern Macedonia and Thrace (North-Eastern Greece) has been developed, called “AMALTHEA” (Amalthea) [24]. The authors designed and implemented a Web lexicographic environment, which accommodates information retrieved from restaurant menus. This infrastructure enabled the development of a thesaurus with information about dishes, their ingredients, recipes, concepts related to food, as well as dietary and cultural.

To incorporate this knowledge base (Amalthea) to the presented pipeline a tree-like structure was implemented. A traversal mechanism granted the dish labels, their parent nodes and their corresponding ID codes. The dish labels were used as the queries to be provided to the Web search engines. We chose not to look for complex relations between dishes within the knowledge base even though label translations, recipes, related ingredients, and dish preparation procedures were available.

B. Food/NonFood classification

To reject irrelevant content from the candidate pool of images, a food/nonFood discriminator was trained via transfer learning. A binary class dataset with 200K colored images was constructed. The one class included 100K images from the ETH-Food101, while the other was a meticulous manual selection (with duplicate exclusion) of non-food images from benchmark datasets [5]. The choice of ETH-Food101 was based on its content, which it was visually aligned to the criteria of desired images.

The chosen deep neural network model was the VGG-16 [25] pretrained on the ImageNet. The accuracy on the test set was 98.59%. This model was used as the food/NonFood discrimination module in the proposed pipeline.

C. Removing duplicate images

To reject duplicate images according to the criteria of undesired content, a deduplication procedure was developed to handle this task. Deduplication is used widely in many applications (forensics, closed-loop systems, etc.). Several approaches have been developed to deal with this task, including, but not limited to, relying on textbook image descriptors [29][30], as well as bespoke deep neural networks [31][37][38].

Typically, a feature extractor combined with a distance measure, compare any two images to find out if they are similar or not. The application of a threshold \( \theta \) defines the degree of similarity, which in this case reduces to being a duplicate or not. A reliable method is needed to (a) bring closer those images that are either exactly the same or geometrically transformed copies of one another, while at the same time to (b) push further away those, which have just a visual resemblance or are completely different (Fig. 2). Moreover, as the purpose of this pipeline is to construct a large-scale dataset, the deduplication method must scale well, as more and more data are collected into the candidate pool. Otherwise, the computational complexity will reach an impractical \( O(n^2) \).

That said, image descriptors coupled with locality-sensitive hashing methods [26] offer the benefits of being fast feature extractors and reliable at detecting most affine transformations. At the same time, they keep the description of an image compact, which is computationally important for
the comparison. In addition, Hamming distance fits nicely as a metric for these descriptors. To tackle the computational complexity a BK-tree structure [28] is exploited, which indexes the binary hashes of the candidate images. Given a query hash and a threshold $\theta$, this structure is used for the purpose of being faster in finding the hashes under that threshold. The complexity of this algorithm is approximately $O(\log n)$.

The deduplication module in the proposed pipeline employs three image hashing methods, namely (a) the average image hash, (b) the perceptual hash and (c) the difference hash. An exclusive BK-tree is constructed for the three descriptors respectively. Each time an image is encountered its features are extracted and then compared with all candidate image hashes within the tree structures. A majority voting approach decides whether an image is duplicate or not; no tie breaker is needed since there are three voters with equal voting weight. The query hash is stored within its respective BK-tree, if and only if there was no duplicate decision outcome (also the query image is added to the candidate pool).

D. Pruning candidate classes

The Web is an uncontrolled environment for the task of collecting representative images in respect to some label. It is expected for some classes to be overpopulated, whereas others be exactly the opposite. In the overpopulated class scenario, the reasonable thing to do is to keep the $K$ most similar images based on some ranking. In contrast, underpopulated classes need a closer inspection to decide upon an action plan, since greedy approaches might result data losses. Usually, this issue is addressed with (a) seeking for more content, (b) discarding underqualified classes and (c) merging similar classes. Automating this procedure is not always trivial.

To deal with this shortcoming this pipeline exploits the aforementioned knowledge base. When a class’s candidate pool doesn’t meet the completion criterion, then it is merged with a pertinent class within the same super-class group (children-nodes under the same parent) according to the Greek food thesaurus. If no other class exists within a certain group, then the underpopulated one is merged with its parent. This pruning procedure is followed until all classes meet the completion criterion.

E. Class assignment based on aggregated rankings

Another shortcoming related to the uncontrolled environment of the Web, is that, given a query, no perfect retrieved pool of samples is expected to emerge. Actually, it is anticipated for some content to be misclassified. Duplicate and non-food retrieved images are discarded already by the respective modules within the pipeline. Thus, the misclassified results contain only food samples fitted mainly in two categories: either they reside to a different class, or they belong to the desired class but only to some small degree.

To assign candidate images to a certain class an aggregation of two rankings is employed in the proposed pipeline. Under the assumption that search engines function as retrieval systems, which provide the results in a relevance order, the first ranking system emerges from their retrieved arrangement. The second ranking system arises from the visual similarity amongst the samples of a candidate pool. For this purpose, a rather simple yet effective image descriptor is devised, which is somehow close to the concept of the dominant color descriptor of the MPEG-7 standard [39], yet is simple and rather effective.

The descriptor by design considers color similarity between images, whilst is invariant to geometrical transformations. It analyzes colored images by matching the actual colors to a number of $h$ hues and $v$ intensity values in the HSV color space. This results in a quantized image with $q = h + v$ channels. From the quantized image two further quantities are calculated per channel: (a) the coordinate of the center of mass and (b) the percentage coverage. Subsequently, these values are normalized within the range [0,1]. At the end, the normalized quantities are put together in a vector forming the final signature. We term this descriptor as Color Gravity Descriptor (CGD), due to its inherent characteristic to form color discs with specific mass and capture the attractions among them.

The procedure of ordering the candidate pool by its visual similarity is as follows: (a) extract the CGD signatures; (c) calculate a single cluster with Chebyshev distance; (d) calculate the mean $\mu$ and standard deviation $\sigma$ of that single cluster; and (e) use the $\mu$ and $\sigma$ to arrange the samples by their distance from the center of the cluster in a weighted manner.

As it happens, the two ranking systems have different arranging criteria. Thus, a method is needed to produce an optimal ranking that maximizes some sort of agreement. The approach followed in this pipeline is to calculate the pairwise disagreement $\tau$ measured with Kendall Tau and then maximize the following loss function[33]:

$$\arg \max_{\tau} \sum d(\bar{\tau}, \tau) \leq \sum d(\bar{\tau}, \tau) \quad \forall \tau \in N \quad (1)$$

This aggregation module can combine any ranking system, but a naive implementation can lead in trying to solve an NP-hard problem [34]. Approximations using weighted graphs have been proposed as well [35].

IV. EXPERIMENTAL VERIFICATION

A. Setting up the pipeline’s parameters

The proposed pipeline consists of several modules, which need certain parameters to work as expected i.e., the desired number $K$ of samples per class, and the threshold $\theta$ per method which separates duplicates from the rest.

To find the optimal threshold for the image hashing algorithms, an experiment is conducted similar to the one proposed by Ke [27]. A random subset is taken comprised of 2000 unique food images from the ETH Food-101 dataset. For each image the following 40 transformations are applied: (a) colorization, (b) contrast adjustment, (c) cropping, (d) despeckling, (e) downsampling, (f) mirroring (flipping), (g) change of format, (h) framing, (i) rotation, (j) scaling, (k) saturation enhancement and (l) intensity adjustment. These transformations result in 82000 $^1$ images, of which the signatures are extracted respectively with each image hashing method. Distances and confusion matrices are calculated for all images. The desired threshold $\theta$ is the one that minimizes the following function (see also Fig. 3):

$$\arg \min_{\theta} |FN - FP| \quad (2)$$

$^1$ Note, the parameters for each transformation are the same as suggested in [27]
To test the thresholds, another sample of 20000 images is taken from the same dataset with mutual exclusion in respect to the training one. Then, 500 of them are randomly selected and undergo the aforementioned transformations. This results in a test set of 40000 images. The achieved cross-validated accuracy for $\theta^{(d)} = 3$ (average image hash), $\theta^{(d)} = 14$ (difference hash) and $\theta^{(d)} = 14$ are 85.57%, 86.72% and 86.5% respectively.

B. GREFood - data acquisition

From the Amalthea ontology 361 labels of dishes served in Greek restaurants were extracted. The pipeline queried two search engines (Google and DuckDuckGo), in order to populate these classes with candidate samples. The food/nonFood and deduplication modules rejected initially all the undesired content. A total of 401231 images were collected from the Web. In total, 128K were rejected as duplicate (32%) and 107K as non-food (26.8%). Note that, the deduplication module was not applied on the non-food images. Since the purpose of this pipeline is to produce a balanced dataset, the need for the pruning and the class assignment modules is evident in Fig. 4, which shows many underpopulated classes within a group of semantically similar. After pruning the classes according to the procedure described previously, there were 169 classes left. The pruning module by design merges underpopulated classes with other relevant classes in the same group or their parent nodes; this is repeated indefinitely until the completeness criterion is met. We chose to not let this happen in this work and we parameterized the module to stop the merging procedure when a class was merged with its parent. Thus, underdeveloped classes were removed completely (4 classes).

The final module of this pipeline is the one which assigns the candidate samples into their respective classes by aggregating a visual similarity and an indexing ranking system. Therefore, 165 classes were kept that met the dataset completion criterion (in this case 550 images per class), as shown in Fig. 5.

C. Food recognition

A visual example of the GREFood dataset, in which the samples were collected and curated by the proposed pipeline is shown in Fig. 6. To test its fitness for dish recognition, a classification experiment was conducted using the vanilla VGG-16 architecture pretrained on the ImageNet.

The GREFood dataset was split into 8-folds from which training and testing instances were produced. Affine transformations were applied on each training set. To avoid overfitting, regularization mechanisms are employed i.e., Dropout and batch normalization. The performance of classification on the GREFood validation sets is 64.8(±7.27)% The just above the average accuracy score indicates the classifier struggled to perform remarkably. Food co-occurrence and between classes’ similarity are probable complications. However, classification tasks are attainable, in this dataset which has been constructed in a fully automatic manner, with no human innervation whatsoever.

V. CONCLUSIONS AND FUTURE WORK

A pipeline for constructing a balanced thematic dataset with data acquired from the Web is proposed in this paper. The pipeline uses state-of-the-art and off-the-shelf methods to find, collect and clean data with large intra-class variations. A food dataset was constructed (GREFood). Limitations related to the proposed approach fall into two categories mainly. The first relates to the error carryover from the incorporated modules. The other concerns to the decisions related to the pipeline’s policies like the deduplication module which strictly prohibits duplicate images to exist even in different classes and the rejected images by the assignment module are not being re-used. External factors affecting the performance of this pipeline come from the content availability on the Web and the relations between dishes mined from the Amalthea ontology. We consider improvements are possible. The deduplication strategy can be examined empirically.
RESEARCH Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code:T1EDK-02015). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for the experiments of this research.

ACKNOWLEDGMENTS

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code:T1EDK-02015). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for the experiments of this research.

REFERENCES

[1] NG, Hong-Wei; WINKLER, Stefan. A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). IEEE. 2014. p. 343-347.

[2] Rotte, R., Timotheo, R., & Van Gool, L. (2015). Dec: Deep expectation of apparent age from a single image. In Proceedings of the IEEE international conference on computer vision workshops (pp. 10-15).

[3] Weyand, T., Araujo, A., Cao, B., & Sim, J. (2020). Efficient near-duplicate detection and sub-image retrieval. In ACM multimedia (Vol. 4, No. 1, p. 5).

[4] Križhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

[5] Križhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[6] Christoudoulidis, S., Anthimopoulos, M., & Mougiakakou, S. (2015, September). Food recognition for dietary assessment using deep convolutional neural networks. In International Conference on Image Analysis and Processing (pp. 458-465). Springer, Cham.

[7] Van Dorst, B., Mehta, J., Bekaert, K., Rouah-Martin, E., De Coen, W., Dubrue, P., ... & Robbens, J. (2010). Recent advances in recognition elements of food and environmental biosensors: a review. Biosensors and Bioelectronics, 25(4), 1178-1194.

[8] Pavlidis, G., & Markantonatou, S. (2020). Gastronomic tourism in Greece and beyond: A thorough review. International Journal of Gastronomy and Food Science, 100229.

[9] Aizawa, K., & Ogawa, M. (2015). Foodlog: Multimedia tool for healthcare applications. IEEE Multimedia, 22(2), 4-8.

[10] Min, W., Jiang, S., Liu, L., Rui, Y., & Jain, R. (2019). A survey on food computing. ACM Computing Surveys (CSUR), 52(5), 1-36.

[11] Min, W., Liu, L., Wang, Z., Luo, Z., Wei, X., Wei, X., & Jiang, S. (2020, October). Isfa food-500: A dataset for large-scale food recognition via stacked global-local attention network. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 393-401).

[12] He, Y., Xie, C., Khamma, N., Boushey, C. J., & Delp, E. J. (2013, July). Food image analysis: Segmentation, identification and weight estimation. In 2013 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[13] Farinella, G. M., Allegra, D., Molisanti, M., Stanco, F., & Battisti, S. (2016). Retrieval and classification of food images. Computers in biology and medicine, 77, 23-39.

[14] Chen, M., Dhinag, K., Wu, W., Yang, L., Sukthankar, R., & Yang, J. (2009, November). PFID: Pittsburgh fast-food image dataset. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 289-292). IEEE.

[15] Bossard, L., Guillaumin, M., & Van Gool, L. (2014, September). Food–101–mining discriminative components with random forests. In European conference on computer vision (pp. 446-461). Springer, Cham.

[16] Joutou, T., & Yanai, K. (2009, November). A food image recognition system with multiple kernel learning. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 285-288). IEEE.

[17] Hoashi, H., Joutou, T., & Yanai, K. (2010, December). Image recognition of 85 food categories by feature fusion. In 2010 IEEE International Symposium on Multimedia (pp. 296-301). IEEE.

[18] Matsuda, Y., & Yanai, K. (2012, November). Multiple-food recognition considering co-occurrence employing manifold ranking. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 2017-2020). IEEE.

[19] Kawano, Y., & Yanai, K. (2014, September). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In European Conference on Computer Vision (pp. 3-17). Springer, Cham.

[20] Wang, X., Kumar, D., Thome, N., Cord, M., & Precioso, F. (2015, June). Recipe recognition with large multimodal food dataset. In 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (p. 1-6). IEEE.

[21] Matsuda, Y., Hoashi, H., & Yanai, K. (2012, July). Recognition of multiple-food images by detecting candidate regions. In 2012 IEEE International Conference on Multimedia and Expo (pp. 25-30). IEEE.

[22] Matsuda, Y., & Yanai, K. (2012, November). Multiple-food recognition considering co-occurrence employing manifold ranking. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) (pp. 2017-2020). IEEE.

[23] Kaur, P., Sikka, K., Wang, W., Belongie, S., & Divakaran, A. (2019). Foodx251: A dataset for fine-grained food classification. arXiv preprint arXiv:1907.06167.

[24] Markantonatou, S., Toraki, Z., Minos, P., Vagalopoulou, A., Stamou, V., & Pavlidis, G. (2021). AMAAΘEIA: A Dish-Driven Ontology in the Food Domain. Data, 6(4), 41.

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

[26] Chi, L., & Zhu, X. (2017). Hashing techniques: A survey and taxonomy. ACM Computing Surveys (CSUR), 50(1), 1-36.

[27] Ke, Y., Sukthankar, R., Huston, L., Ke, Y., & Sukthankar, R. (2004, October). Efficient near-duplicate detection and sub-image retrieval. In ACM multimedia (Vol. 4, No. 1, p. 5).

[28] Burkhard, W. A., & Keller, R. M. (1973). Some approaches to best-match file searching. Communications of the ACM, 16(4), 230-236.

[29] Sikora, T. (2001). The MPEG-7 visual standard for content description—an overview. IEEE Transactions on circuits and systems for video technology, 11(6), 696-702.

[30] Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814.

[31] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Phlbinn, J., ... & Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1386-1393).

[32] Min, W., Liu, L., Luo, Z., & Jiang, S. (2019, October). Ingredient-guided cascaded multi-attention network for food recognition. In Proceedings of the 27th ACM International Conference on Information Retrieval on Multimedia (pp. 1331-1339).

[33] Kemeny, J. G. (1959). Mathematics without numbers. Daedalus, 88(4), 577-591.

[34] Bartholdi, J. J., Tovey, C. A., & Trick, M. A. (1989). The computational difficulty of manipulating an election. Social Choice and Welfare, 6(3), 227-241.

[35] Conitzer, V., Davenport, A., & Kalagnanam, J. (2006, July). Improved bounds for computing Kemeny rankings. In AAI (Vol. 6, pp. 620-626).

[36] Ciocca, G., Napolitano, P., & Schettini, R. (2017, September). Learning CNN-based features for retrieval of food images. In International Conference on Image Analysis and Processing (pp. 426-434). Springer, Cham.

[37] Ioannis Kansizoglou, Nicholas Santavas, Loukas Bampis and Antonios Gasteratos, “HASeparator: Hyperplane-Assisted Softmax”, 19th IEEE International Conference on Machine Learning and Applications (ICMLA2020), 14-17 December 2020, Miami, Florida, USA, (Virtual Conference).

[38] Ioannis Kansizoglou, Loukas Bampis and Antonios Gasteratos, “Deep Feature Space: a geometrical perspective”, IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear.

[39] Chang, S. F., Sikora, T., & Purl, A. (2001). Overview of the MPEG-7 standard. IEEE Transactions on circuits and systems for video technology, 11(6), 688-695.