UNIVERSITÀ DEGLI STUDI DI CATANIA
Dipartimento di Matematica e Informatica
Dottorato di Ricerca in Matematica e Informatica XXX Ciclo

Dario Allegra

Food Understanding from Digital Images

Tesi di Dottorato di Ricerca

Prof. Filippo Stanco

Anno Accademico 2016 - 2017

“*Someday, someone will best me. But it won’t be today, and it won’t be you.*”

“Last Word”, *Magic: The Gathering.*

Abstract

In the last decade food understanding from digital media has become a challenge with applications in many different domains. On the other hand, food is a crucial part of human life and what people eat strongly affects their health and characterize their identity. For this reason food plays a key role in world economy. The focus of my Ph.D. thesis is the study of food image understanding from the perspective of Computer Vision and Machine Learning. As first original scientific contribution I propose an approach to perform discrimination between food VS non-Food images. For this study I adopted the One-Class Classification paradigm which allows to build a binary classifier by performing learning from the samples of one class only. Specifically, a One-Class Support Vector Machine is trained by employing food images. The second contribution of my work is related to food retrieval and classification. Since food is intrinsically deformable and presents high variability in appearance, this task is very challenging and requires an in-depth study of images representation. To this aim I propose a new representation model related to the neuroscientific notion of Anti-Texton. The third problem considered in this thesis is about food volume and carbohydrates estimation. Last but not least, new food datasets has been introduced for the scientific community. To address Food VS Non-Food problem, three datasets has been employed: the public UNICT-FD889 food dataset and two new datasets downloaded by Flickr. The latter two, include respectively 4805 food images and 8005 non-food ones. Moreover, UNICT-FD889 has been extended from 889 to 1200 classes and annotated across 8 groups: Appetizer, Main Course, Second Course, Single Course, Side Dish,Dessert, Breakfast, Fruit. To evaluate volume estimation performance a novel dataset of 80 different plates has been built. This dataset includes RGB images, as well as depth map and 3D models. In thesis appendices, I report the other works produced during my Ph.D studies and also a comprehensive discussion on Cultural Heritage preservation and exploitation through modern technologies. These works mainly focus on 3D model reconstructions, semantic annotation platforms and virtual unrolling of papyri.

Acknowledgements

I would like to take this opportunity to thank Professor Giovanni Maria Farinella for his guidance, his enthusiasm and invaluable suggestions he has made me during my Ph.D. pursuit. Also because of him, I have been able to achieve the results described in this thesis. I am very grateful for the time he devoted to me.

My thanks also go to PD. Dr. Stavroula Mougiakakou, who has giving me the opportunity to conduct part of the studies reported this thesis in her lab at the ARTORG Center at University of Bern. The months I have spent in her group has made my Ph.D. experience productive and stimulating.

Contents

Chapter 1

Introduction

1.1 Motivation

It is well-known that food plays a fundamental role in people life and global economic. Eating choice are strongly correlated to people culture, financial situation and, most of all, health conditions [1]. Diseases like obesity may even influence the economic situation of a country, because of the direct medical costs, productivity costs, transportation costs and human capital cost [2]. Furthermore, allergic diseases can make very risky the food intake. For these reasons, the current imaging technologies (e.g., smartphones and wearable devices) are taking a dominant role in the food intake monitoring (Fig.s 1.1(a) and 1.1(b)). They make possible the building of automatic and accurate system to assess people's diet and, consequently, to raise society's awareness among the quality of life.

The ability to automatically detect images which depicts food and then the recognition of it, is fundamental to assist people during their daily meals. Automatic food image retrieval and classification could replace the inaccurate manual dietary assessment, that is based on self-reporting. Hence, food understanding engines embedded in mobile devices can be used to create food-logs that help the experts (e.g., nutritionists, psychologists) to understand the behaviour, habits and/or eating disorders of people, especially the ones affect by chronic diet-related diseases. All these critical aspects of dietary habits and food intake, have motivated this thesis. The following section provides a more detailed discussion about the fundamental role of food in human life and the relevant contribution of Computer Vision and Machine Learning technologies to address food-related problems.

(a) (b)

Figure 1.1: (a) An example of wearable camera; (b) a mobile app for automatic diet monitoring.

1.1.1 Food Impact on Human life

People often ignore the impact that food has in their life. They do not have time to take into account the eating healthy and nutritional food. Hence, usually they prefer fast food or a snack loaded with sugar instead of a regular nutritional meal. Unfortunately, an inadequate nutrition is one of the main cause of many chronic diseases such as obesity, diabetes, cancer, osteoporosis, dental diseases and cardiovascular ones [3, 1]. In early 2002, a Joint WHO/FAO Expert Consultation recognized that the growing epidemic of chronic disease that afflicts most of the countries in the world is correlated to dietary and lifestyle changes. Although standards of living have improved, food availability raised and become more diversified, there have also been serious negative effects in terms of inappropriate dietary habits, decreased physical activities and a corresponding increase in diet-related chronic diseases, especially among poor people. The impact of chronic diseases in the society rapidly increasing all over the world. It has been estimated that, in 2001, chronic diseases contributed 60% of the 56.5 million reported deaths in the world and 46% of the global burden of disease. Moreover, this percentage is expected to increase to 57% by 2020. Almost half of the total chronic disease deaths are related to cardiovascular problems. On the other hand, obesity and diabetes are also showing alarming trends: they already affect a large part of the population, and they have even started to appear earlier in life. In most of the regions of WHO (World Health Organization), deaths caused by chronic diseases dominate the mortality statistics [1]. This situation leds the governments, the concerned international agencies as well as non-governmental organizations to address food and nutrition policy, health promotion, and strategy for the control and prevention of chronic diseases.

1.1.2 Food Perception

Global obesity epidemic led a large number of researchers to study human perception of food, the relationship to food choices and amount of food intake and the role of the visual stimuli. In [4, 5, 6] the authors studied the relationship between brain activity, eating habits and food visual perceptions. Killgore et al. [4], correlated orbitofrontal and anterior cingulate cortex activity of 13 women to the view of high-calorie and low-calorie foods. They found out that bodymass index (BMI) is negatively correlated with both cingulate and orbitofrontal activity during high-calorie viewing, and just with the orbitofrontal activity during low-calorie viewing. This suggests a relationship between weight and responsiveness of the orbitofrontal cortex to images which depict rewarding food. In [5], the authors found that maintenance of a reduced body weight was associated with changes in brain activity elicited by food-related visual cues. They perform their test on 6 obese patients and proved that this kind of brain activity can be reduced through leptin administration. Medic et al. [6], examined the food choice and magnetic resonance imaging (MRI) of overweight and lean people during an unlimited buffet. Their aim was to assess the capability of the two groups (lean and overweight people) to evaluate the healthiness of food. Results shown that both are able to well distinguish healthy from unhealthy food. This suggests that obesity can be related on how the presence of food surpasses prior value-based decision-making.

In [7] Delwiche, described how visual cues can affect taste an flavour of food. For example, flavour, can be viewed not just a mere combination of raw materials or chemicals components, but also as a combination of different stimuli. Multiple factors, including visual appearance, can influence the interpretation of the primary stimuli and change the perception of taste, smell, and flavour.

McCrickerd and Forde [8] focused on how visual and smell cues lead food choice. Specifically, they described how the size of food and the amount of food served can effect the food intake. Simply splitting foods like cookies or chocolate bars, so they are viewed as smaller more numerous pieces, results in a reduction of intake of that food without changing palatability. Moreover, there are evidences which indicate that some adults and children choose and consume larger portions when served with larger dishware.

By observing that people seem to give more importance on the expected pleasure from food then the actual food intake, Petit [9] et al. discussed how food-related contents published in social media can help to choice of healthy meal. Seeing food presented in an appetizing and/or "ready to be eaten" way, gives the possibility to the viewer's brain to vividly imagine the consumption experience. Currently, the food industry uses social media to promote their products with good-looking food photos. Hence, the authors claimed that public health prevention and organizations could promote healthy lifestyles by exploiting the same food industry strategies.

The aforementioned works prove that would be interesting and technically possible to use Computer Vision and Machine Learning to extract information on how the food is presented and then try to find a correlation with health statistics.

1.2 Aims and Findings

The aim of this thesis is the investigation of food understanding. This term can be refereed to a set methods and approaches to extract information about food through automatic visual contents analysis. In my thesis I address three different problems:

- binary classification of food vs non-food images;
- retrieval and classification of food images;
- segmentation and volume estimation of food items in a dish.

Additionally, I introduce novel food datasets to address the aforementioned problems.

My approach for food vs non-food classification is based on one-class classification paradigm. This means, that a class only is used to build the mathematical model used to distinguish among two classes. A one-class classifier is trained only on positive samples. In classification phase, it considers all the negative samples like "anomalies". In this thesis, food is considered as positive class, whereas non-food as the negative one.

Food recognition and classification is investigated to study the best representation for food images, i.e. the one that achieves the best performance. Since texture-oriented representations have shown the best results, I propose a new descriptor called Anti-Texton, which considers Textons co-occurrences to improve the state-of-art Texton-based representation.

Finally, I present a database that contains annotated RGB and RGB-D images from 80 different meals served on a round dish accompanied by accelerometer data. Each meal consists of two to four different food items of know weight, volume and nutrient composition. Along with the dataset, the results of different baseline methods for segmentation, depth and volume estimation are discussed.

1.3 Contributions

The main contributions of this thesis, related to the food understanding, are summarised below:

- the original study of food Vs non-Food via One-Class Classification approach;
- the introduction of two new dataset of food and non-food images respectively;
- a comprehensive study of discrimination capability of texture (Texton) for food classification and recognition;
- the introduction of an original descriptor based on the notion of Anti-Texton;
- an extend version of the dataset UNICT-FD889 (i.e., UNICT-FD1200), with 311 new classes and a new labelling across 8 categories;
- a novel food dataset which includes RGB-D images and 3D models, to address volume and nutrients estimation;
- a baseline on the new dataset, for segmentation, depth estimation and volume estimation tasks.

The contributions about Cultural Heritage, reported in Appendix A are:

- the proposal of new method to perform virtual unrolling;
- the introduction of a new platform of semantic annotation, to present the studies conducted on Morgantina Silver Treasure;
- an accurate comparison of different 3D scanners, evaluated on an architectural element;
- a study of the compatibility between two parts of a Kouros through modern technologies.

The contributions of this thesis have been published in international conference, journal and book chapters.

International Journals:

- D. Allegra, E. Ciliberto, P. Ciliberto, G. Petrillo, F. Stanco, C. Trombatore. "X-ray Computed Tomography for virtually unrolling damaged papyri". Applied Physics A, 2016, Vol. 122(3). DOI: 10.1007/s00339-016-9796-1.
- G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, S. Battiato. "Retrieval and Classification of Food Images" Computers in Biology and Medicine, 2016, Vol. 77 Pages 23-39. DOI: 10.1016/j.compbiomed.2016.07.006.
- M. F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, G. Milazzo, F. L. Maria Milotta, L. Pellegrino, S. Schiavone, F. Stanco. "Integrated three-dimensional models for noninvasive monitoring and valorization of the Morgantina silver treasure (Sicily)" Journal of Electronic Imaging, 2016, Vol. 26(1). DOI: 10.1117/1.JEI.26.1.011015.
- F. Stanco, D. Tanasi, D. Allegra, F. L. M. Milotta, G. Lamagna, G. Monterosso. "Virtual anastylosis of Greek sculpture as museum policy for public outreach and cognitive accessibility" Journal of Electronic Imaging, 2017, Vol. 26(1). DOI: 10.1117/1.JEI.26.1.011025.

International Conferences:

- G. M. Farinella, D. Allegra, F. Stanco, S. Battiato. "On the exploitation of one class classification to distinguish food vs non-food images". In: Lecture Notes in Computer Science. Vol. 9281, 2015, pp. 375 – 383. DOI: 10.1007/978-3-319-23222-5_46.

- D. Allegra, E. Ciliberto, P. Ciliberto, F. L. M. Milotta, G. Petrillo, F. Stanco, C. Trombatore, "Virtual Unrolling Using X-Ray Computed Tomography", European Signal Processing Conference, 2015, pp. 2864-2868. DOI: 10.1109/EUSIPCO.2015.7362908.

- M. F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, F. L. M. Milotta, S. Schiavone, F. Stanco. "Archaeometric characterization and 3D survey: new perspectives for monitoring and valorisation of Morgantina silver Treasure (Sicily)". International Conference on Metrology for Archaeology, 2015, Vol. 1.

- D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F. L. M. Milotta, C. Santagati, F. Stanco. "Low Cost Handheld 3D Scanning for Architectural Elements Acquisition", Smart Tools and Apps in computer Graphics, 2016. DOI: 10.2312/stag.20161372.

- F. Stanco, D. Tanasi, D. Allegra, F. L. M. Milotta. "3D Digital Imaging for Knowledge Dissemination of Greek Archaic Statuary", Smart Tools and Apps in computer Graphics, 2016. DOI: 10.2312/stag.20161373.

- D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella and S. Mougiakakou. "A Multimedia Database for Automatic Meal Assessment Systems"International Workshop on Multimedia Assisted Dietary Management, 2017.

International Book Chapters:

- D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F. L. M. Milotta, C. Santagati, F. Stanco. "Hand Held 3D Scanning for Cultural Heritage: experimenting low cost Structure Sensor scan", A. Ippolito, M. Cigola (editors), Handbook of Research on Emerging Technologies for Cultural Heritage, IGI Global, pp. 475-499 (2017). DOI: 10.4018/978-1-5225-0675-1.ch016.

The publications not related to this thesis have been reported in Appendix B.

1.4 Computer Vision for Food Understanding

Although food understanding has been largely addressed in the last decade by Computer Vision researchers, it has a long history. Following the path of food-related Computer Vision works, from the beginning in 1977, it is possible to coarsely define four different areas:

- **Food detection and recognition for automatic harvesting:** automatic detection and recognition of fruits and vegetables are useful to enhance robots affordable and reliable vision systems in order to improve the harvesting process in terms of quality and speed;
- **Food quality assessment for industry aims:** in the late 80s, industrial meals production knew a large scale expansion, especially in developed countries. Consequently, the evaluation of produced food quality with vision systems became an interesting and valuable challenge;
- **Food logging, dietary management and food intake monitoring:** as described in Section 1.1, the growth of the number of people affected by diseases caused by a non healthy diet led the researchers to study the problem. From late 90s, the focus was moved to the usage of Computer Vision solutions to help food experts (e.g., nutritionists) for the monitoring and understanding the relationships between patients and their meals;
- **Food classification and retrieval:** the large and fast spreading of mobile cameras, together with the birth of social network services gave the possibility to upload and share pictures of food. For these reasons, in recent years, classification and retrieval of food images become more and more popular.

In Fig. 1.2 it is reported a time-line that shows the periods on which the interest for a certain area was growing and have got highest popularity by taking into account the published papers in literature over the years.

Despite most of the solutions proposed in the different areas overlap, the main goals of the developed systems are different. Hence, if a certain accuracy obtained by

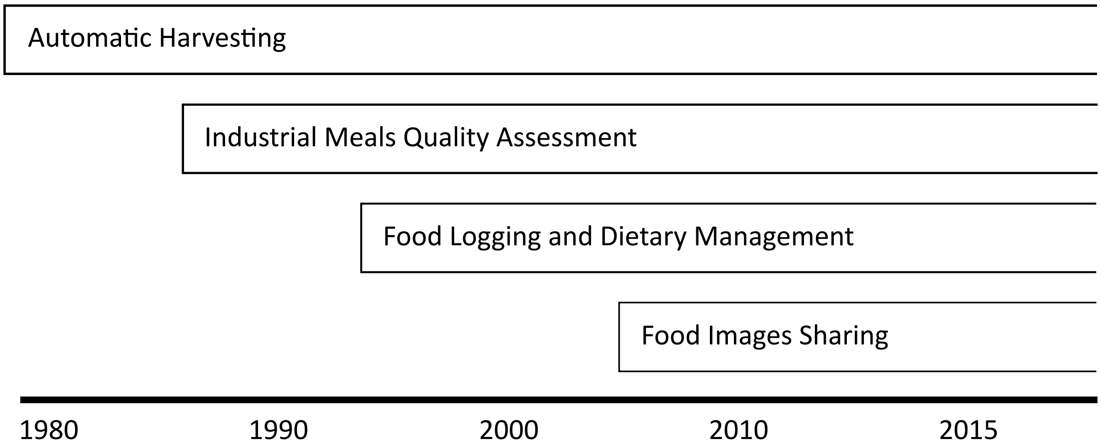


Figure 1.2: Food image analysis tasks employed during the years.

a system for the detection and recognition of food for automatic harvesting could be acceptable by a robotic industry, there is no guarantee that the same performance were sufficient in systems for diet monitoring, i.e. for patients with diseases like diabetes or food allergy. For these reason it has been chosen to categorize works about food in the aforementioned areas.

In the following subsections a detailed review the state-of-the-art works in the identified research areas is provided, in order to remark the importance of Computer Vision contribution.

1.4.1 Food detection and recognition for automatic harvesting

Fruits harvesting can be addressed by different techniques, however it is really important that they do not cause damages to the fruit and/or to the tree/branches. Hence, accurate systems for fruits/vegetables detection and recognition from images are desirable in order to perform this task correctly. One of the first Computer Vision solution has been proposed by Parrish and Goksel in 1977 [10], and it is related to apples detection. The system they designed consists of a B/W camera and an optical red filter. First of all the acquired image is binarized through thresholding operation and smoothed to mitigate noise and artifacts. Then, the region roundness is estimated by measuring the difference between the longest horizontal and vertical segments inside the region itself. Finally, an image area is classified as apple through a density estimation procedure and then a thresholding step.

In 1988 Levi et. al [11] implemented AID, a robot vision system for oranges recognition. A pseudo-grey image is obtained by means of an electronic filter used for image enhancement. The value of each pixel, coded using 6 bits, is proportional to the difference between the pixel hue value and a reference hue value. Then, the gradient is computed using a standard Sobel filter to get the magnitude and directions in two separate images. To correctly detect the oranges location, a gradient template previously computed is used. Specifically, a matching between the detected gradient and the template is performed. This approach allows to achieve an accuracy of 70% in fruits detection.

Another orange recognition method is proposed in [12]. It is based on colour images, in particular the Hue and Saturation components of each pixel are used as dimensions of a two-dimensional feature space. Then, two thresholds based on the maximum and minimum values for each component are used to define a region in the feature plane. Each pixel inside this region is classified as orange. With this method approximately 75% of the pixels are correctly classified. The authors extend their study employing a Bayesian classifier [13], and exploiting the RGB values rather than the Hue and Saturation components. Also in this case the goal is to separate fruit pixels from the background pixels. The performed tests show an accuracy of 75%.

A machine video system for melon harvesting [14] was developed by The Purdue University (USA) and The Volcani Center (Israel). This system is able to analyse binary image to locate the melons and estimate their size. Basic operations like shape and textures analysis are performed in order to obtain multiple candidate regions from the original image. Subsequently, thanks to prior knowledge on the domain, the candidates are evaluated to discard false positive and multiple detections by achieving a true positive rate of 84%.

In 1995 The Italian institute CIRAA designed a robotic system called AGROBOT [15]. The main aim was automatizing greenhouse operations. The images are acquired through a colour monocular camera and are segmented via thresholding on the Hue and Saturation histograms. This system is also able to extrapolate information about the 3D geometry of the scene through stereo matching. The performances of AGROBOT shown high quality results: 90% of correctly detected ripe tomatoes; the main error causes in this system are related to the occlusions.

Jiménez et al. [16], exploit the 3D information obtained with a laser scanner to perform automatic harvesting of spherical fruits. Thanks to the laser scanner, it is possible to map the points of the scene in the 3D world using spherical coordinates. Then, for each points also the laser energy attenuation value is stored. Hence, the scene can be described by four images: the azimuth angles, the elevation angles, the distance from the sensor (i.e. a depth map) and the attenuation values. These images are processed, and taking advantage of the information retrieved by the scanner, four new images are produced in output. Three of them are actually used for the orange recognition: one is an enhancement of the previous image representing the distance from the sensor, the other two encode respectively the apparent reflectance and the reflectance of the surfaces. The image analysis procedure is based on these two information. Firstly, the apparent reflectance image is binarized through a threshold to separate the background from the foreground; then the remaining pixels are clustered using the Euclidean distance. The detected clusters with a low number of pixels are immediately rejected as “not-fruit”. This step is performed to eliminate the possibility of random small areas of a highly reflective non-fruit object. However this reflectance-based approach is not able to detect fruits whose reflectance is under 0.3. For this reason, the authors proposed to use the Circular Hough Transform on the distance image to detect fruits based on shape assumption.

1.4.2 Industrial food quality assessment

The food quality inspection is not strictly related to domain of dietary food monitoring, nevertheless it concerns food image analysis. Computer Vision systems have been used to perform quality assessment since this task is very critical for food industry that have to produce products that satisfy the customers.

In [17] Munkevik et al. describe an approach to check the quality of industrial cooked meals. Firstly, they proposed to segment the food and then extract 18 different features from the segmented image. The selected features allow to represent different properties. Among them, the size of the food items on the plate, the overlapping between different food items, the shape of the food item and also information about the colour. Finally, a Self Organizing Feature Map [18] is employed to learn the model of a meal. The authors propose method extension in [19], where they consider a larger number of food items an exploit an Artificial Neural Network (ANN) to improve classification performances.

In 2007, Kilic et al. [20] addressed a beans quality classification problem. The dataset used for the experiments consists of 511 images with a variable number of beans. Morphological operators are employed for image segmentation, than the first 4 order statistic are computed on the RGB channels. To asses the beans quality a score based on three levels for both, integrity and colour, has been proposed. Although, $3 \times 3 = 9$ possible combinations can be defined with this quality score, the authors decided to use 5 combinations only. Each combination is considered as a different class. Finally, the classification is performed by using ANN and splitting the dataset in the following way: 69 beans images for training, 71 for validation and 371 for testing.

The quality of pizza production has been addressed in several works, as the ones of Du and Sun [21, 22]. The proposed algorithms are intended to inspect three different pizza properties: shapes, toppings and sauce spread. While the approach in [21] faces only the evenness of the topping, the method described in [22] is more complex and involves also other parts of the pizza to be evaluated. To perform quality assessment by exploiting shape, geometrical features such as the area ratio, aspect ratio, eccentricity, roundness and also some coefficients of the Fourier transform have been considered. Concerning the topping and the sauces, HSV colour histograms and Principal Components Analysis (PCA) are used. Classification is performed by considering four quality levels for the shape and five for topping and sauce spread. To build the classification model a set of binary Support Vector Machine (SVM) organized in a Directed Acyclic Graph (DAG) has been employed. The dataset used for experiments includes 120 images for the shape, 120 images for the sauce and 120 images for the topping.

Finally, a review about the methods for food quality assessment is presented in [23, 24, 25]. The authors address the different acquisition systems as well as the features that can be employed in different tasks. Last but not least, the machine learning algorithms used to perform the decision for this task are analysed.

The inspection of the food quality is usually addressed in constrained environment, with a few food classes and low variability. For this reason, very simple features such as colour or shape information are enough to face the problem and achieve very good results. This kind of scenario is different from the one where images of food are acquired during real meals of a patient or they are downloaded from a social networks. A generic system for food intake monitoring have to be able to work in low constrained scenario without prior knowledge. Differently than an industrial factory where the ingredients, the quantity and the appearance of food are known in advance, in a generic food understanding problem there are many variables. High number of food classes and ingredients, the food mixing as well as illumination, orientation, different acquisition devices and so on, make this task very challenging.

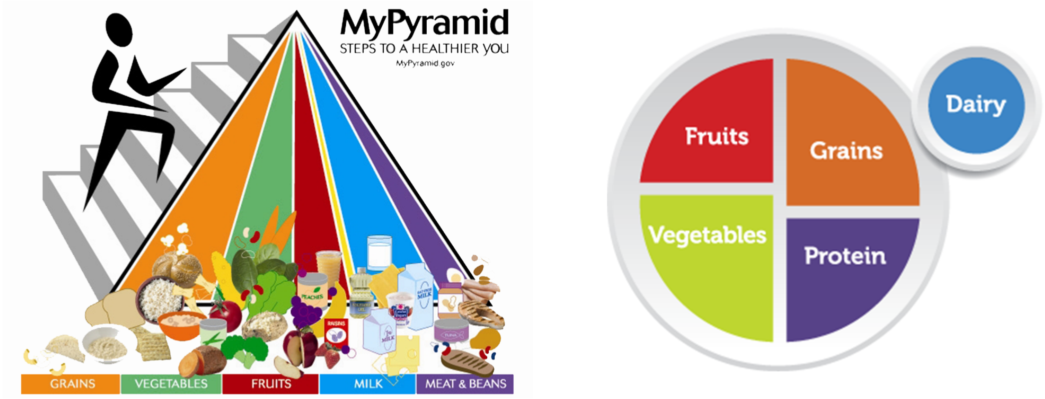
1.4.3 Food logging, dietary management and food intake monitoring

Diet monitoring is a critical aspect of the human health, since can help to reduce chronic disease risks such as diabetes or obesity. For this reason, since the ’70, computer science has been exploited to assist the medical teams in dietary assessment of the patients. However, the first systems for food logging and intake monitoring were calculators for nutrition values that exploited standard food list [26, 27]. Hence, they did not use the Computer Vision techniques.

Since it has been proved that food diaries are effective instrument, in the last decade Computer Vision researchers have put effort to propose reliable tools to improve the automatic detection and recognition of food images, as well as the nutritional merit evaluation. These types of tools can increase self-awareness of eating habits, moreover to add photographs to the written diary have a more effective impact on the patients. A discussion about the state-of-art systems for food logging is given below.

FoodLog[1] [28, 29, 30, 31, 32] is an Internet application that gives the possibility to acquire and stores information regarding daily meals. The main aim of this system is to help the users to keep note of their meals and, above all, to correctly balance the main nutrients coming from different kinds of food (e.g., carbohydrates,

[1] http://www.foodlog.jp



(a) (b)

Figure 1.3: (a) MyPyramid model and the five classes: grains, vegetables, meat, fruits, milk; (b) The currently MyPlate model.

protein, etc.). The application enables the user to upload one or more pictures on a remote folder, where the all the information are stored.

Kitamura et. al proposed FoodLoog in [28]. The images which include food items are detected by using colour features based on HSV and RGB, as well as the shape of the plate. Food detection is performed by training a SVM classifier according to the following strategy: the images are divided in 300 blocks and each block is classified as one of the five nutritional groups described in the "My Pyramid" official model (grains, vegetables, meat & beans, fruits, milk) or as "non-food". However this model has been replaced in 2011 by "MyPlate" model [2]. The old "MyPyramid" model and the new "MyPlate" are depicted in Fig. 1.3

In 2009, Kitamura et al. [30] extended their previous work by exploiting more local features. Colour information are coupled with SIFT descriptors [33] by selecting keypoints with three different methods (Difference of Gaussians, centres of grid, centres of circles). Further improvements are proposed in [31], by including a pre-classification step and the customization of the food image estimator. Finally, in [32] the Support Vector Machine classifier is replaced by a Naive Bayesian one.

Shroff et al. [34] proposed a system to help people affected by diabetes to follow their dietary rules. The authors employ two different kinds of features: objected-related features like colour, size, texture, shape; context features such as time of

[2] https://www.choosemyplate.gov/

the day or user preferences. ANN classifier is used by the authors to prove that the context information lead an improvement in the accuracy of the monitoring system.

The work of Puri et al. [35] focuses on food recognition and 3D volume estimation. Firstly the photos, captured under different lighting conditions and poses, are normalized by color and scale by using a particular calibration card placed besides the food items. For features selection they employ an Adaboost-based algorithm that combines colour (in RGB and LAB space) and texture information (Maximum Response filters). The goal is to perform a segmentation by classifying the different food items in a plate. The final classifier is obtained as a linear combination of several weak SVM classifiers, one for each feature. For 3D reconstruction they use RANSAC [36] to estimate pose and then, dense stereo matching for depth estimation.

Another work in which 3D reconstruction is exploited is the one of Dehais et al. [37]. The 3D model is used for food volume estimation. Stereo pairs are used to computer disparity map and then a dense points cloud is built and aligned with respect to the estimated table plane. This algorithm is designed to work by employing a specific marker placed on the table. By assuming the different food items in the plate are already segmented, each food segment is projected on the 3D model for volume computation. They define the volume as the integral of the distance between the surface of each segment and either the plate (identified by its rim and reconstructed shape), or the table (identified by the reference pattern).

In [38] the authors categorise food from video sequences taken in a supervised environment. The dishes are placed on a table covered with a black tablecloth. They take into account an elliptical Region-of-Interest (ROI) and extracted different kind of descriptors such as MSER [39], SURF [40] and STAR [41]. Hence, the images are represented exploiting the Bag of Words paradigm and vocabulary with 10000 visual words built by using K-means clustering. Subsequently each data point is associated with the closest cluster using the Approximated Nearest Neighbour algorithm. To capture information about colour, histogram in the HSV space is computed inside the ROI and combined with the aforementioned descriptors. The final aim is to classify the dish in a specific frame of the sequence. In the proposed approach each unclassified frame is compared with frames that are already classified. To do this, a similarity score is computed for both, the Bag of Words representation and the colour histograms. The score for the first representation is computed by exploiting the term frequency-inverse document frequency (tf-idf) [42] technique, while for the colour similarity, the correlation coefficient between the $|L_1|$-norm of two histograms is used. Finally, the two scores are linearly combined with different weights to obtain a global score for the considered frame. Moreover, since the calories for the reference dish are known, this score allow to roughly quantify the difference of them in the two frames.

Food intake estimation is also studied in the work of Liu et al. [43] where a wearable system equipped with a camera and a microphone is proposed. The microphone is used to detect chewing sounds, so that the Computer Vision part of the framework can be activated. To identify frames which contains food, they propose to use a simple approach based on ellipse detection and colour histograms. After the ellipse is found, it is split in four quadrants and, for each of them, the colour histogram is computed in the C-color space [44]. Finally, the food consumption evaluation is performed by computing the difference between the histogram of subsequent frames.

1.4.4 Food classification and retrieval

In order to recognise food depicted in images, two computation strategies can be usually considered: classification and retrieval. In both cases the task is to identify the category of a new food image observation on the basis of a training set of data. The main difference between the two approaches stay in the mechanism used to perform the task. In case of classification the training set is used just to learn the decision function by considering the representation space of the images. Hence, the training images are represented as vectors in a feature space through a transformation function (e.g., Bag of Visual Word approach by considering SIFT or Textons features [45, 46]) whereas a learning mechanism is used to train a classifier (e.g., a Support Vector Machine) to discriminate data belonging to different classes. After that, the training dataset is discarded and a new observation can be classified by considering the employed feature space and the trained classification model. In case of retrieval, the training set is maintained and the identification is performed comparing the images through similarity measures (e.g., Bhattacharyya distance) [47] after their representation in the feature space.

In [48], a framework for food classification of japanese food is proposed. The approach is trained and tested on a dataset with 50 classes. Three kinds of features are extracted and used: a) Bag of SIFT; b) Colour Histograms c) Gabor Filters [49]. The keypoint sampling strategy on which the SIFT descriptor have been computed is implemented with three different ways: using the DoG approach, by random sampling and using a regular grid. To compute Color Histograms, the images are first divided in 2×2 regions, and for each region a 64-bin RGB histogram is calculated. The region-based histograms are then concatenated into a 256-bin. In a similar way, the images are split in 3×3 and 4×4 blocks to compute Gabor Filters responses. The employed Gabor filters take into account four different scales and six orientation, so for the whole image a 216 or 384-dimensional vector arises as result of the extraction step. While Color Histograms and Gabor Filters provide a representation of the images by themselves, SIFT keypoints are clustered generating two different vocabularies with 1000 and 2000 codewords and the images are represented using the Bag of Words paradigm. Summing up, for each image 9 different representation are provided, one coming from the Color Histograms, two from the Gabor Filters with different blocking schemes and six from the combination of sampling strategies and vocabulary size for SIFT features. Classification is performed using a Multiple Kernel Learning SVM (MKL-SVM) [50]. In [51] the dataset is extended up to 85 classes, and 8 variants of Histogram of Oriented Gradients (HOG) [52] are introduced as new features. Moreover, the χ^2 kernel is employed as a kernel function in the MKL-SVM. An extended version of the dataset, containing 100 food items, has been used in [53] where candidate regions are identified using different methods (whole image, Deformable Part Model (DPM) [54], a circle and the segmentation method proposed in [55]). The final segmentation arises by integration of the results of the aforementioned techniques. For each candidate region, four sets of features are computed: Bag of SIFT and Bag of CSIFT [56], Spatial Pyramid Representation [57], HOG and Gabor Filters. Then a MKL-SVM is trained for each category, and a score is assigned to every candidate region. The experiments are conducted on images containing both single and multiple food-item. In successive work [58] the same approach is used, but the scores assigned by the classification algorithm are re-arranged applying a manifold learning technique to the candidate regions.

The dataset used in [58, 53] is called UEC FOOD 100 (Fig. 1.4) and is an extension of the dataset presented in [51, 48]. On this dataset, other approaches have been tested. For instance, pre-trained Convolutional Neural Networks (CNN) [59] are used in [60] for feature extraction. The CNN features are coded using the Fisher Vectors technique [61], and then the classification is performed by means of SVM. Raví et al. [62] exploited jointly different features in a hierarchy to obtain real-time food intake classification. The hierarchy of features encodes, in some way, the complexity of the images: on simple classes, the classification will rely on the features at the first level, while on more complex classes more features will be used. To represent the images, the Fisher Vector [63] technique is employed, and PCA is applied as in [64]. To perform classification, a linear SVM is trained using the one-vs-rest strategy. The UEC FOOD 100 has been extended to 256 categories (UEC FOOD 256) in [65] using a so-called "foodness classifier" and transfer learning on images coming from crowdsourcing. UEC FOOD 100 and UEC FOOD 256 have been employed by Yanai et al. [66] to finetune a pre-trained deep convolutional neural network (pre-trained with 2000 categories in the ImageNet).

Another dataset used in literature is the Pittsburgh Food Image Dataset (PFID) [67]. This dataset is composed by 4545 still images, 606 stereo pairs, 303 videos for structure from motion (360° videos), and 27 privacy-preserving videos of eating events of volunteers. The images portrays 3 instances of 101 food items, bought in 11 different fast food chains. In [67], a baseline for future experiments is provided. The authors use color histograms and Bag of SIFT features to train a multi-class SVM. In [68], an ingredient based segmentation is performed using a Semantic Texton Forest [69]. Hence, pairwise statistics of local features are computed on the segment connecting two points, and specifically: a) orientation; b) midpoint; c) between-pair; d) distance. Moreover, two joint features are considered (Distance + Orientation and Orientation + Midpoint). SVM with a χ^2 kernel is employed for classification purpose. The PFID is also used for calories estimation in [70]. SIFT are extracted and a cosine-based distance function is used for matching. Rankings on food categories can be obtained in two ways: 1) a ranking based matching, based on top T items of each frame-based rankings; 2) a count-based matching based on sum of keypoint matching counts over all video frames. Zong et al. [71] locate the keypoints using the SIFT detector, applying the Local Binary Pattern (LBP)

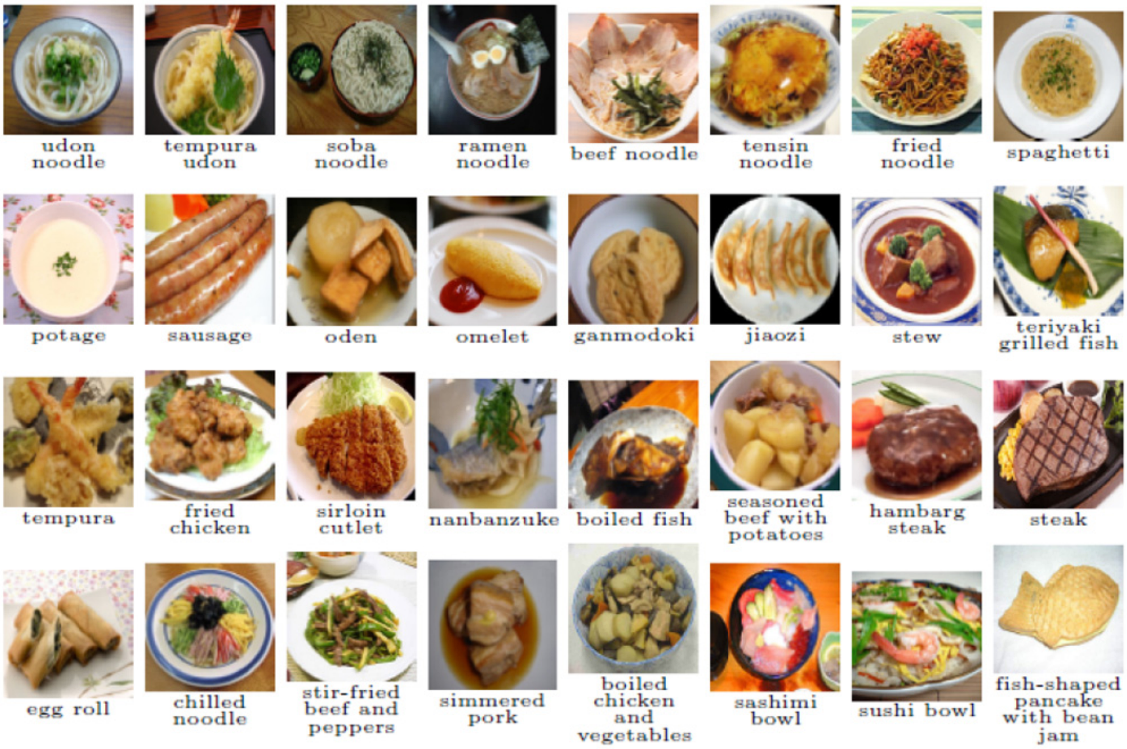


Figure 1.4: A sample of 32 images from 32 different classes of UEC FOOD 100 dataset.

[72]. Then they employ a BoW model, using a codeword filtering function to select the most discriminative words in the vocabulary. Dictionary creation is performed in a class-based manner. To provide spatiality, the shape context descriptor [73] is calculated on the image space, considering the words as keypoints. The images are classified by means a cost function which takes into account the Bhattacharyya distance and the shape context matching cost. Nguyen et al. extended the previous mentioned approach introducing the Non-Redundant Local Binary Pattern (NRLBP) [74] and propose two strategies to classify the images: the first makes use of a SVM, the second is based on a cost function. Farinella et al. propose two different approaches on the PFID: one [75] is based on the representation of food images as Bag of Textons. Textons are computed using the responses of MR4 filters, then clustered in a class-based fashion obtaining a visual vocabulary. In the

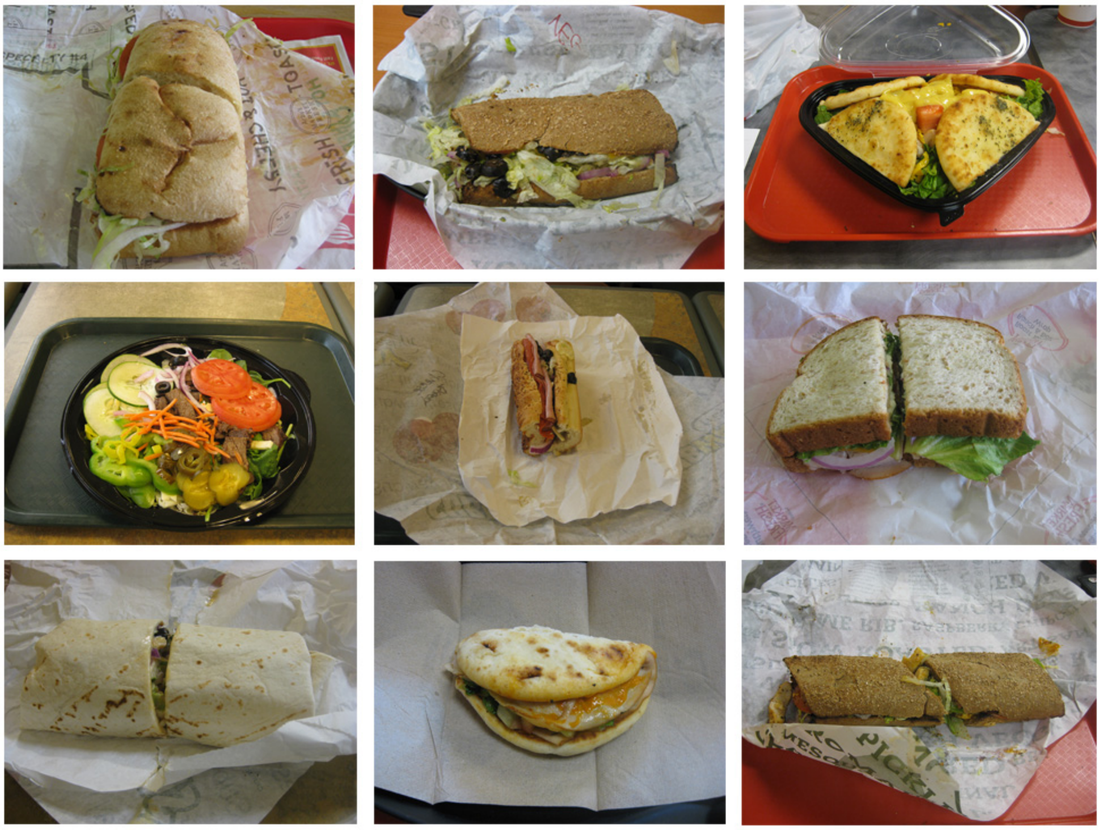


Figure 1.5: A sample of 9 images from PFID dataset

other approach [76] SIFT and SPIN [46] features are computed over a dense grid, and multiple runs of the k-means algorithm are performed separately for SIFT and SPIN. The vocabularies obtained in output are used as input for an Expectation-Maximization based consensus clustering technique [77]. In both approaches, SVM is used as classifier. The method proposed in [78] combines different descriptors calculated on patched centred on the keypoints detected by the Harris-Laplace detector. For each feature, a visual codebook with 1000 words is built, and for each set a gaussian kernel is computed. The resulting kernels are used as input to train a Sequential Minimal Optimization (SMO) MKL-SVM. A small sample of PFID dataset is depicted in Fig. 1.5.

Bosch et al. propose a method for food identification based on global and local features [79]. As global features, they use: 1) 1^{st} and 2^{nd} moment statistics computed on the color channels of the image; 2) entropy statistics; 3) predominant color statistics. As local features, they consider small patches, and calculate the following features: 1) local color statistics; 2) local entropy color; 3) Tamura features; 4) Gabor filters; 5) SIFT descriptor; 6) Haar wavelets; 7) Steerable filters; 8) DAISY descriptor [80]. While the global features are used as input for a SVM with a RBF kernel, the Bag of Words approach is used with local features. Classification, in this case, is done using a Nearest Neighbour algorithm. This approach was tested on a subset of the dataset created at Purdue University [81]. The Purdue Food Dataset is an extension of the USDA Food and Nutrient Database for Dietary Studies (FNDDS), created having in mind the goal of augmenting *"an existing critical food database with the types of information needed for dietary assessment from the analysis of food images and other metadata"*.

Rahmana et al. in [82] present a dataset with 209 acquired using a iPhone3, to be used for retrieval purposes. They propose, as a baseline, Gabor filter variants to ensure scale and rotation invariance to their algorithm. However, they perform also a classification task, grouping the categories in 5 groups (Bread, Cereal, Veg, Fruit, Fast).

Another system for mobile food recognition is proposed in [83]. Here, color histograms on the RGB space are computed on 3×3 blocks and a dictionary with 500 visual words is built on SURF descriptors, to enclose local features in the general description of the image. To classify the images, a linear SVM with explicit embedding [84] is employed. It is interesting to note that the authors propose a system able to suggest the direction to which the camera should be moved, in order to improve classifier accuracy. Also, a dataset with 50 categories containing 100 images each is presented.

A Computer Vision system for Chinese food identification has been proposed by Chent et al. in [85]. The authors work on a database composed by 50 categories of ready-to-eat Chinese meals, with 100 images per category. On each image, the following features are extracted: 1) SIFT with sparse coding; 2) LBP with

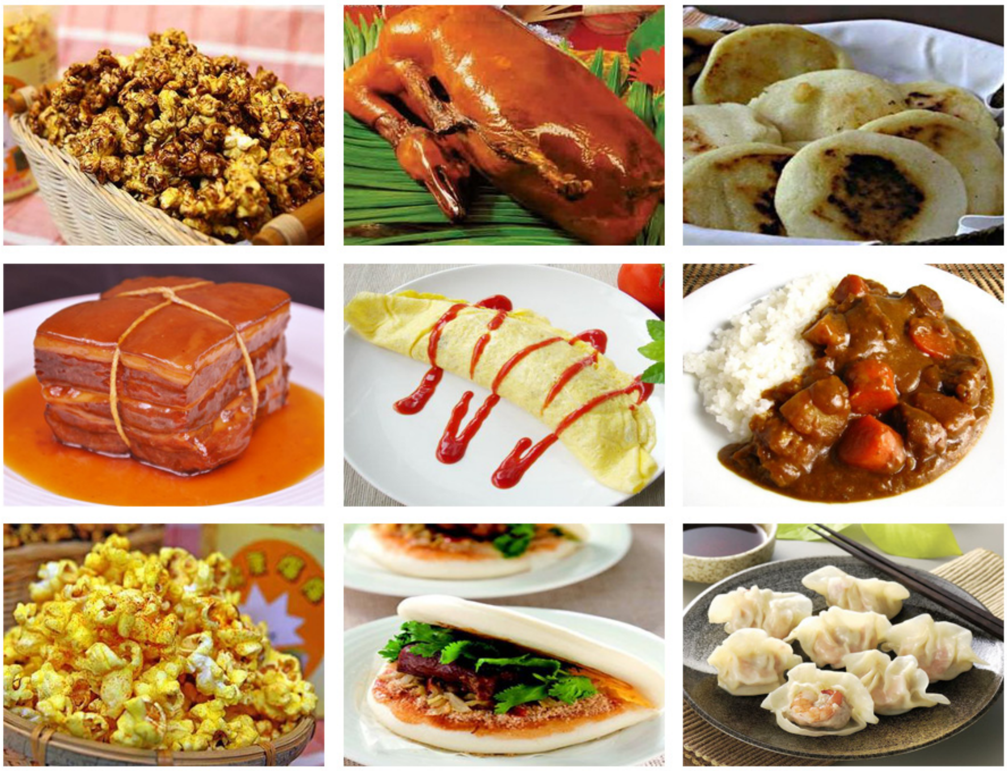


Figure 1.6: A sample of 9 images from Chen et al.' dataset.

multi-resolution sparse coding; 3) color histograms; 4) Gabor textures. A SVM is trained for each feature using 5-fold cross validation; the fusion is done using the Multi-Class AdaBoost algorithm. Marginally, the authors propose also a quantity estimation technique using Microsoft Kinect, but this approach has been tested only on a single item of "hot & sour soup". A sample of this dataset is reported in Fig. 1.6.

A food recognition system integrated on a chopping board is the topic of the work by Pham et al. [86]. In this work, an imaging system composed by a matrix of optical fibres is placed under an appropriately prepared chopping board. The sensor acquires the image and afterwards a 64-dimensional color histogram and a 64-dimensional vector of Bag of SURF features are computed. The algorithms used to classify the images are kNN and SVM. The training and testing phases make use of a dataset composed by 1800 pictures of 12 food ingredients.

Random Forest (RF) [87] are used in [88] for mining discriminative regions. Superpixels are generated from the images and dense SURF and color histograms are computed and encoded using Fisher Vectors [61]. These descriptors are supplied to the RF for training. Once the RF has been trained, the leaves constitute the set of candidates for the components. Using a probability-based distinctiveness function, the most discriminative leaves are selected. Hence, a linear binary SVM is trained for each class, using the samples lying in the most discriminative leaves as positive samples and hard negative samples to speedup the learning process. Alongside with the algorithm, the authors present a novel dataset, called Food-101, composed by 1000 images for each one of the 101 most popular dishes on `foodspotting.com`. Some images of these dataset is shown in Fig. 1.7.

In [89] Xin et al. propose UPMC Food-101 (Fig. 1.8), a new dataset of 101000 images to address the recipe recognition problem. This dataset includes the same 101 categories of Food-101 and 1000 new images for each one. Google Image Search engine is exploited to retrieve 1000 images for each of the categories, moreover for all the images the related HTML textual description is collected. To benchmark the dataset, Bag of Word and CNN approaches are employed and textual information are embodied to improve classification performance.

Other food dataset include images and related geocontext information, such as GPS coordinates, restaurant where the dish is cooked an so on. Herranz et al. [90] propose a probabilistic model to combine locations, restaurants and visual features by exploiting a reduced set of the dataset collected by Ruihan et al.[91] from Institute of Computing Technology, CAS. To each of the restaurants are associated the related geographical coordinates to uniquely locate it and a menu which includes at least 3 dish categories. Then, for each of these categories, more than 15 images are included.

The UNICT-FD899 [92] has been acquired by users with a smartphone in four years during meals (i.e., iPhone 3GS or iPhone 4) in unconstrained settings (e.g.,

Figure 1.7: A sample of 35 images from Food-101 dataset.

different backgrounds and light environmental conditions). Each dish has been acquired trough a smartphone multiple times to introduce photometric (e.g., flash vs no flash) and geometric variability (rotation, scale, point of view changes). The overall dataset contains 3583 images acquired with smartphones. The dataset is designed to push research in this application domain with the aim of finding a good way to represent food images for recognition purposes. The first question the authors try to answer is the following: are we able to perform a near duplicate image retrieval (NDIR) in case of food images? Note that there is no agreement on the technical definition of near-duplicates since it depends on "how much" variability (both geometric and photometric) the system can tolerate. For instance, some approaches define the near duplicate of an image as the images obtained transforming

Figure 1.8: A sample of 15 images from UPMC Food-101 dataset.

the original by means of slight common editing, such as contrast equalization, scaling, cropping, etc. Other techniques (e.g., [93, 94])consider as near duplicate the images of the same scene but with different viewpoint and illumination. In [92], the authors consider this last definition of near duplicate food images to test different image representations on the proposed dataset. Then, they benchmark the proposed dataset in the context of NDIR by using three standard image descriptors: Bag of Textons [95], PRICoLBP [96] and SIFT [33]. Results confirm that textures and colors are fundamental properties. The experiments performed point out that Bag of Textons representation is more accurate than the other two approaches for NDIR. UNICT-FD889 dataset is a collection of food images acquired by users in real cases of meals. Each plate of food has been acquired multiple times (four in the average) to guarantee the presence of geometric and photometric variabilities (Fig. 1.9). It

Figure 1.9: A sample of 96 images included in UNIC-FD889 dataset.

is designed to arouse research in this application domain with the aim of finding a good way to represent food images for recognition purposes.

A comparative analysis on features and classifiers is the core of [97]. The authors test several features, basically related to three aspects (color, texture, local regions) and two classifiers (kNN, Vocabulary Tree [98]) on a novel dataset composed by 42 classes, with a total of 1453 images.

FRIDa dataset has been proposed in [99] and includes 877 images belonging to 8 different categories: natural-food, transformed-food (e.g., cooked food), rotten-food (e.g., moldy fruits), natural-non-food items (e.g., pinecone), artificial food-related objects (e.g., fork, spoon), artificial objects, animals (e.g., butterfly), and scenes (e.g., mountains). This dataset has been validated on a sample of 73 standard variables (e.g., ambiguity, familiarity, etc.) as well as variables related to food items (e.g., distance from eatability, perceived calorie content, etc.). In Fig. 1.10, some

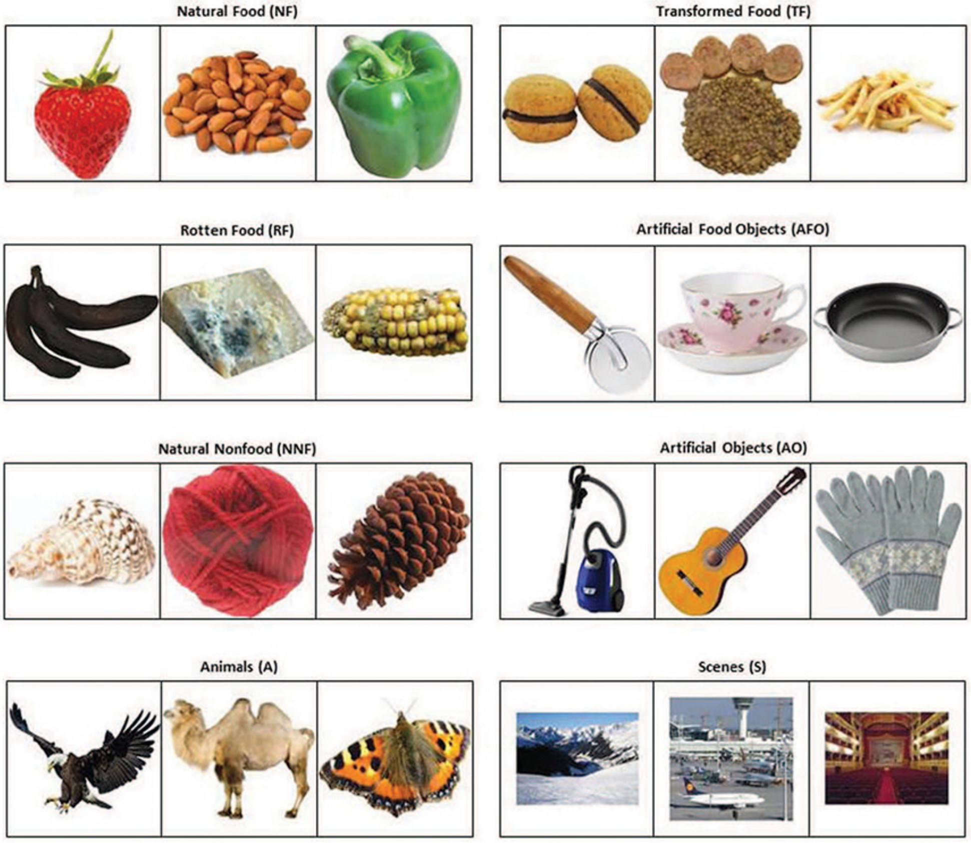


Figure 1.10: A sample of 3 images for the each of the 8 classes in FRIDa dataset.

images of this dataset is shown.

In [100] Pouladzadeh et al. introduced FooDD (Fig. 1.11). It is a dataset of 3000 images across a large variety of food photos taken from different devices and under different illumination conditions. The authors have used color segmentation and k-mean clustering in order to perform food segmentation; then, they have employed Cloud SVM and deep neural network for recognition and calories estimation.

Table 1.1 summarizes the main features of the publicly available datasets reported in the state-of-the-art works in the last years.

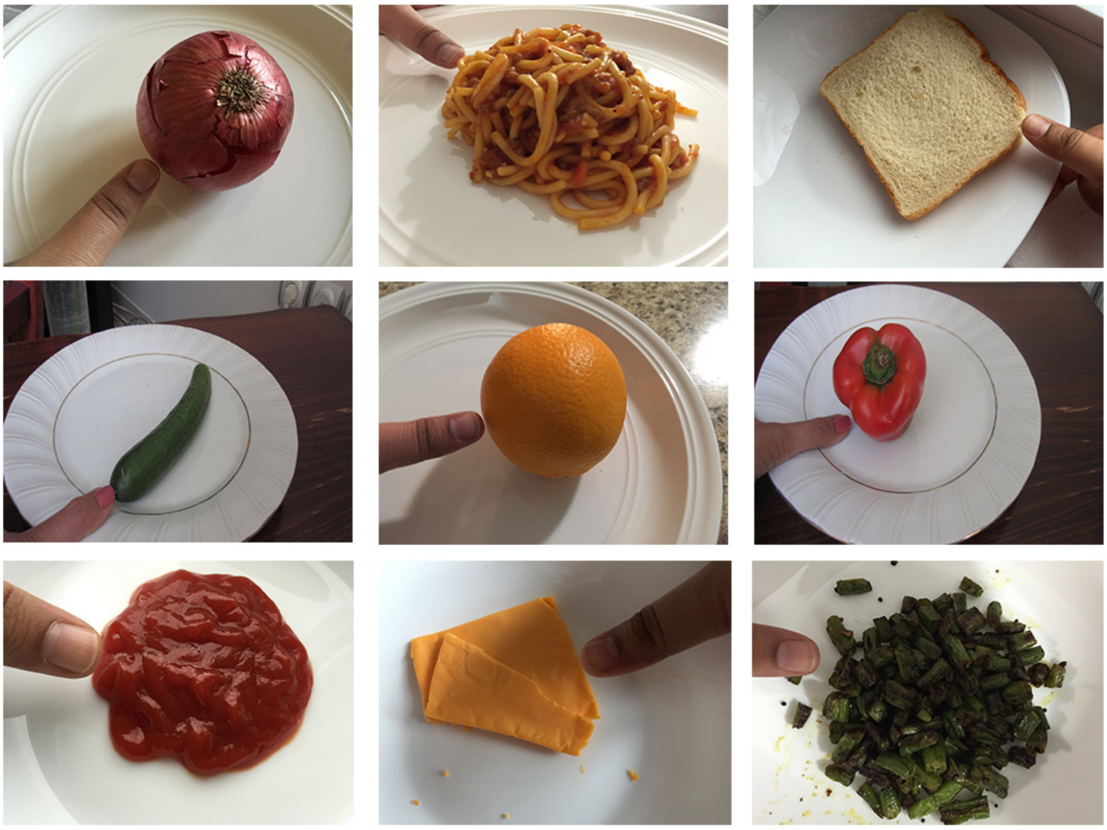


Figure 1.11: A sample of 9 samples included in FooDD dataset.

Table 1.1: Publicly available food datasets.

Dataset	**Presented in**	**Classes**	**Img per Class**	**# of Img**
UEC FOOD 100	[53]	100	≈ 100	9060
UEC FOOD 256	[65]	256	≈ 100	31651
PFID	[67]	101	18	1818
FRIDa	[99]	8	ND	877
NTU-FOOD	[85]	50	100	5000
ETHZ Food-101	[88]	101	1000	101000
UNICT-FD889	[92]	899	3/4	3583
FooDD	[100]	23	ND	3000
UPMC Food-101	[89]	101	1000	101000
CAS Dataset	[90]	ND	ND	117504

Chapter 2

Food VS No-Food Via One Class Classification

To build food monitoring systems that can automatically collect a food images the automatic classification between food vs non-food images is the first problem to solve.

This problem has been taken into account in [30, 101]. In [30] Kitamura et al. presented a food-logging web system which consider this task to analyse the food intake balance and visualise a food log. They used both global and local features to represent images and SVM to classify them. Circle detection and color information are exploited as feature to identify the presence of dishes in images. However in real scenario not all the images of food include the plate. Moreover the shape of the plate can be different. Kagaya et al. [101] used deep learning for food detection and recognition. As in [30], to perform a proper training of the method proposed in [101] both food images and non-food ones have to be employed. This means that to train a food detector, the variability of the non-food classes have to be captured in dataset used for training purpose. Despite could be simple to collect images of food (e.g., by considering the current available food dataset or images downloaded from website dedicated to food), to build a proper representative dataset of non-food images can be a challenging task. Differently than cited works, I investigate one-class classification approach (OCC)[102] to recognize when an image is belonging to the food class. Multi-class classification methods, such as the ones proposed in the aforementioned works, aim to classify an unknown image into one of several predefined categories (two classes in case of food vs non-food classification). One-class classification approaches allow to obtain a model of a single class, so

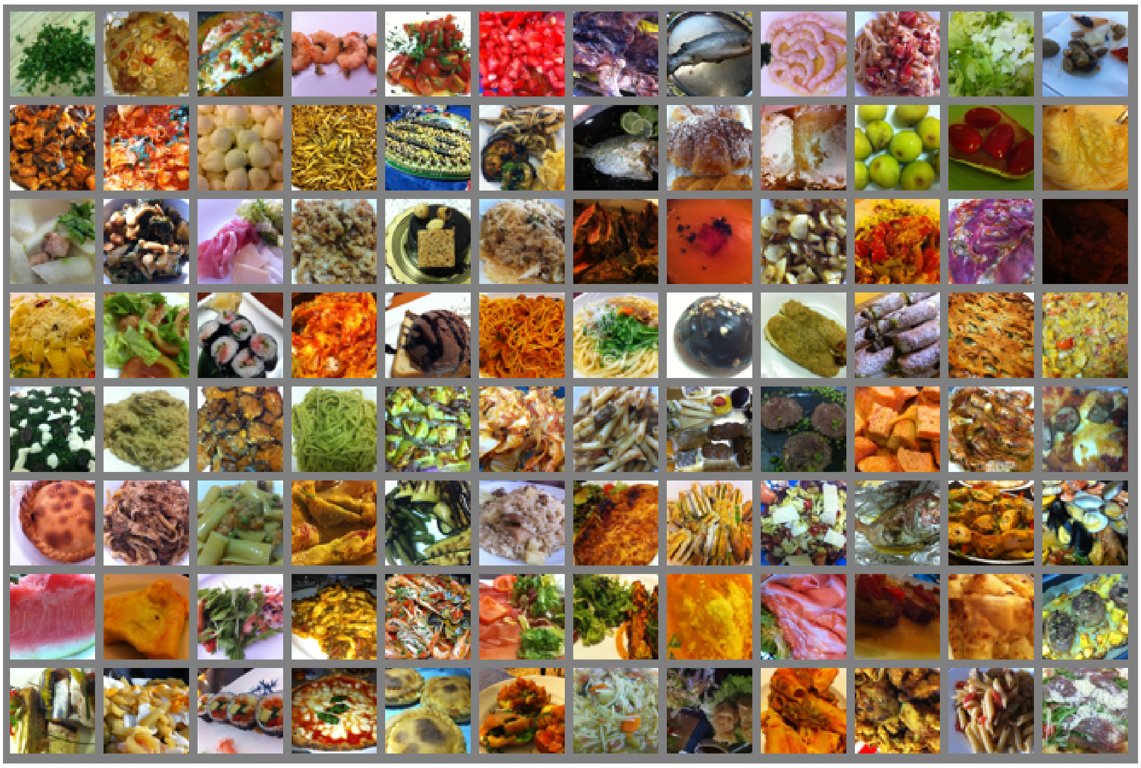


Figure 2.1: A sample of 96 images included in UNIC-FD889 dataset.

the images that do not fit the model are labelled as an “anomaly” with respect to that class. In this study, Bag of Words is employed to represent images by using three different descriptors: Textons, Scale-invariant feature transform (SIFT) and Pairwise Rotation Invariant Co-occurrence Local Binary Pattern (PRICoLBP). One-class classification is performed by using one-class Support Vector Machine (OSVM). To learn about the food class I have used the UNICT-FD889 dataset, since it presents variability and considering that the images are collected in real meal scenario with a mobile phone. Some samples of this dataset is reported in Fig.2.1. I also used two more datasets for testing purpose which can be used as benchmark to compare food vs non-food classification algorithms.

2.1 Proposed Method

I considered the one-class classification paradigm (OCC) for food vs non-food classification problem [102, 103]. One-class classification algorithms learns about the class to be identified assuming that representative training data of all the possible classes are not available or very difficult to obtain. My choice is motivated by the fact that in a training phase could be simple to have example of what a food image looks like, but it is very difficult to define all the images classes related to the non-food class. If one considers the problem of detecting food frames in videos acquired with wearable glasses, the non-food class is composed by all the possible scene that a human can observe in his life. This motivated us to perform a benchmark experiment, where the unique class to be use for learning purpose is the food one. As training data to represent the food space I have used the UNICT-FD889 dataset introduced in [92]. It contains 3583 food images belonging 889 different classes that are taken in real scenarios during meals with an iPhone. (Fig. 2.1). To test the discriminative capability of the approach, two more datasets of food and non-food images respectively composed by 4805 and 8005 images have been considered. These two dataset have been downloaded by Flickr (Figs. 2.2 and 2.3)

I employed three different image descriptors as baseline for the experiments: Bag of SIFT features [57, 33, 104], PRICoLBP [96] and Bag of Textons [57, 95]. These descriptors have been chosen for the good results they exhibited in task as Texture recognition and Food retrieval [95, 96, 92].

For Bag of SIFT I have considered a dense sampling on a grid with spacing of 8 pixels. For each grid point a 16×16 patch is extracted and SIFT descriptor is computed consider the colour domain [105]. The codebook to be used for a Bag of SIFT representation has been obtained through K-Means clustering with $K = 2200$. To obtain the PRICoLBP representation I used the original code provided by the authors at http://qixianbiao.github.io. For both, grey and colour images I set radius 2, neighbour points 8 and template 2. With these parameters I have a 1180-dimensional vector for gray whereas 3540-dimensional vector for RGB images. To compute the Bag of Textons representation I employed the MR8 bank of filter and the Schmidt one, in grey and Lab colour domain. I used a vocabulary of 2200 Textons to represent images.

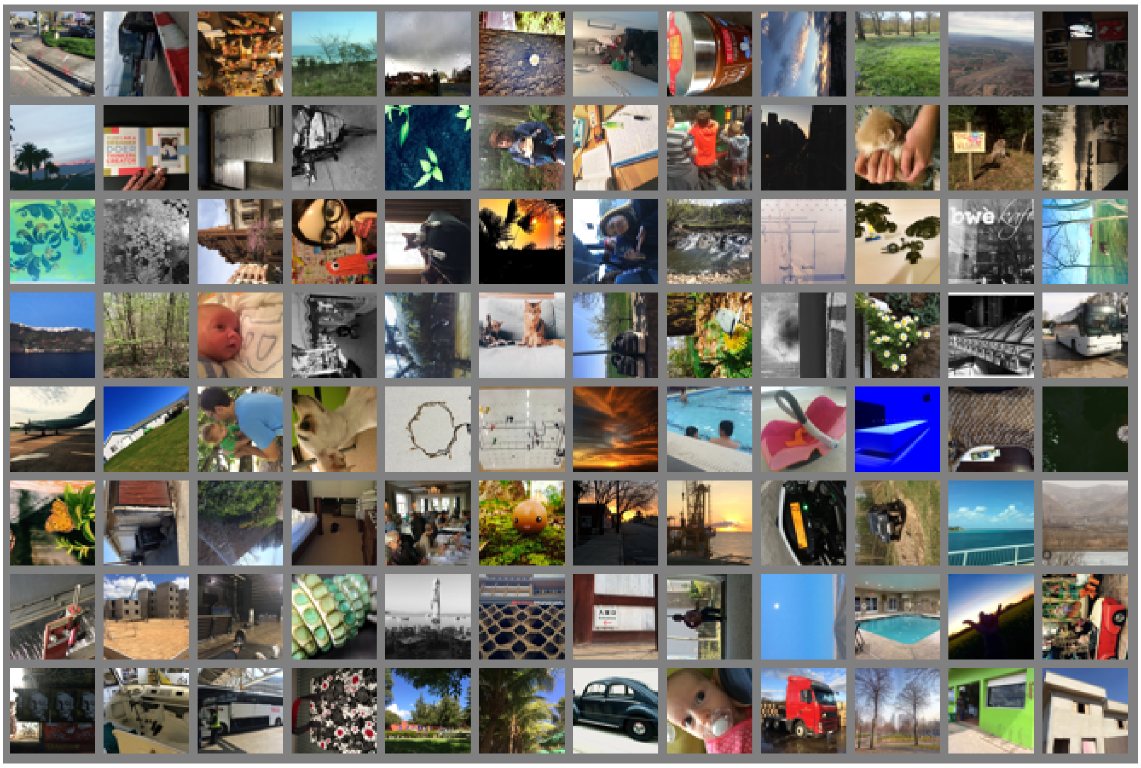


Figure 2.2: A sample of 96 images belonging to the non-food dataset downloaded by Flickr.

2.1.1 One-Class classification

The One-Class Classification (OCC) problem is different from the well-known binary classification problem, since in OCC the negative class is neither present nor properly sampled. OCC can be successfully applied in different scenarios, like detecting machine faults or homepage classification. For instance, a classifier should detect when abnormal or faulty behaviour is occurring in a machine. The samples related to the normal working of the machine (i.e., positive training data) are easy to collect. Conversely, most of the faults will not occur (or will rarely occur), so one will have small or no data for the negative class. In the case of a homepage classifier (is a webpage "homepage" or not?), it is relatively easy obtain samples of homepages (positive examples). However collecting samples of non-homepages (negative examples) can be very challenging. Actually, it could be hard represent the negative concept uniformly and this might involve human bias.

Since the negative data is limited or even absent, in OCC tasks only one side of

Figure 2.3: A sample of 96 images belonging to the food dataset downloaded by Flickr.

the classification/decision boundary can be determined by using the data. For this reason, one-class classification results in a harder problem of conventional multi-class classification. Shortly, it is hard to define how closely the decision boundary should fit in each of the directions around the data, on the basis of one class only.

One-Class Classification has been also referred in literature as Outlier Detection, Novelty Detection or Concept Learning. These terms are related to different application of OCC. In this Chapter is described the first use of OCC for food Vs non-food classification.

Specifically, I have chosen OSVM algorithm since it has been successfully employed in task like: Handwritten Digit Recognition, Information Retrieval, Face Recognition Applications, Medical Analysis, Bioinformatics, Spam Detection, Anomaly Detection and Machine Fault Detection [102]. Differently from other methods, it is able to find an hypersphere that encompasses all the positive samples which belong to the training set (i.e., the food images); then, the new samples geometrically inside the hypersphere are classified as positive, while the outside ones are classified as negatives. Moreover, since it is an SVM-based method, it can exploits kernel trick to efficiently find non-linear boundaries to separate data. Finally, as reported in Chapter 1, SVM-based approaches have shown good performance in many Computer Vision tasks related to the food domain. These facts have motivated my choice.

2.1.2 Bag of SIFT

SIFT algorithm allows to detect visual interest points and describes them such that the final descriptor results invariant to scale, rotation, illumination changes and partially invariant to affine distortion [33, 104]. SIFT are usually extracted sparsely or densely [57, 106] from gray-scale or color images. After SIFT extraction the set of descriptors can be used for matching purposes or to build an image representation based on the Bag of Words paradigm (BoW). I tested both representation approaches in the experimental phase.

To build the BoW SIFT representation we use a dense regular grid to compute the SIFT descriptor of a patch. At this point, a clustering algorithm is used to quantise descriptors space extracted on training images to create a visual words vocabulary. To represent image, each point of the regular grid is associated to nearest visual word. When the visual vocabulary is computed, each image in the training and test set can be represented as a distribution of visual words. A grid with spacing of 8 pixel and a patch of 16×16 is used during dense sampling on the three RGB channels. K-means clustering is exploited to compute the visual words vocabulary with different sizes. The SIFT descriptors are computed independently for each colour channel. A Bag of SIFT is obtained for each color channel and the three visual word distributions are concatenated in a unique descriptor. The final descriptor is an histogram of the SIFT-based visual words, i.e. clustered keypoints. The use of this descriptor is motivated by the fact that I SIFT keypoints are able to describe local patch, and I expect that food images a higher number of certain patches. In my experiments VLFeat [107] library has been used to extract SIFT keypoints.

2.1.3 Pairwise Rotation Invariant Co-occurrence Local Binary Pattern

Pairwise Rotation Invariant Co-occurrence LBP descriptor (PRICoLBP) focuses on encoding spatial co-occurrences and pairwise orientations of the well-known Local Binary Pattern (LBP) features [108]. It preserves the relative orientations of LBP features pairs in order to obtain rotational invariance. To compute the PRICoLBP descriptor, I employed the original implementation provided by the authors which is available online[1]. I exploited PRICoLBP on both gray and color domain. In the experiments I set the radius 2, neighbour points equal to 8 and the template equals to 2. This results in two kinds of PRICoLBP descriptors of 1180 and 3540 components to represent grey and colour images respectively. This descriptor has been chosen because of the best performance achieved in [96] among several descriptors.

2.1.4 Bag of Textons

Textons have been introduced by Julesz as the putative unit for the visual perception during pre-attention processing [109]. A computational model for Textons can be obtained trough the responses of the grey or colour image to a bank of filters [110]. Filter responses of the training images are quantised through clustering procedure. Hence, each cluster centroid can be considered a Texton and a set of them compose a visual codebook [95]. To represent images each filtered pixel is associated with one of the Texton in the codebook considering a similarity metric (I use L^2 distance). Finally, the histogram of the distribution over the different Textons of an image is built. I considered different configurations involved in the Textons extraction pipeline: grey and Lab colour domain; MR8 (Minimum Response 8) and Schmid filter banks. As similarity measure between two Texton distributions, I used the χ^2 distance. A comprehensive discussion about Textons and filter banks is reported in Chapter 3. Because of the good performance achieved in [92] on the UNICT-FD889 dataset, I have been led to employ it in food vs non-food classification.

[1] http://qixianbiao.github.io/

2.2 Experimental Settings

As preprocessing step the images have been resized to 320 × 240 pixels. Moreover, the vocabulary size for both, Bag of SIFT and Bag of Texton has been fixed to 2200. To test this approach two different kind of experiment have been performed. In the first experiment. I have employed two different sub-set of UNICT-FD889 dataset for training and testing purpose. Moreover, for a proper evaluation of different descriptors, the experiments have been repeated three times and the average performance have been reported. To built the test set, a single image for each class with more than 2 elements has been chosen. This results in a set of 728 food images. The entire test set consists of 728 food images plus all the 8005 non-food images downloaded by Flickr. The rest of 2855 food images are used to train the OSVM classifier.

In the second experiment I used the whole UNICT-FD889 dataset to perform training (3583 images). For testing purpose I have used the same dataset of non-food images employed for the first experiment (8005 images) and one more food dataset (4805 images) obtained from Flickr by downloading (and visually reviewing) images with the tag “food”. This experiment is more challenging than the first one since the food images used in the training and the once used in the testing phases look very different. Despite the Flickr images with “food” tag are related to images containing food, these can contain also other objects not belonging to the food class (e.g., sometime the percentage of the pixels related to food are much less than the once of the background and other objects). There is also a huge variability in the scale of the food plates, as well as photometric variability, and there are examples of dishes which never appear into the training dataset. In this experiment I considered only Textons and SIFT descriptors on color domain since they obtained the best performances in the first experiments.

Finally, a coarse grid strategy to find the best parametrization for SVM classifier has been employed. Relying on coarse grid result, I select a Sigmoid kernel with $\gamma = 10^{-5}$ and an OSVM tolerance $\nu = 0.35$. Parameter ν is a value which is used to tune the tolerance towards outliers for OSVM algorithm.

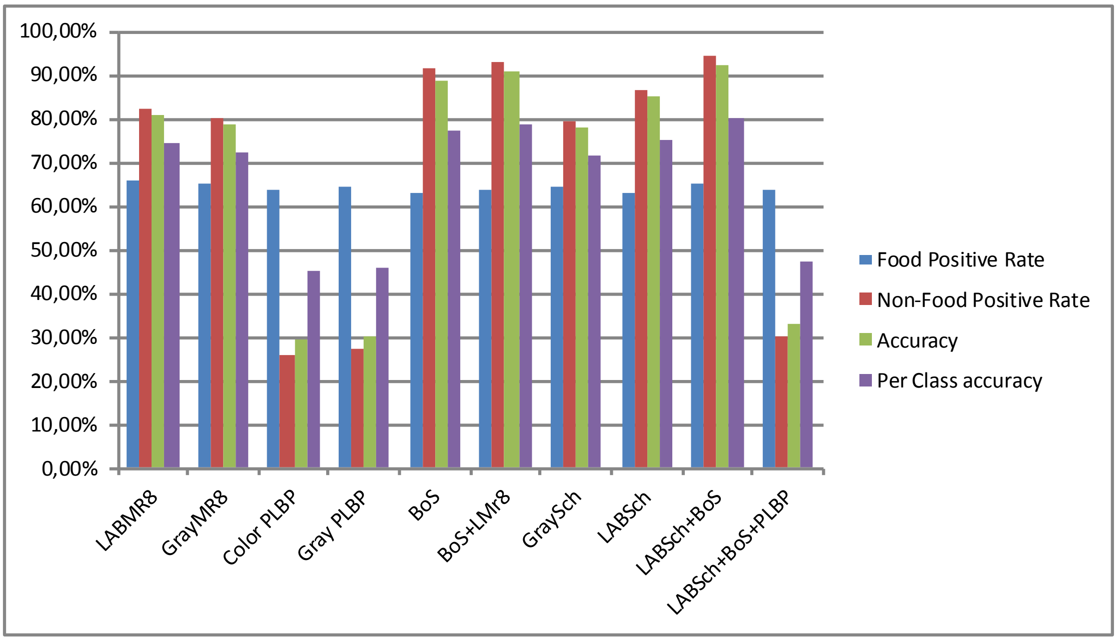
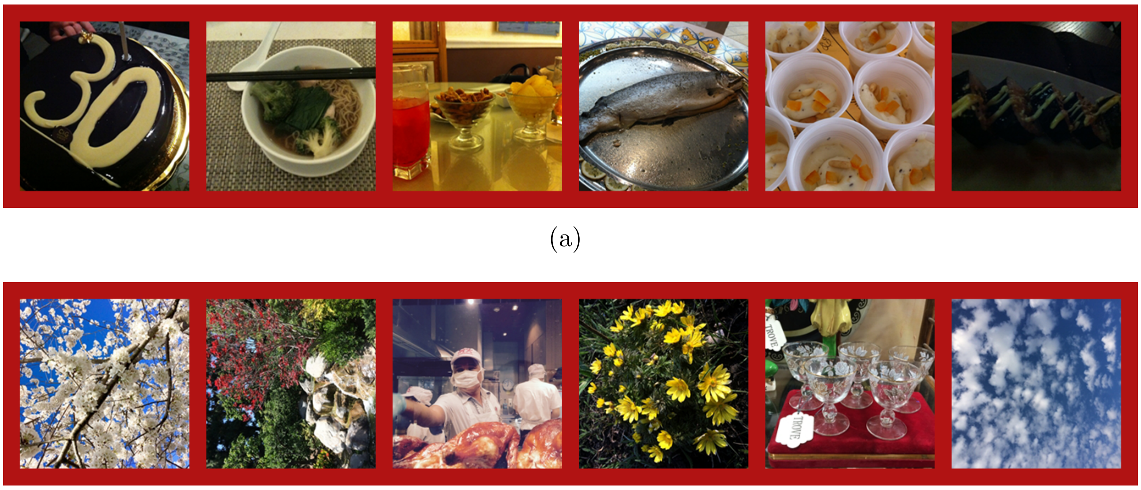


Figure 2.4: The results of food vs non-food classification for the first experimental setting.

2.3 Results and Discussion

The results obtained in this first experiment are shown in Fig. 2.4. Food classification rate is similar for all representations, while non-food classification rate is strongly dependent from the descriptor, and varies between 26.21% and 94.44%. As in [92], Textons outperform PRICoLBP (PLBP). Moreover, the Schmid bank of filters (LABSch) seems to outperform MR8 filters. The best performances are obtained when the Bag of SIFT representation (BoS) is employed. Colour domain helps all the descriptors except PRICoLBP. Since SIFT and Textons capture different image's aspects (i.e., SIFT summarises spatial histograms of gradients, whereas Textons encode textures) I have tested a simple concatenation of Bag of SIFT and Bag of Textons. This test shows an improvement in the discrimination capability (94.44% for non-food and 65.43% for food). The achieved results encourage the usage of multiple descriptors to have a low false positive rate (i.e., very few images of non-food class misclassified as food). Some example of misclassified images are shown in Fig. 2.5.



(a)

(b)

Figure 2.5: Misclassification examples for (a) food class and (b) non-food class related to the first experiment.

As last test I have concatenated Bag of Schmid Textons, Bag of SIFT and PRI-CoLBP (all obtained considering color domain). The results confirm that PRI-CoLBP do not add useful information for food vs non-food classification (Fig. 2.4).

The results obtained in this second experiment are shown in Fig. 2.6). Also in this case seems that combination of the descriptors can help for the task under consideration. Some examples of misclassified images is reported in Fig. 2.7. It is important highlight once more that in the performed experiments images of non-food class have not been used for training the classifier. Considering the results of the two experiments, it is clear that by learning from the food class only it is possible to achieve low false positive rate for food vs non-food classification already with simple image representations. This means that in a possible wearable systems for food monitoring which have to automatically collect food images there will be few outlier to be manually removed by nutrition experts. Note that the trade-off between good true positive rate and low false positive rate can be tuned by the parameters used in one class classification. On the other hand, the classification accuracy of the food class is still to low to be considered useful to monitor the food intake and the behaviour of a person. By considering the two experiments the main observation with respect to this last aspect is that, when the food class to be recognized is

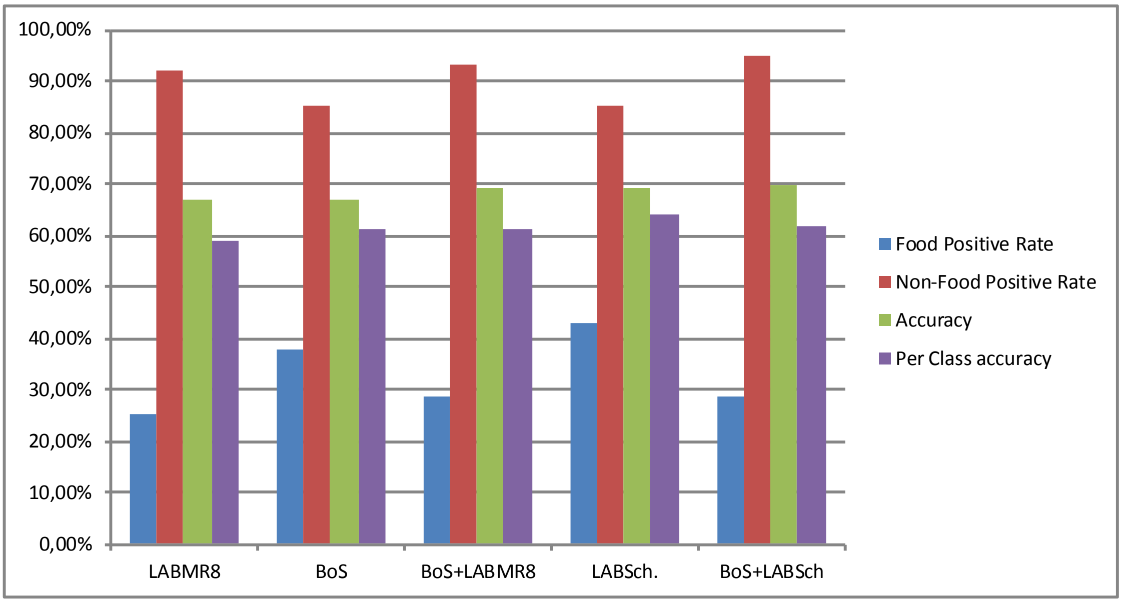
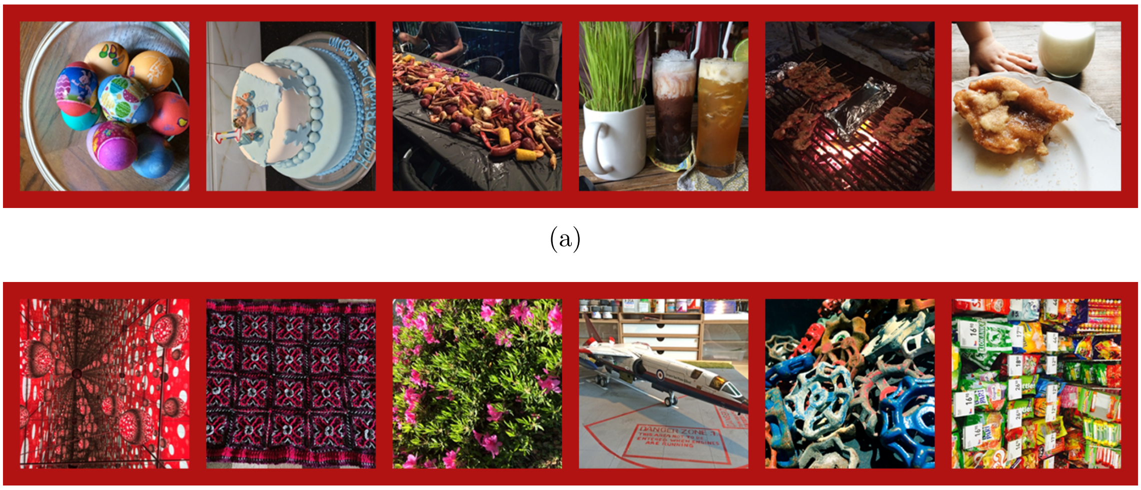


Figure 2.6: The results of food vs non-food classification for the second experimental setting.

represented in the training, the food classification performance is higher (i.e., when images of food used for testing are visually similar to the once in the training, despite high geometric and photometric variabilities). Also, the combination of different features can contribute to have a better discrimination. My conjecture for food vs non-food classification is that by considering a big representative food image dataset, where the food dishes appear in a appropriate scale (i.e., the food plate is the main or the only object, as usually occurs when a user snap a food during meals), and by considering appropriate image representation, the food vs non-food classification become a feasible task.

One more consideration made from the outcome of the experiments is related to the description level to consider the task about of food vs non food classification. Sometime in literature this problem has been called food detection [102]. As demonstrated by the second experiment, when images contain food but also background and other objects, the food vs non-food classification become more difficult. This is mainly because the whole image is considered during food vs non-food classification. On the other hand, a food detector have to be able to localize where the food appears rather than classify the whole image (i.e., draw a bounding box in the part



(a)

(b)

Figure 2.7: Misclassification examples for (a) food class and (b) non-food class related to the second experiment.

of the image in which the food appear discarding the other parts).

Chapter 3

Retrieval and Classification of Food Images

Food has a high variability in appearance and it is intrinsically deformable. This makes classification and retrieval of food and hence an interesting challenge for Computer Vision researchers. The image representation used to automatically understand food images plays the most important role. Despite many approaches have been published (see Section 1.4.4), it is difficult to find works where different techniques are compared on the same dataset. This makes difficult to figure out peculiarities of the different representations, as well as to understand which is the best representation method for food retrieval and classification.

To find a suitable representation of food images it is important to have representative datasets with a high variety of dishes. Although different retrieval and classification methods have been proposed in literature, most of the datasets used so far have not been designed having in mind the study of a proper image representation for food images. Many food datasets are composed by images collected through the Internet (e.g., downloaded from Social Networks), where a specific plate is present just once; there is no way to understand if a specific type of image representation is useful for the classification and retrieval of a specific dish acquired under different points of view, scales or rotation angles. Also the food images collected through the Internet have usually a low resolution and have been processed by the users with artistic or enhancement filters.

In this chapter I address the problem of food image representation for retrieval and classification purposes. Additionally, a new dataset designed for the study of the representation of images is introduced. The proposed dataset, called UNICT-FD1200, is composed by 1200 different food plates acquired by users during real meals. Each food plate has been acquired multiple times and in different light conditions to guarantee both high geometric and photometric variability. Building on top of the work in [92] I employ a bag of words like representation based on Textons features [95] to represent the images of food. Then I present in-depth analysis of the main properties of bag of Textons representation pipeline to point-out which colour domain, bank of filters and vocabulary size are more suitable to tackle the retrieval and classification in this specific domain. I also propose a new image representation building on the perceptual concept of Anti-Textons discussed in [111] by Williams and Julesz. The proposed Anti-Textons features extend Textons by encoding spatial information during feature extraction.

3.1 Proposed Dataset

The research in the field of Computer Vision needs a large amount of organized data in order to test the algorithms for task such as detection, recognition and so on. Unfortunately it is not always easy to collect meaningful data for the different tasks. In particular, in the case of food classification and retrieval for food intake monitoring, can be very difficult build a representative dataset. Actually food comes in many forms and it is naturally deformable, so a representative dataset should contain different variabilities. Moreover it is important whether the data are acquired in real meal scenario rather than collected from the web, where images of food are usually posted to show the best aspect of a dish and, some time, are post processed for this scope. As discussed in previous Chapter 1, different datasets have been proposed in literature. However, most of them are build by collecting images downloaded from internet [53, 58, 88, 89], contain food images acquired with constrained laboratory settings [67, 100] (e.g., variabilities related to light conditions and background are not considered), consider very simple food plates [99], or include only food from one nationality [85].

Considering the aforementioned limitations of the datasets currently available for testing purposes, in one of my previous work I have introduced the UNICT-FD889 dataset [92], which is a collection of food images acquired during real meals, useful for the study of the image representation to be used for food image retrieval purposes. This dataset is available online at the URL http://iplab.dmi.unict.it/UNICT-FD889/.

In this Chapter I presents an extension of UNICT-FD889. Specifically, I included more dishes as well as the labels related to the following 8 categories: *Appetizer, Main Course, Second Course, Single Course, Side Dish, Dessert, Breakfast, Fruit.* Images depicting mixed food (e.g., fish with salad), are labelled with multiple labels (e.g., Second Course and Side Dish). The new dataset is composed by 4754 images related to 1200 distinct dishes of food of different nationalities (e.g., English, Japanese, Indian, Italian, Thai, etc.). Each plate has been acquired multiple times (four in the average) to guarantee the presence of geometric and photometric variabilities. All the food photos have been taken in the last five years during real meals by using a mobile camera in unconstrained settings, such as different backgrounds and light conditions. This is a significant characteristic which is mandatory to test food understanding algorithms on real scenario data. At the best of my knowledge, all the other state-of-art datasets, except UNICT-FD889, include photos retrieved by web in semi-automatic way or acquired under laboratory settings. The mobile cameras used for the acquisition are iPhone 3GS, iPhone 4 and iPhone 5 with a max resolution (e.g., equals to 2448×3264 for the iPhone 5). The UNICT-FD1200 dataset is thought to help research in the field of food-understanding with the aim to study the best representation to use for food images. It can be used to test food image retrieval as well as food classification by considering the aforementioned classes. Fig. 3.1 shows image samples randomly selected from the UNICT-FD1200 dataset, whereas Fig. 3.2 can be useful to assess the multi-view acquisition as well as geometric and photometric variabilities.The UNICT-FD1200 dataset is available for research purposes at the URL http://iplab.dmi.unict.it/UNICT-FD1200/.

3.2 Image Representation

To benchmark the proposed dataset I used the three features employed in Chapter 2: SIFT, PRICoLBP and Textons. I exploited SIFT to represent the food images as set of features to be used together with a matching scheme during classification and retrieval, as well as to build a representation based on the bag of words paradigm.

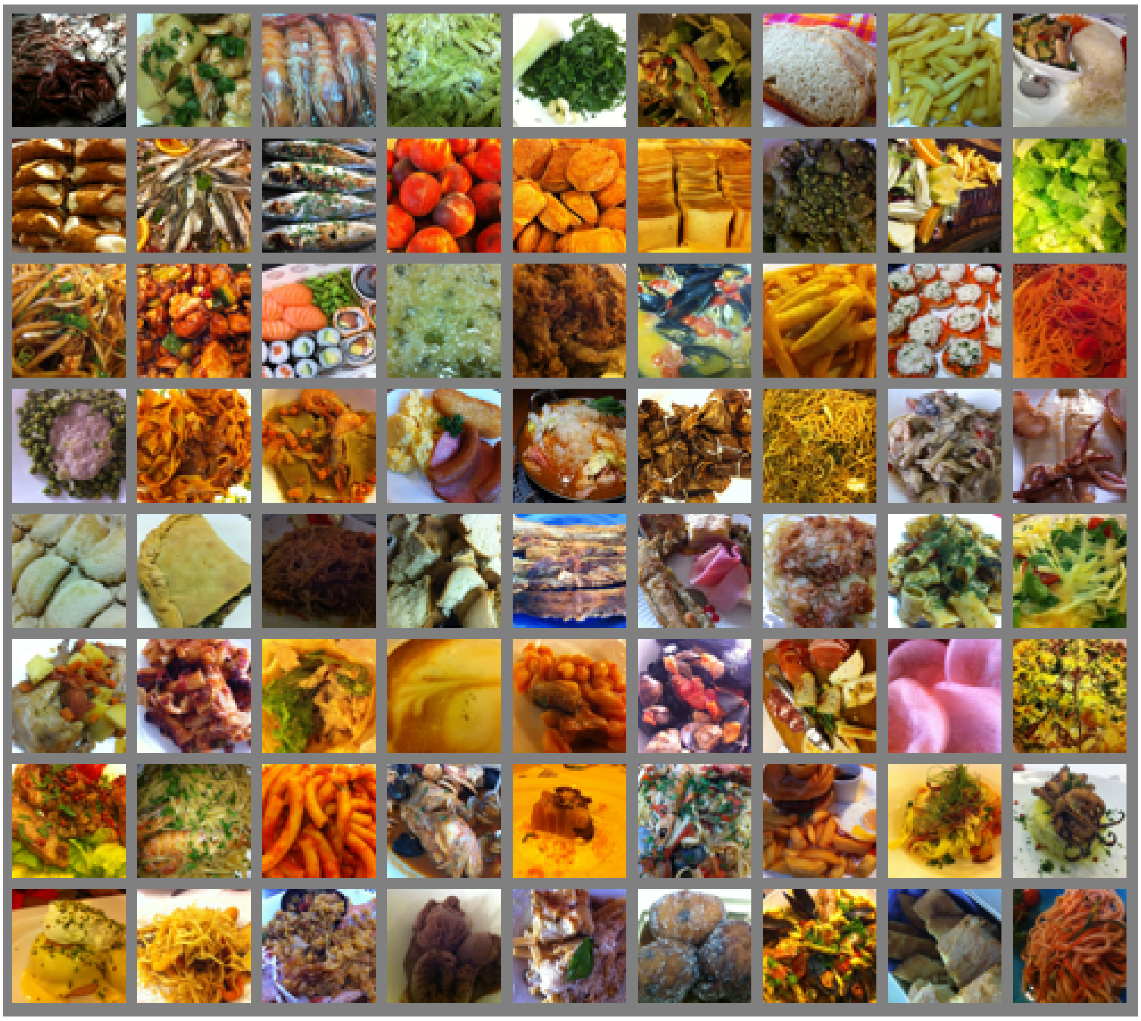


Figure 3.1: A sample of 72 images belonging to UNICT-FD1200

The PRICoLBP features have been included into the comparison since they have been recently proposed and tested on food dataset [96]. I considered Bag of Texton representation because its capability to describe texture information. Despite the simplicity of the Bag of Textons representation, it has obtained good results in the context of food classification and retrieval [92, 75]. Finally, I propose a new image representation based on the perceptual concept of Anti-Textons [111, 112, 113] to encode spaces between Textons. The proposed image representation outperforms all the others approaches. It is important to note that all the aforementioned representation methods are invariant or partially invariant to the illumination. Texton-based

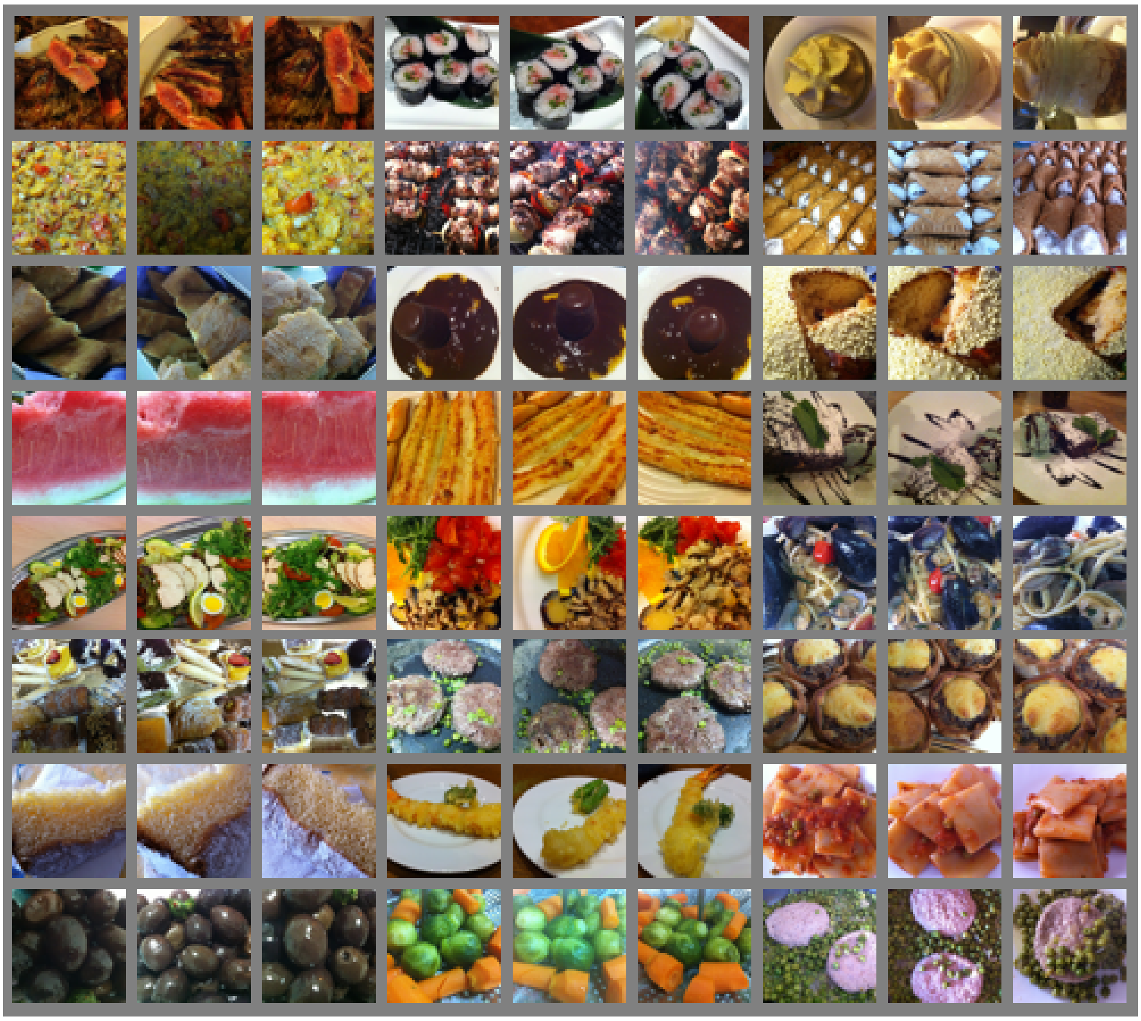


Figure 3.2: A sample of 24 dishes belonging to UNICT-FD1200. For each of them, three instances are reported.

representations perform two normalization steps to strongly reduce the illumination effect (pre-processing normalization at mean 0 and variance 1 and post-processing normalization according with the Weber's law [95]). PRICoLBP is a variation of LBP, which is invariant to global illumination changes [108]. Concerning SIFT, in the extraction process, the last normalization step employed to build the descriptor guarantees linear and non-linear illumination invariance [104]. Consider descriptors with illumination invariance property is mandatory because images into the proposed dataset have been acquired under different light conditions.

3.2.1 SIFT and PRICoLBP Representation

SIFT and PRICoLBP have been you employed with the same modality described in Chapter 2. Additionally, in this study, SIFT has been also tested for matching purposes. In this case the SIFT of a query image are matched to the keypoints of all the images in the training set. The query image is associated to the image of the training dataset with the highest number of matchings. Since the SIFT matching algorithm assigns a score to each matched point based on the quality of the match, I also consider to inversely weight each matched keypoints by taking into account the similarity between the SIFT descriptors of the matched keypoints. I consider both grey and color domain. In the RGB domain the SIFT features are extracted and matched independently on each color channel, then the sum of the matching for the three channels is considered to compute the similarity index.

3.2.2 Bag of Textons

Differently from the work reported in Chapter 2 I considered more different configurations involved in the Textons extraction pipeline to highlight which bank of filters, color domain, normalization procedure and size of the vocabulary are the most appropriate in the application context discussed in this chapter. In the following I details the different 49×49 filter banks tested in this work (LM, MR8, MR4, Schmid) and LINC normalization strategy.

Leung-Malik

The Leung-Malik (LM) filters bank [110] consists of 48 filters (Fig. 3.3), among which smoothing filters, edge detectors and bar detectors. There are 4 Gaussian filters, first and second derivatives of Gaussian at 6 orientations and 3 scales, 8 Laplacian of Gaussian filters. The scale σ of the Gaussian functions is between 1 and 10.

Maximum Response Filter Banks 8

The Maximum Response 8 (MR8) filters are derived by the Root Filter Set (RFS) which consists of 38 filters similar to the LM filters [110]. After the convolution with the 38 filters only 8 response are selected. As in LM filter bank, MR8 contains

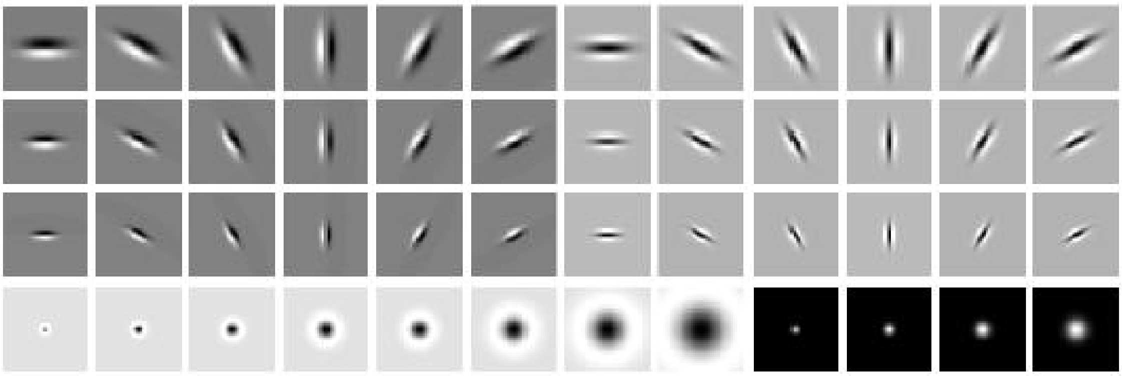


Figure 3.3: The 48 filters of Leung-Malik filter bank.

filters with different scales and orientations. However, only the maximum response is selected across orientations of a specific filter (e.g., edge filter) in order to achieve rotation invariance. The 38 filters consists of a Gaussian filter and a Laplacian of Gaussian filter with scale $\sigma = 10$, first derivative of Gaussian filters at 3 scales and 6 orientations, second derivative of Gaussian filters with the same scales and orientations of the first derivative of Gaussian filters. In Fig. 3.4 is given an example of the extraction of the MR8 responses from the 38 RFS filters.

Maximum Response Filter Banks 4

The Maximum Response 4 (MR4) is a subset of the MR8 filters which is built considering a single scale for the edge filters and bar filters [110]. Hence the filter bank to be applied contains 14 filters but 4 responses only are selected to get rotation invariance (In Fig. 3.4).

Schmid Filters

The Schmid filter bank [114] consists of 13 isotropic filters defined by the equation:

$$F(r, \sigma, \tau) = F_0(\sigma, \tau) + \cos\left(\frac{\pi \tau r}{\sigma}\right) e^{-\frac{r^2}{2\sigma^2}} \tag{3.1}$$

where σ is the filter scale in pixel and τ a value which is proportional to the number of concentric rings in the kernel. $F_0(\sigma, \tau)$ is added to obtain a zero DC component

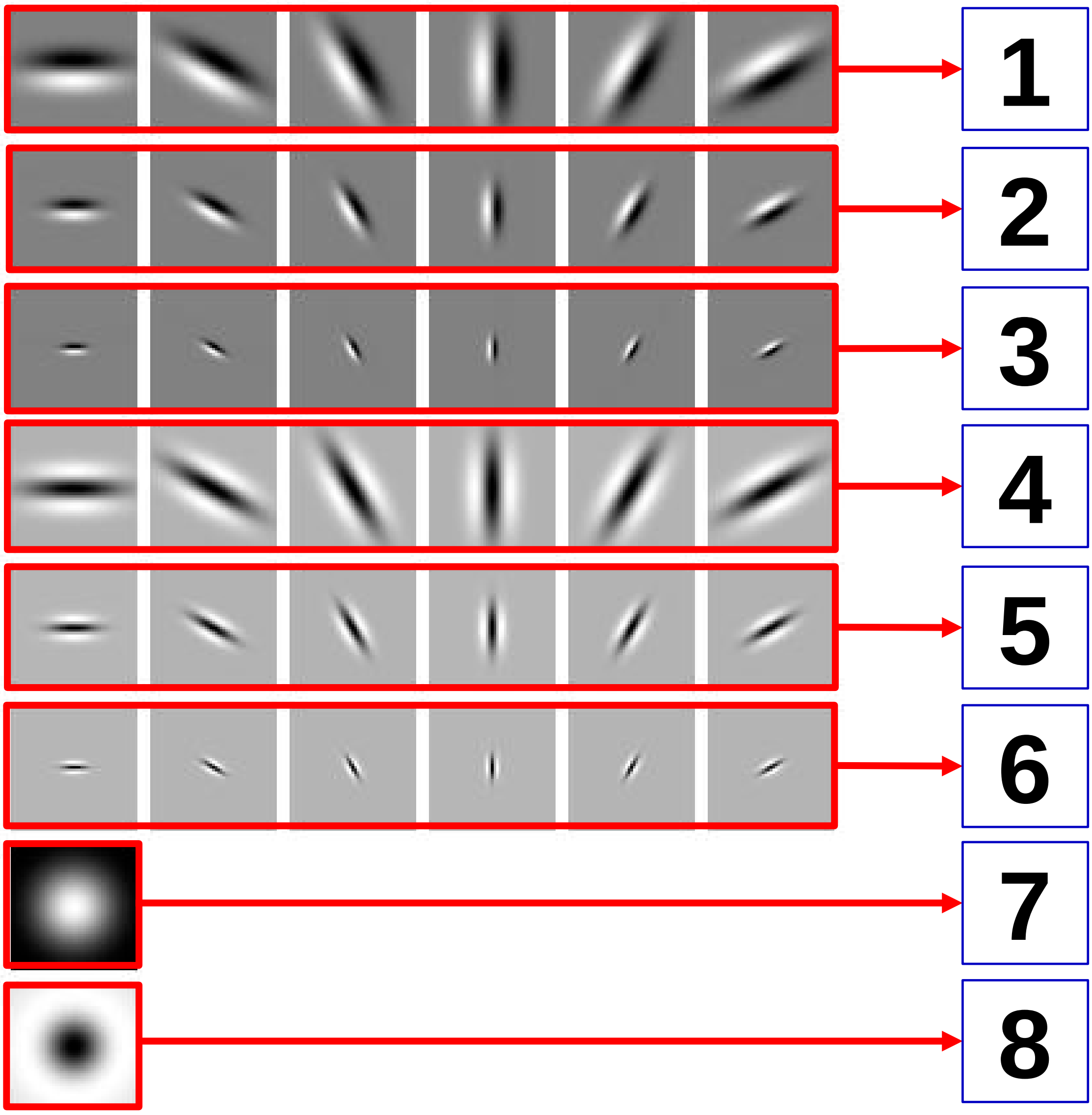


Figure 3.4: An example where a 49×49 patch from which the MR8 responses are computed. For each orientation of a specific type of filter, only the maximum response is chosen.

for the filter with (σ, τ) pair taking values (2,1), (4,1), (4,2), (6,1), (6,2), (6,3), (8,1), (8,2), (8,3),(10,1), (10,2), (10,3) and (10,4). Those filters are shown in Fig. 3.6.

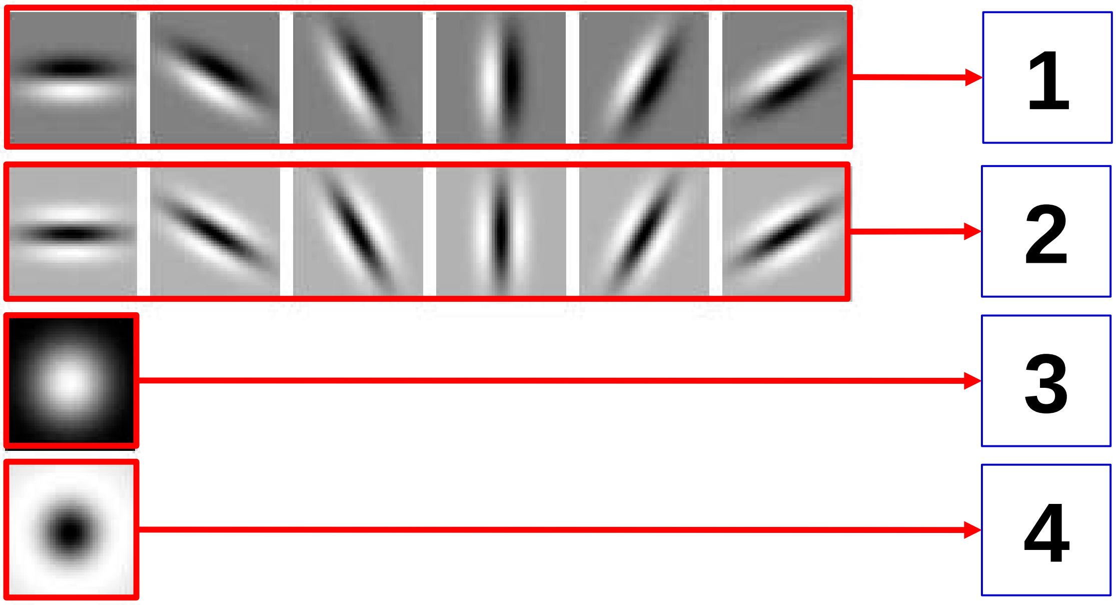


Figure 3.5: The filter responses are collapsed to guarantee the rotational invariance obtaining MR4 responses.

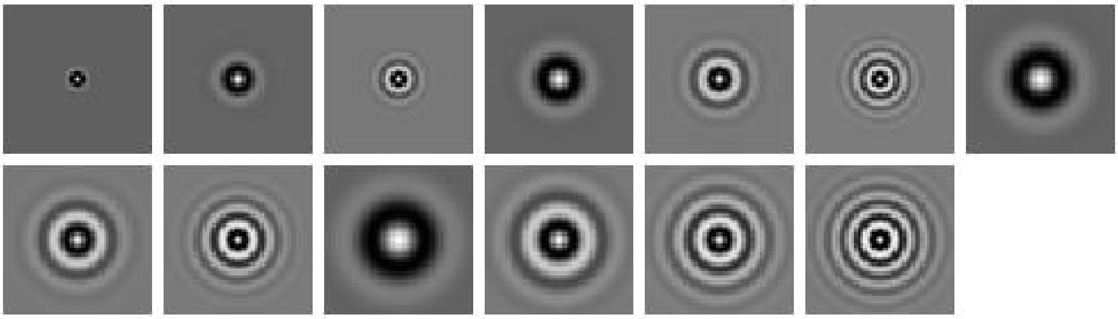


Figure 3.6: The 13 Schmid isotropic filters.

Local Intensity-normalized Colors (LINC) Filters

To achieve invariance to local intensity changes the Local Intensity-normalized Color procedure has been proposed in [115]. The authors proposed to use opponent color space and a normalization of the filter responses. Specifically, for each filter response the Gaussian filter response for first channel at the same scale σ is exploited in order to obtain local intensity normalization. Despite LINC normalization has been proposed for MR8 filter bank, I have employed the procedure considering both MR8 and Schmid bank of filters.

3.2.3 Anti-Textons descriptor

Bag of Textons representation has shown good performances in the context of food classification and retrieval [92]. However, this representation does not take into account the spatial relation between the visual words. This is because, in the bag of words paradigm, only the first order statistics of the visual words are used as image descriptor. I propose to exploit the spatial information around each Texton to build a more discriminative representation of food images. This idea is supported by the study presented in [113] where the authors defined the concept of anti-textons as the space between two Textons. Anti-Textons concept have been introduced in literature by Williams and Julesz in [111, 112] for the purpose of texture segregation (i.e., segmentation).

At the best of my knowledge there is only a single attempt to find a suitable computational procedure to compute Anti-Textons for texture segmentation [113]. Differently from previous works I introduce a computational approach to compute Anti-Textons distribution for the purpose of image representation. The proposed method assumes that a textons vocabulary with N codewords has been obtained from the set of training images. Once the visual vocabulary is obtained, the Anti-Textons computation pipeline shown in Fig. 3.7 is applied to represent an image. The Anti-Textons representation is computed considering the following steps:

- The Textons map for an image I is computed. For each pixel the Textons map store the corresponding Texton ID.
- For each Texton with ID i ($i = 1, ..., N$) a binary map is produced. The binary map B_i for the Texton i, contains 1 in the position where the Texton i occurs and 0 in all the other positions. At this stage, N binary maps are computed.
- The Distance Transform [116, 117] for each map B_i is computed. This results in a "saliency" map where the points close to the Texton i are less salient than the further ones. I use this saliency map to establish how much each Textons into the Textons map can be considered Anti-Textons with respect the Texton

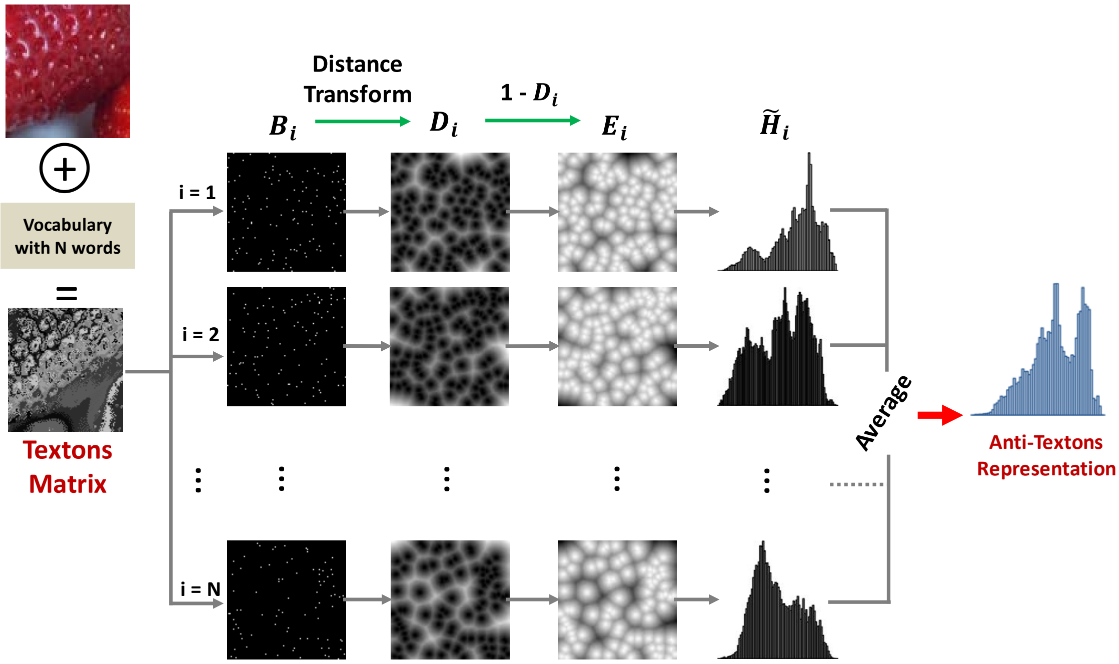


Figure 3.7: Anti-Textons representation pipeline (see text for details).

i. Each saliency map is normalized by dividing by its max value. I refer to the normalized map for Texton i with the symbol D_i. The maps D_i are inverted by computing $E_i = \mathbf{1} - D_i$. This is the way I encode the space between two Textons of the same class i.

- As next step each map E_i is used to weight the original Texton map to obtain the final Anti-Textons distribution for Texton i. In particular, I compute the histogram H_i as follows: $H_i(k) = \sum_{\mathbf{x}} E_i(\mathbf{x})B_k(\mathbf{x})$, where $\mathbf{x}$ is the coordinate in the Texton Map. The normalized histogram $\widetilde{H_i}$ (at sum 1) represents the Anti-Textons distribution for the Texton i.

- Finally, average all the N computed histograms $\widetilde{H_i}$ in order to produce the Anti-Textons representation for the image I.

The experiments confirm that the proposed Anti-Textons representation outperforms the other representation.

3.3 Experimental Settings and Results

This Section focuses on the settings and the quality measures used to compare the image representations presented in the previous Section and tested on the dataset proposed in Section 3.1. I performed both, retrieval and classification tests. For retrieval purpose, all the 4754 images of the UNICT-FD1200 dataset have been resized to 320 × 240 pixels. For a proper evaluation of the representation methods, I performed the experiments three times with different training sets and test sets. All results are obtained by averaging among the three different tests. All the representation approaches are compared by using the same training set and test sets. To build a training set I selected a single image for each of the 1200 dishes. Hence, a training set is composed by 1200 different images. The intersection between the three training sets is empty. For each test set, I used the rest of the images. The dataset, as well as details useful to proper replicate the experiments with the considered training and test sets are available at URL http://iplab.dmi.unict.it/UNICT-FD1200/. For each image representation, the results are obtained by averaging over the three runs. In the case of the retrieval a run consists in a group of queries composed by the test images for which we need to find the corresponding image in the training set. The retrieval performances are measured using the quality metric $P(n)$ which is based on the top-n criterion:

$$P(n) = \frac{Q_n}{Q} \tag{3.2}$$

where Q is the number of queries (test images) and Q_n the number of correct queries among the first n retrieved images. In this case $P(1)$ results in the classification accuracy measure of the system. As index to describe the whole retrieval result I decided to use the Mean Average Precision (MAP) described in [118].

For classification purposes, I consider the same three training and test sets employed for retrieval purpose and a $1-NN$ classifier with χ^2 distance. Because of the images can have multiple labels (up to 2 labels), two performance metrics have been considered: as first measure, the intersection between the labels of the query image and the labels of the nearest retrieved images (according to the $1-NN$ criteria). If the intersection is not empty, we count a positive match. Only the overall accuracy is computed in this case. For the second classification test, all the multi-labeled images are removed. In this way, the training set is reduced from 1200 to about 965 images and the test set from 3479 to 2799. With a single label, it is possible to build a standard confusion matrix for evaluation purpose. In the following subsections I detail both the performed experiments and the obtained results.

To prove the validity of the proposed descriptors I have done some experiments on the UNICT-FD889 and I compared the results with the one reported in [92] and [119]. Finally, a comparison on a different dataset (i.e., Menu-Match [120]) and a test with CNN features have been performed.

3.3.1 Global Textons Vs Class-Based Textons

Bag of Textons representation obtained in two modalities has been tested: class-based and global. For the class-based representation I consider each image in the training set as a class because it is related to a specific plate. Then, 10 Textons per image have been extracted by using K-Means algorithm to quantise the space related to the considered categories. Hence, the vocabulary can be build by collecting all the extracted Textons. Since the training set is composed by 1200 images, the vocabulary contains 12000 visual words. In the global approach all the filter responses of the training set are considered to build the final vocabulary through K-Means clustering with K=12000. I have performed several test by using MR4 filter banks in gray domain and different vocabulary size for the global approach. The results (Table 3.1) show that there is no meaningful difference between the class-based approach and the global one. Since the construction with the global approach allows to perform tests at varying of the final vocabulary in a simple way, I have chosen this modality to build the visual codebook for the all other experiments presented.

3.3.2 Gray Textons Vs Color Textons

As next experiment, I decided to compare Textons representation in gray domain with respect to the one obtained considering RGB domain. To this aim, I choose to apply the MR4 filter bank to each color channel and then concatenate the responses obtained for the difference channels. Hence, considering the MR4 filters we obtained features in 4-dimensional space for the gray domain and features in a

Table 3.1: Accuracy and Mean Average Precision for Global Textons VS Class-based Textons by using MR4 filters in gray domain.

Representation		**Accuracy**	**mAP**
12000	Textons (MR4) - Gray - Class Based	**29,16%**	**36,93%**
12000	Textons (MR4) - Gray - Global	28,94%	36,56%
6000	Textons (MR4) - Gray - Global	28,56%	36,39%
3000	Textons (MR4) - Gray - Global	28,30%	36,06%
1500	Textons (MR4) - Gray - Global	28,41%	36,33%
750	Textons (MR4) - Gray - Global	28,16%	35,99%
375	Textons (MR4) - Gray - Global	27,73%	35,58%

12-dimensional space for the color domain. The P(n) graph in Fig. 3.8 shows that a great improvement has been achieved by using color information. For this reason, I guess that the color information is critical for a good representation of food images. Considering P(1), which correspond to the recognition accuracy of the system, the gray representation obtains 28.94% whereas considering color domain an accuracy of 68.14% is obtained.

3.3.3 SIFT based representation

I test SIFT descriptor in both, gray and RGB color domain. To retrieve images, I have used two similarity measures. The first one is based on the number of matched points, while in the second one each matching is weighted by taking into account the matching quality score. The approach with weighted measure outperforms the one where only the number of matched points are considered. Also in this case, the plots in Figs. 3.9 and the Table 3.2, show that the descriptors in color domain outperform the gray ones for both the SIFT measures employed. Considering the weighted measure in color domain, I obtained the best accuracy for SIFT based representation, that is 63.52%. Nevertheless, this result does not outperform the previous results obtained with MR4 filter bank in color domain.

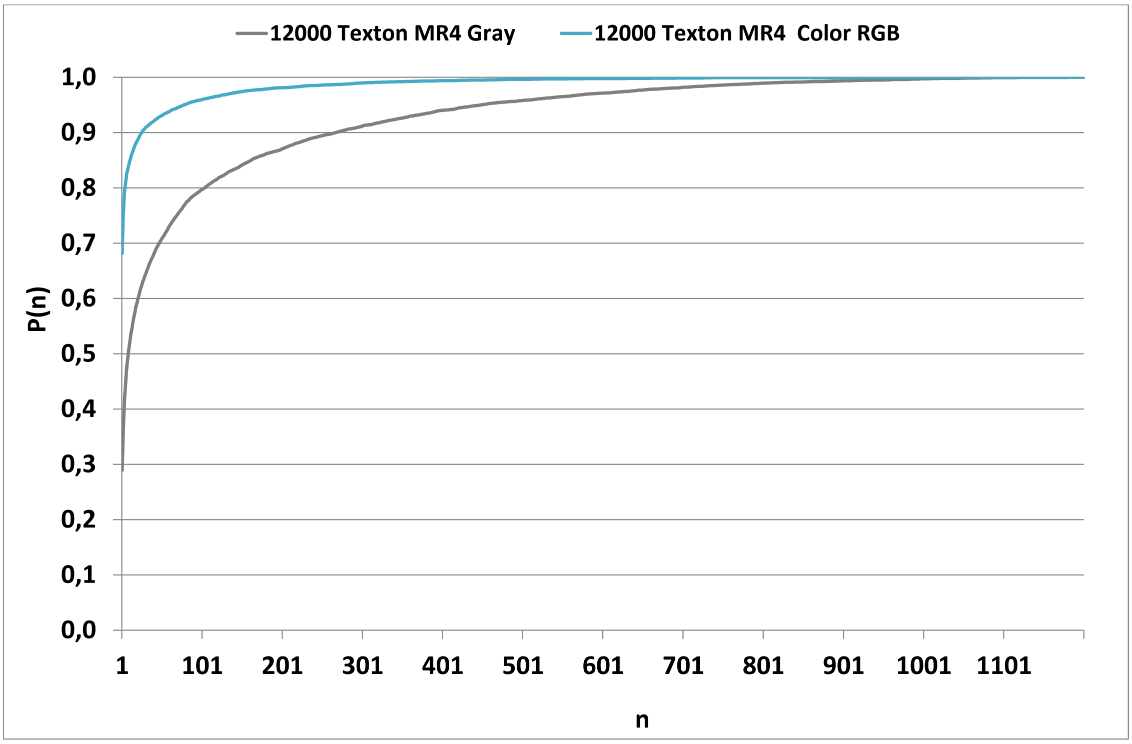


Figure 3.8: P(n) curves related to Gray Textons and RGB Color Textons.

Table 3.2: Accuracy (P(1)) and Mean Average Precision for SIFT matching approach and SIFT matching with weighted scheme approach.

Representation	Accuracy	mAP
SIFT - RGB - Score	**63,52%**	**67,30%**
SIFT - RGB - Match	61,15%	64,46%
SIFT - Gray - Score	54,26%	57,73%
SIFT - Gray - Match	51,67%	54,78%

3.3.4 PRICoLBP based representation

This descriptors can be described as a histogram of CoLBP pattern to encode textures in a rotational invariant way. Since, PRICoLBP has been used for food classification with promising results [96], I take into account it in the comparison. Result are presented in Fig. 3.9. PRICoLBP in color domain is better than PRICoLBP in gray domain. However once again the best results are still obtained using Bag of Textons approach with MR4 filter bank and 12000 visual word in RGB color

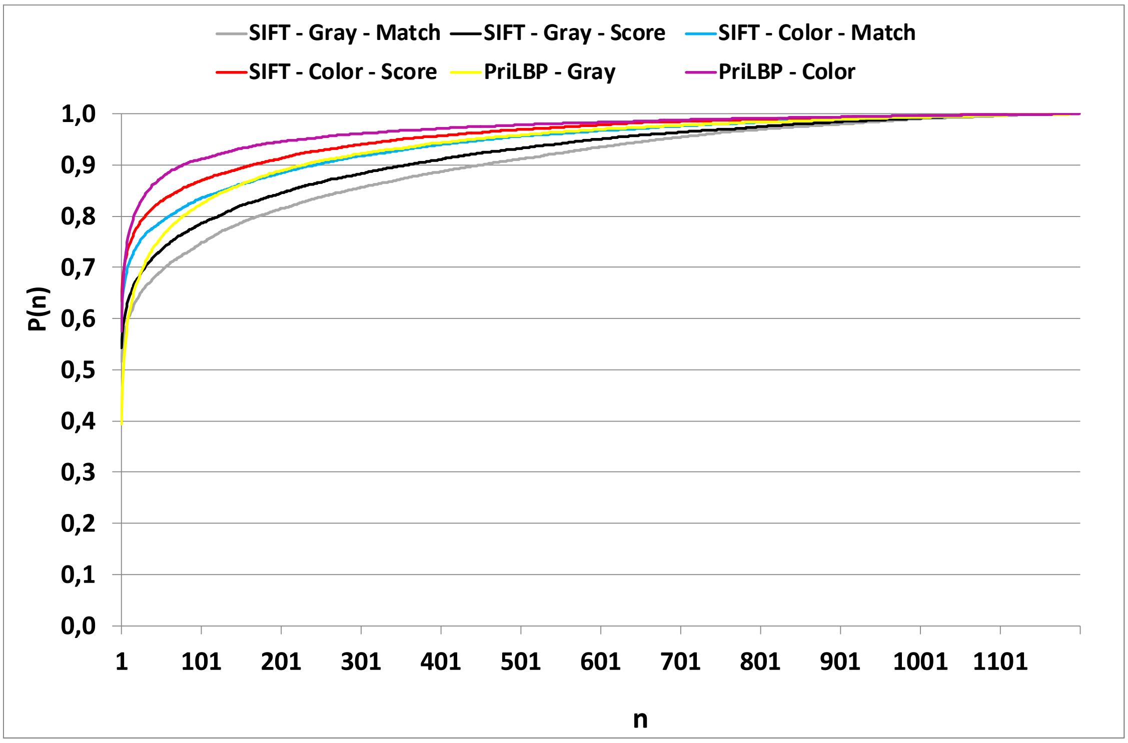


Figure 3.9: P(n) curves for SIFT matching approaches, SIFT matching with weighted scheme approach and PRICoLBP features in gray and RGB color domains.

domain. Hence, I decide to to focus on Bag of Textons representation for the next experiments.

3.3.5 Dimension of the visual vocabulary

The vocabulary size is one of the parameters of the retrieval system to consider to better understand the retrieval performances when the number of visual words used to represent the food images is reduced. For this purpose, I performed tests by using MR4 filter bank in RGB color domain with different numbers of visual words: 12000, 6000, 3000, 1500, 750, 375. In Table 3.3 are reported the performances of the tests where the number of visual words is reduced. Despite the retrieval accuracy decrease, no high drops are observed. This is reasonable because when the vocabulary is reduced, some discriminative visual words could be lost. Nevertheless a smaller vocabulary results in a better use of the resources (e.g., memory, CPU). However

Table 3.3: Accuracy (P(1)) and Mean Average Precisionfor different vocabulary size for the Bag of Textons representation considering MR4 filters and colour domain.

Representation		Accuracy	mAP
12000	Textons (MR4) - RGB - Global	**68,14%**	**73,99%**
6000	Textons (MR4) - RGB - Global	66,60%	72,65%
3000	Textons (MR4) - RGB - Global	65,48%	71,62%
1500	Textons (MR4) - RGB - Global	63,03%	69,69%
750	Textons (MR4) - RGB - Global	60,53%	67,50%
375	Textons (MR4) - RGB - Global	56,58%	63,91%

Table 3.4: First P(n) values ($n = 1 \ldots 10$) related to Bag of Textons representation obtained with different filter banks in RGB domain.

Representation		P(1)	P(2)	P(3)	P(4)	P(5)	P(6)	P(7)	P(8)	P(9)	P(10)
12000	Textons (MR4) - Color - Global	68,14%	74,16%	77,17%	79,30%	80,70%	81,80%	82,79%	83,41%	84,02%	84,70%
12000	Textons (MR8) - Color - Global	71,55%	77,41%	80,20%	81,81%	83,11%	84,21%	85,07%	85,77%	86,33%	86,84%
12000	Textons (Schmidt) - Color - Global	**75,74%**	**80,79%**	**83,16%**	**84,43%**	**85,68%**	**86,68%**	**87,49%**	**88,10%**	**88,68%**	**89,20%**
12000	Textons (LM) - Color - Global	61,69%	68,24%	71,59%	73,69%	75,35%	76,63%	77,79%	78,93%	79,73%	80,56%

for the next comparisons, I decided to use the vocabulary size that guarantee the best performance (12000 words).

3.3.6 Filter banks

In the performed test I considered Bag of Textons representation in RGB domain by using three more filters banks: MR8, LM and Schmid (see Section 3.2 for details). Tables 3.4, 3.5 and Fig. 3.10 report an improvement for MR8 and Schmid filters banks with respect to MR4. On the other hand the LM filter bank has shown the worst performances. I guess this is because Leung-Malik set is not rotationally invariant. This idea is coherent with the best performance obtained with the Schmid set, which consists of 13 symmetric filters. The retrieval system employing Schmid bank of filters in RGB color domain obtained an accuracy of 75.74% and a MAP of 80.43%

3.3.7 Bag of SIFT Vs Bag of Textons

For a proper comparison between Textons features and SIFT features I decided to test the Bag of Words paradigm using SIFT descriptors with a vocabulary size of

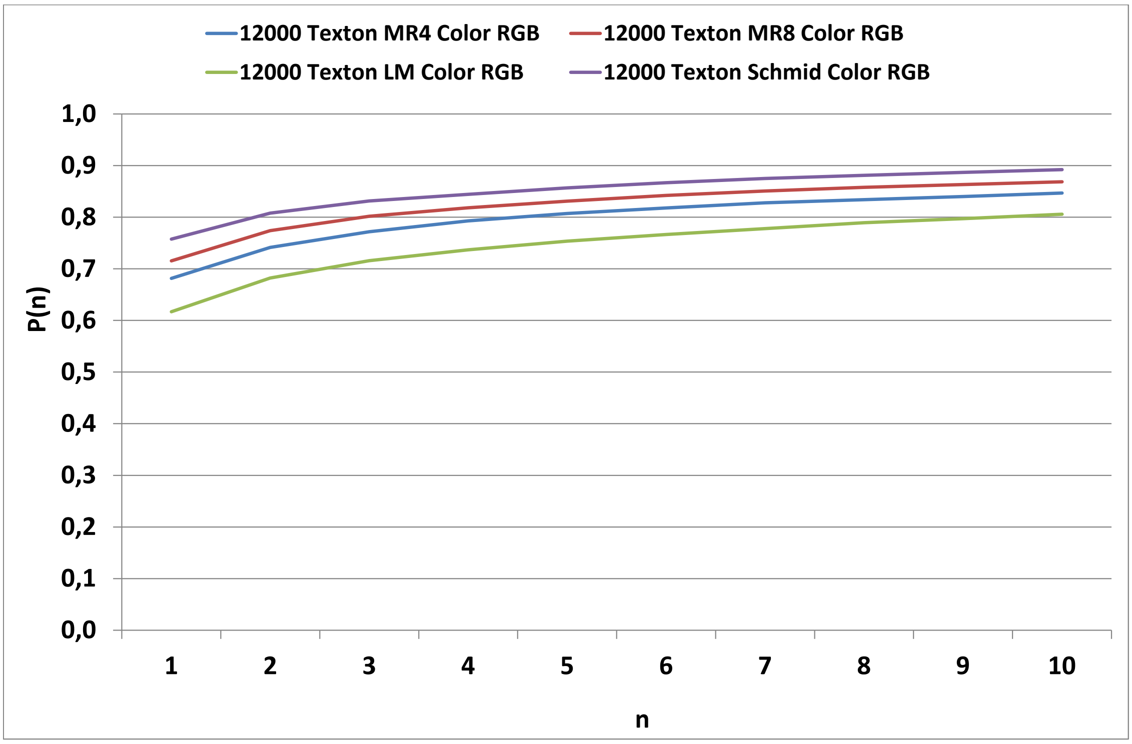


Figure 3.10: P(n) curves related to Bag of Textons representation obtained with different filter banks in RGB domain.

Table 3.5: Accuracy and Mean Average Precision related to Bag of Textons representation obtained with different filter banks in RGB domain.

Representation		Accuracy	mAP
12000	Textons (MR4) - RGB - Global	68,14%	73,99%
12000	Textons (MR8) - RGB - Global	71,55%	77,00%
12000	Textons (Schmidt) - RGB - Global	**75,74%**	**80,43%**
12000	Textons (LM) - RGB - Global	61,69%	68,22%

12000. Considering the work [105] where Bag of SIFT have been used for food classification purpose I used a dense sampling on a grid with spacing of 8 pixels. A 16×16 patch is extracted and SIFT descriptor is computed considering the three RGB channels as described [105]. To make more fair the comparison with respect to the Bag of Textons representation I repeated the Bag of Textons tests by using MR8 bank of filters, color domain, 12000 visual word but considering the same 8×8

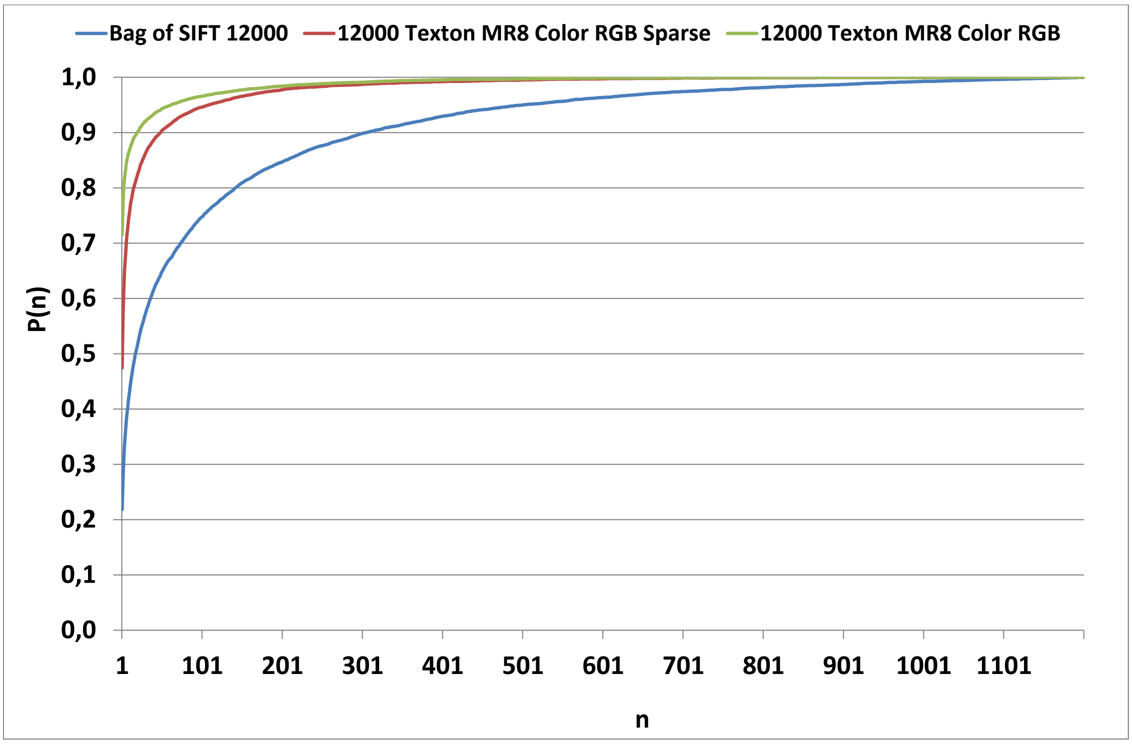


Figure 3.11: P(n) curves for Bag of Textons and Bag of SIFT representations in RGB domain.

Table 3.6: Accuracy and Mean Average Precision for Bag of Textons and Bag of SIFT representations in RGB domain.

Representation		Accuracy	mAP
12000	Textons (MR8) - RGB - Global	**71,55%**	**77,00%**
12000	Bag of SIFT	21,81%	29,14%
12000	Textons (MR8) - RGB - Global - 8×8	47,45%	57,00%

sampling used for SIFT descriptors. The results in Fig. 3.11 and Table 3.6, show that Bag of Textons approach without spatial sampling Bag of SIFT representation. It is interesting to notice that bag of Textons approach outperforms with a large margin Bag of SIFT also when spatial sampling is used.

Table 3.7: Accuracy and Mean Average Precision of Bag of Textons representation with different color spaces.

Representation		Accuracy	mAP
12000	Textons (MR8) - Lab - Global	85,04%	88,39%
12000	Textons (MR8) - LINC - Global	83,10%	86,93%
12000	Textons (MR8) - RGB - Global	71,55%	77,00%
12000	Textons (Schmidt) - Lab - Global	**87,44%**	**90,06%**
12000	Textons (Schmidt) - LINC - Global	84,32%	87,84%
12000	Textons (Schmidt) - RGB - Global	75,74%	80,43%

3.3.8 Color Space

Finally, I consider to change the color space used into the Bag of Textons representation to achieve further improvements in the retrieval performances. To this aim, I exploited the L*a*b* color space and the opponent color space. In the first case, we simply transform the pixel value of an image from the RGB color space to the L*a*b* one. The Textons are computed in the standard way, as described in 3.3.2. In the second case, we use the opponent color space and the normalization procedure described in [115]. In particular, it has been considered the procedure called Local Intensity-normalized Colors (LINC). The normalization is made by dividing each filter response by the Gaussian filter response (with the same σ value). I tested the MR8-LINC method proposed in [115]. Moreover I have adapted the algorithm has been adapted to extract the LINC version of the Schmid filter banks (Schmid-LINC). As shown in Table 3.7, the best performance are achieved using Schmid filter banks computed in the L*a*b* color space with an accuracy of 87.44% and a MAP equal to 90.06%.

3.3.9 Visual Analysis

In order to understand the different discriminative capabilities among the employed representations, I performed a visual analysis of the results. For this purpose, we have included 5 representations in the analysis: Bag of Textons computed with Schmid filters in L*a*b* color space and 12000 visual words, MR8 filters in L*a*b* with 12000 visual words, MR8 in RGB space with 12000 visual words and sparse sampling with step 8, Bag of SIFT, and SIFT based representation with matching

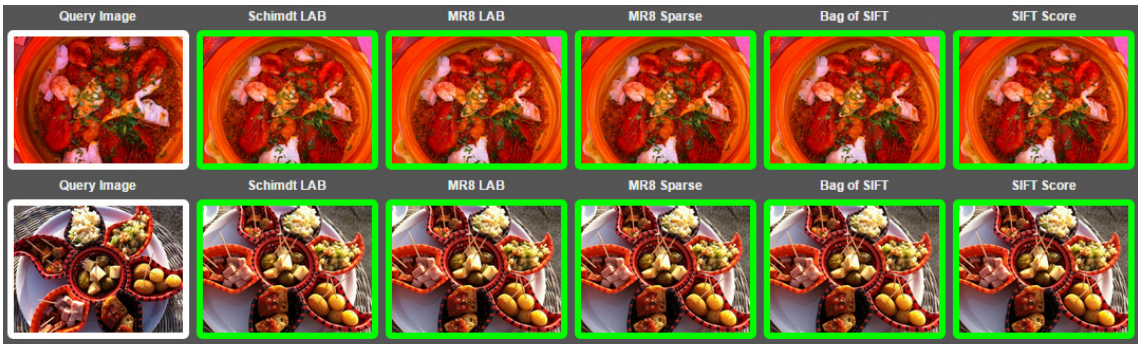


Figure 3.12: A visual comparison where all the considered representations have a positive match.

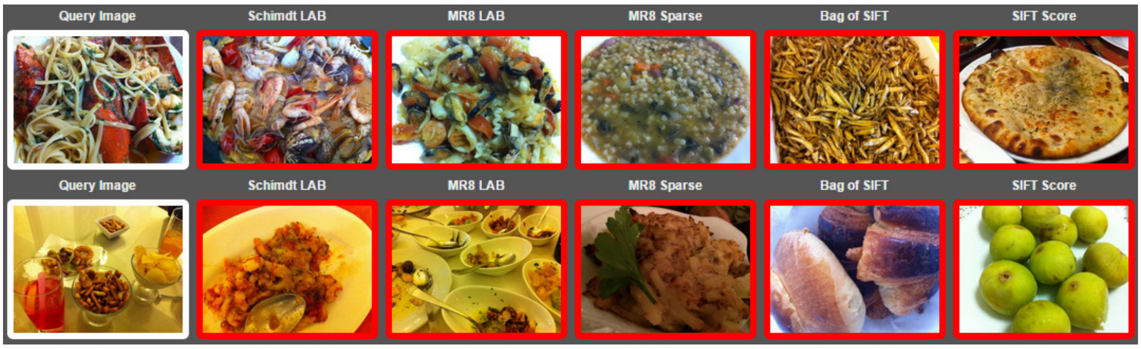


Figure 3.13: A visual comparison where all the considered representations fail.

scheme. Here some interesting result of one of the three test for all the 5 representation have been reported The complete visual comparison,is available at the URL http://iplab.dmi.unict.it/UNICT-FD1200/. In Fig. 3.12 are shown two queries where all the representations have a positive match. On the contrary, in Fig. 3.13, are shown queries where all the representations fail. Since I find out that the Schmid based representation outperforms all the other ones, we selected some queries where this representation had a positive match but all the other ones fail (Fig. 3.14). In Fig. 3.15 are shown, the only 2 queries where the Schmid based representation fails whereas all the other ones have a correct match.

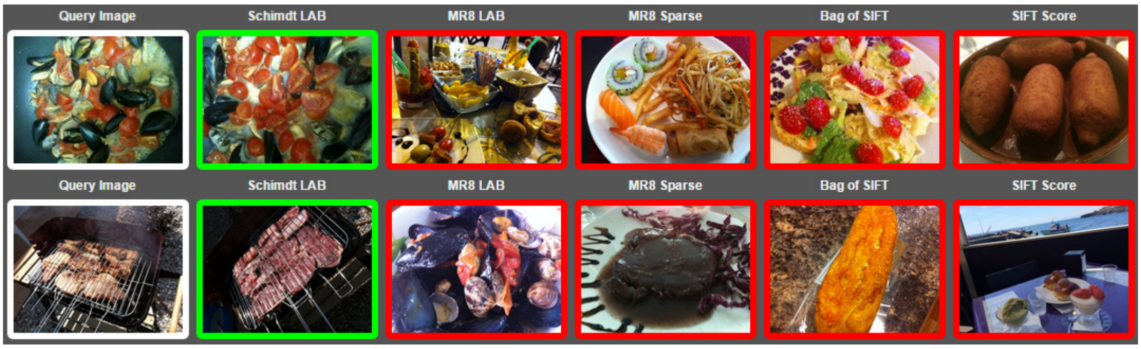


Figure 3.14: A visual comparison where only the Schmid L*a*b* representation gives a correct match.

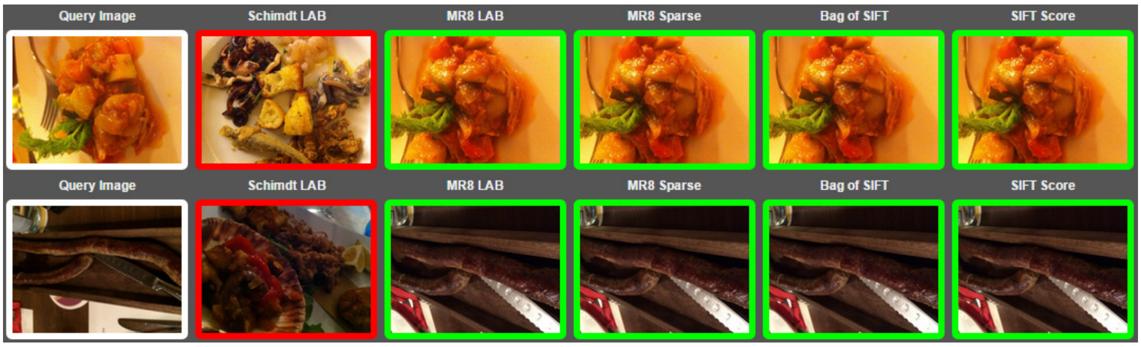


Figure 3.15: The only 2 queries where the Schmid L*a*b* representation fails.

3.3.10 Experiments on the UNICT-FD889

To compare the results reported in [92] with the ones of this chapter, I perform an experiment on the UNICT-FD889 dataset using the representation which has obtained the best results on the UNICT-FD1200 (i.e., Bag of Textons with Schmid filter bank in L*a*b* color space and codebook of 12000 words). The results in [92] are outperformed with an improvement of at least 26% for the accuracy and more than 20% for the MAP score as reported in Table 3.8. Recently in [119], the authors propose a Random Forest classification algorithm on the UNICT-FD889. The proposed representation outperform also the results reported in [119].

Table 3.8: Accuracy and Mean Average Precision of the representation used in [92] and the Bag of Textons representation with Schmid filters.

Representation		Accuracy	mAP
8890	Textons - Gray - Global	27,70%	35,98%
1100	Textons - RGB - Global	60,17%	67,46%
	SIFT - RGB - Score	58,12%	62,74%
	PriCoLBP - RGB	56,33%	63,52%
12000	Textons (Schmidt) - Lab - Global	**86,17%**	**89,21%**

3.3.11 Anti-Textons Results

So far I have presented different experiments which have pointed out that Bag of Textons representation, obtained considering Schmid filters on L*a*b* domain, obtains the best performances on the UNICT-FD1200 dataset. One contribution of this work, is the introduction of a novel representation based on the concept of Anti-Textons, in order to encode spatial information in the classic Bag of Textons representation. To demonstrate the performances of Anti-Textons representation, I have compared the different filter banks to compute the Bag of Textons representation on L*a*b* space with a very small number of visual words equal to 375. As confirmed by the results reported in Table 3.9, Anti-Textons representation involved the best results in all of the configurations. Moreover it is interesting to note that the results obtained considering only 375 visual words with Anti-Textons representation and Schmid filters (85.01%) is close to the one obtained when 12000 visual words are employed (87.44% - see Table 3.7) which has a higher cost in terms of representation storage and similarity computational time during retrieval. On the other hand, the computation of Anti-Textons representation is more expensive with respect to the original Textons based representation since it has to encode the spatial information among textons.

3.3.12 Classification experiments

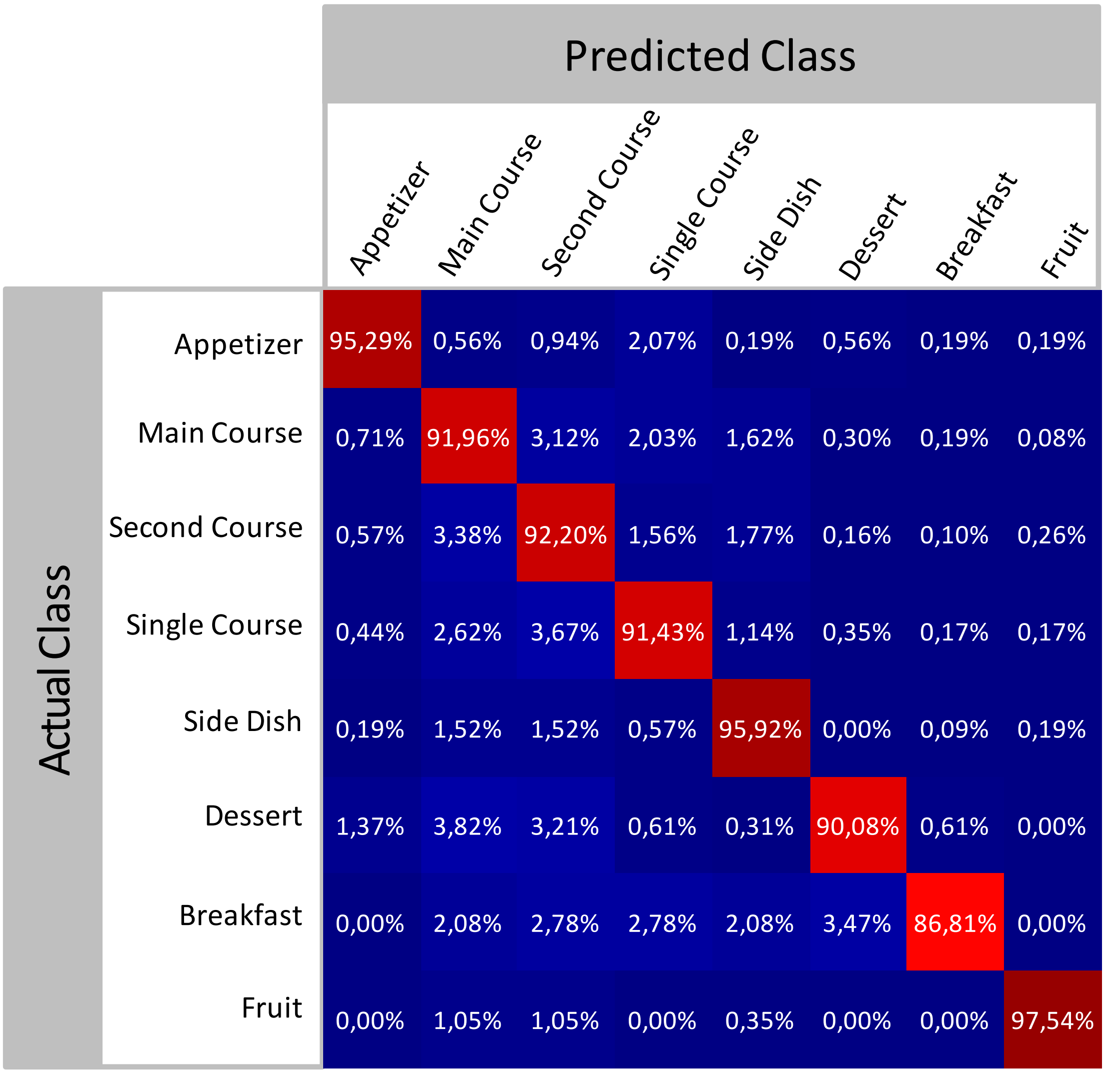
In previous sections I have presented different tests to assess the performances of a retrieval system at varying of features and parameters. As point out by the experiments, an accuracy of 87.44% and a MAP of 90.06% can be achieved on the UNICT-FD1200 dataset by exploiting Schmid Textons computed on the L*a*b*

Table 3.9: Accuracy and Mean Average Precision of the Bag of Texton and Anti-Textons representations with 375 visual words in L*a*b domain.

	Representation	**Accuracy**	**mAP**
375	Textons - LM Lab	74,75%	80,15%
375	Anti-Textons - LM Lab	76,23%	81,39%
375	Textons - MR4 Lab	77,18%	82,11%
375	Anti-Textons - MR4 Lab	78,40%	83,05%
375	Textons - MR8 Lab	80,83%	85,12%
375	Anti-Textons - MR8 Lab	82,21%	86,17%
375	Textons - Schmid Lab	83,77%	87,30%
375	Anti-Textons - Schmid Lab	**85,01%**	**88,22%**

domain with a large vocabulary of 12000 visual words. Moreover tests pointed out that the Anti-Textons representation improve the results in every configuration used. Another task we can consider in the UNICT-FD1200 dataset is classification. As detailed in Section 3.1 each image of the UNICT-FD1200 is labelled with one or two of the following classes: Appetizer, Main Course, Second Course, Single Course, Side Dish, Dessert, Breakfast, Fruit. To perform the classification test I have considered the best Bag of Textons representation mentioned above. For a proper evaluation, I have performed two kind of experiments by using $1-NN$ classifier and χ^2 distance. First, to consider the fact that images can have multiple labels (e.g., Second Course and Side Dish) as evaluation criteria we count a positive match for the query i when $T_i \cap P_i$ is not empty. Let be T_i the set of the true labels for the query image i, and P_i is the set of the predicted labels. The average classification accuracy obtained by using Bag of Textons was 93.04%. Despite this strategy could produce too much positive match, I want remark that the multi-labelled images of UNICT-FD1200 have no more than 2 labels. As second evaluation, the training sets and test sets have been reduced by removing the images with multiple labels. Classification results for this test are reported in the confusion matrix in Fig. 3.16. In this case, the accuracy was 92.60%.

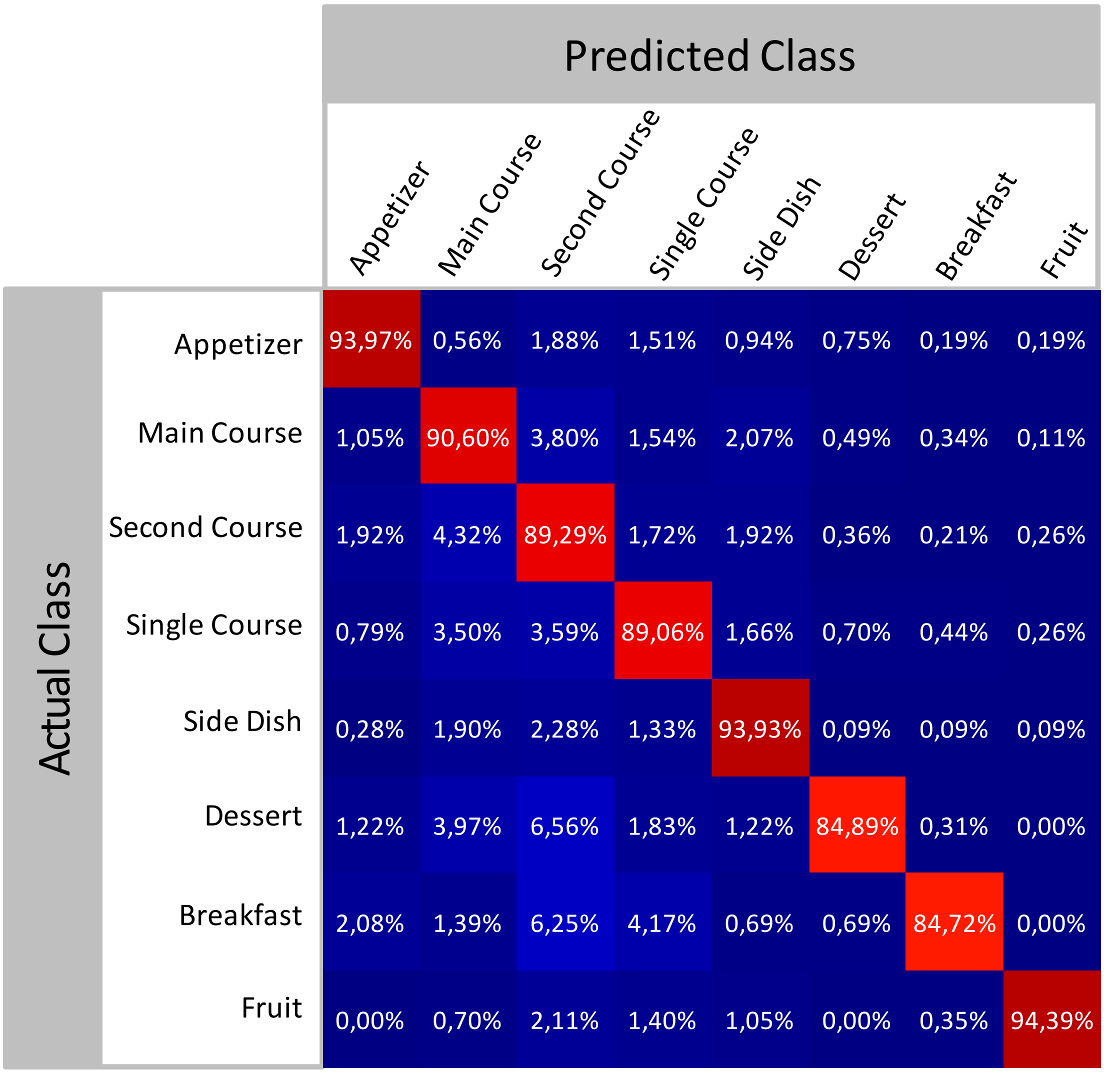
I have also performed classification tests by using the proposed Anti-Textons representation (Schimd filters, L*a*b* color space) with a codebook of 375 elements. In order to compare properly the standard Bag of Textons approach, with the respect



Actual Class \ Predicted Class	Appetizer	Main Course	Second Course	Single Course	Side Dish	Dessert	Breakfast	Fruit
Appetizer	95,29%	0,56%	0,94%	2,07%	0,19%	0,56%	0,19%	0,19%
Main Course	0,71%	91,96%	3,12%	2,03%	1,62%	0,30%	0,19%	0,08%
Second Course	0,57%	3,38%	92,20%	1,56%	1,77%	0,16%	0,10%	0,26%
Single Course	0,44%	2,62%	3,67%	91,43%	1,14%	0,35%	0,17%	0,17%
Side Dish	0,19%	1,52%	1,52%	0,57%	95,92%	0,00%	0,09%	0,19%
Dessert	1,37%	3,82%	3,21%	0,61%	0,31%	90,08%	0,61%	0,00%
Breakfast	0,00%	2,08%	2,78%	2,78%	2,08%	3,47%	86,81%	0,00%
Fruit	0,00%	1,05%	1,05%	0,00%	0,35%	0,00%	0,00%	97,54%

Figure 3.16: The confusion matrix for the classification tests related to food. The image representation used is the Bag of Textons with Schmid filter bank in L*a*b* color space and codebook of 12000 words.

to Anti-Textons representation, the same test using Bag of Textons have been repeated by using a vocabulary of 375 visual words. The accuracy obtained with Bag of Textons was 90.42% whereas Anti-Textons representation has got an accuracy of 91.21% confirming its effectiveness. In 3.17 and 3.18) note that the Anti-Textons representation, with only 375 visual words, is able to reach an accuracy very close to the Bag of Textons with a vocabulary of 12000 Textons (91.21% Vs 93.04%).

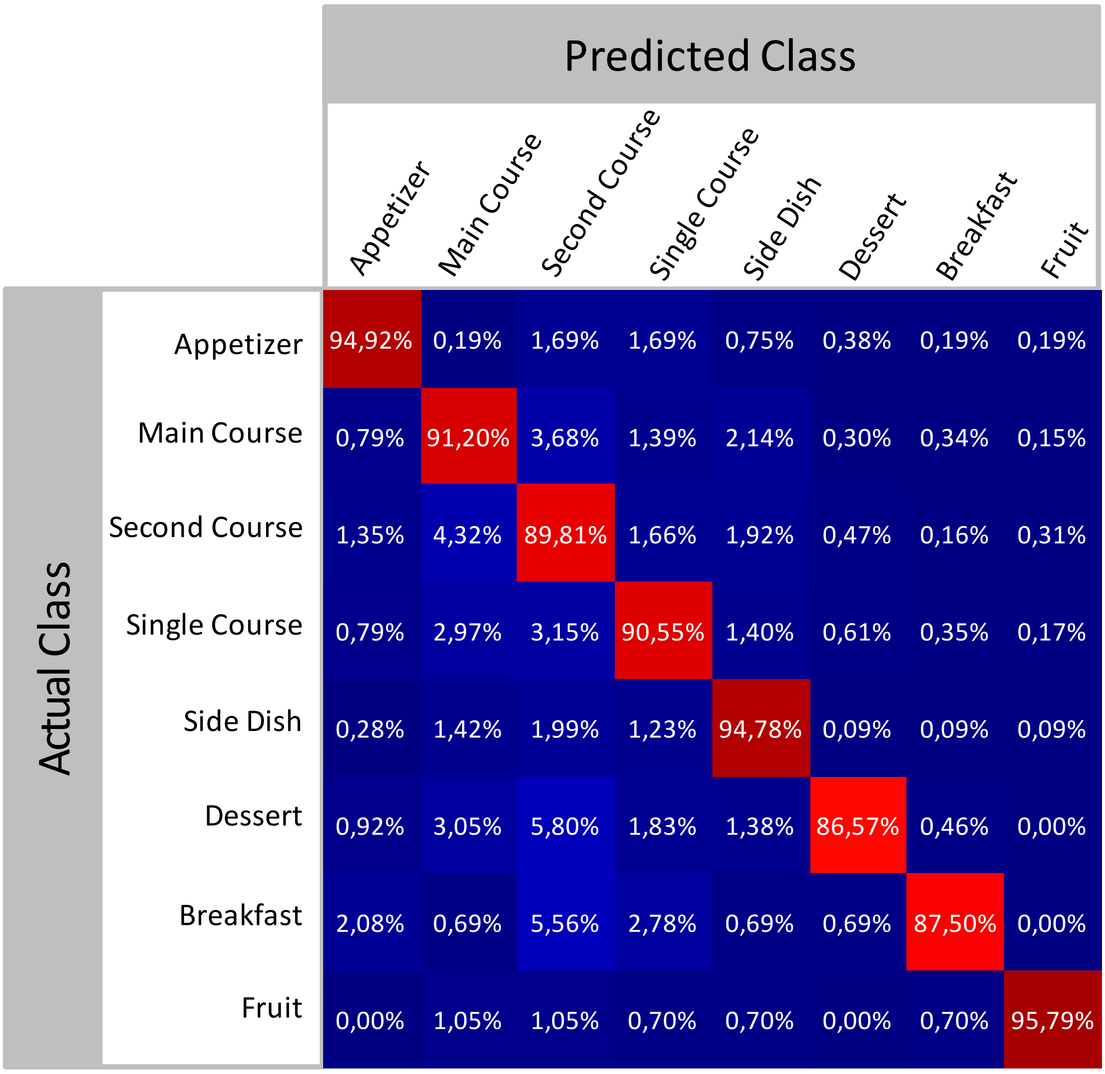


Actual Class \ Predicted Class	Appetizer	Main Course	Second Course	Single Course	Side Dish	Dessert	Breakfast	Fruit
Appetizer	93,97%	0,56%	1,88%	1,51%	0,94%	0,75%	0,19%	0,19%
Main Course	1,05%	90,60%	3,80%	1,54%	2,07%	0,49%	0,34%	0,11%
Second Course	1,92%	4,32%	89,29%	1,72%	1,92%	0,36%	0,21%	0,26%
Single Course	0,79%	3,50%	3,59%	89,06%	1,66%	0,70%	0,44%	0,26%
Side Dish	0,28%	1,90%	2,28%	1,33%	93,93%	0,09%	0,09%	0,09%
Dessert	1,22%	3,97%	6,56%	1,83%	1,22%	84,89%	0,31%	0,00%
Breakfast	2,08%	1,39%	6,25%	4,17%	0,69%	0,69%	84,72%	0,00%
Fruit	0,00%	0,70%	2,11%	1,40%	1,05%	0,00%	0,35%	94,39%

Figure 3.17: The confusion matrices for the classification tests related to food. The employed image representations is Bag of Textons with Schmid filter bank in L*a*b* color space and a codebook of 375 visual words.

3.3.13 GoogLeNet classification

For a proper evaluation of the proposed representations I have performed the classification experiments by employing a CNN-based method. Specifically, to perform tests I fine tuned GoogleNet [121]. Results show an accuracy for the CNN method of 51.41% which is much lower than the accuracy obtained with the representations proposed in this chapter. This is not a surprise, because the CNN-based methods



		Predicted Class							
		Appetizer	Main Course	Second Course	Single Course	Side Dish	Dessert	Breakfast	Fruit
Actual Class	Appetizer	94,92%	0,19%	1,69%	1,69%	0,75%	0,38%	0,19%	0,19%
	Main Course	0,79%	91,20%	3,68%	1,39%	2,14%	0,30%	0,34%	0,15%
	Second Course	1,35%	4,32%	89,81%	1,66%	1,92%	0,47%	0,16%	0,31%
	Single Course	0,79%	2,97%	3,15%	90,55%	1,40%	0,61%	0,35%	0,17%
	Side Dish	0,28%	1,42%	1,99%	1,23%	94,78%	0,09%	0,09%	0,09%
	Dessert	0,92%	3,05%	5,80%	1,83%	1,38%	86,57%	0,46%	0,00%
	Breakfast	2,08%	0,69%	5,56%	2,78%	0,69%	0,69%	87,50%	0,00%
	Fruit	0,00%	1,05%	1,05%	0,70%	0,70%	0,00%	0,70%	95,79%

Figure 3.18: The confusion matrices for the classification tests related to food. The employed image representations is the proposed Anti-Textons method with Schmid filter bank in L*a*b* color space and a codebook of 375 visual words.

usually need a huge amount of data for a proper training. It is important to note the there are real possible cases where a dataset like the one proposed, should be acquired for retrieval task purpose. We can imagine the task of retrieving the food images during the meal in a canteen (e.g., images of food offered by a company). Since building a ground-truth dataset has a high cost, the company should be able to acquire just few images of the plates offered considering the diet of the week (which is usually known in advance) to set the classification system. The recognition/retrieval system can work exactly in the way I presented in this chapter. Moreover, it is important to note that for training purpose is not possible crawl huge amount of images from the web. Indeed, the diet of the canteen is fixed and the company has to learn on the plates of their menu. The proposed solution allows to build a quite robust retrieval system that not only recognize the food, but also can give information on ingredients and calories of the plates (because this information are known by the company and are already present in their databases of the diets per portion. The system has to be able only to associate pictures to the different records of the receipts). In such cases the proposed approach for food classification can be suitable since CNN cannot be successfully applied.

3.3.14 Experiments on the Menu-Match Dataset

To properly assess the proposed method I performed experiments by employing another food dataset. Specifically, I have considered the Menu-Match dataset introduced in [120] and compared the proposed approach with respect the approach described in [120]. The authors of [120] proposed a system which provides automatic classification by using priors about the provenance of food plate depicted in the acquired images. Specifically, the system is able to recognize food plates which are served in a predetermined set of restaurants. Thanks to GPS coordinates stored in the query image metadata, most of the restaurants can be discarded. The Menu-Match dataset contains 646 multi-labelled food images across 41 food categories, which have been acquired with six different mobile devices by five photographers in three different restaurants, in order to guarantee a considerable photometric variability. I evaluated the proposed approaches on the Menu-Match dataset (GPS coordinates have not been used) to compare the performances with respect to the one obtained in [120]. In the original work, the acquired image was represented by employing Bag of Words paradigm and six different kinds of features, among which: color features, Histogram of Oriented Gradients (HOG), Scale-Invariant Features Transform (SIFT), Local Binary Pattern (LBP) and Textons with MR8 filters bank. All the aforementioned features were encoded through locally-constrained linear encoding method (LLC) and finally joint in a unique feature vectors. Since, Menu-Match dataset contains multi-labelled images, the top-5 average recall has been proposed by the authors as evaluation metric, while a one-vs-all SVM has been employed for training and classification. The experiments in [120], report a top-5 average recall of 83.00% with a 30720 dimensional feature vector. I performed tests employing the same Training-Testing protocol proposed by the authors, using the proposed Bag of Textons and Anti-Textons with a vocabulary of 1024 visual words. The experiments pointed out that the proposed Bag of Textons representation in L*a*b* domain and Schmid filters bank, outperforms the representation suggested in [120] obtaining a top-5 recall of 84.05%. A further boost in the performances has been obtained with the proposed Anti-Textons representation (85.82%).

3.4 Discussion

In this chapter the problem of Food Image Analysis has been taken into account. I have focused on the problem of food image retrieval and classification. The new dataset UNICT-FD1200 has been introduced for the study of food image representation and different tests have been done to compare state of the art representation approach. Another contribution is the introduction of a computational approach to encode the perceptual concept of Anti-Texton in order to consider spatial information into the Bag of Textons approach. Experiments have pointed out that Textons based representation computed in a L*a*b* domain considering the Schmid filter banks achieve good performances on both retrieval and classification tests. Finally, I have demonstrated that the proposed Anti-Textons representation is able to improve the results based on the Bag of Textons paradigm. Future works can consider the exploitation of more complex representation as well as a different level of classification (e.g., ingredients) to better describe a food plate. Moreover, considering the achieved results, systems based on retrieval mechanisms can be also built to deal with the problem of food intake monitoring and calories estimation.

Food understanding has become more and more of interest for both research community and society. There is a general consensus that multimedia assisted dietary management systems can be useful to improve the quality of life. To this aim will be important to build systems able to automatic answer different questions from food images:

1. which kind of food is in the image?

2. what are the ingredients of the detected food?

3. does it contain allergic ingredients (e.g. nuts)?

4. which is the volume of the food?

5. how many calories I will assume with this plate?

The above questions pose many challenges. As first, it will be important to build and share benchmark labelled datasets in order to test and compare the different solutions. Common evaluation methods on the benchmark datasets should be proposed to better assess the performances of the systems with respect to the different tasks (e.g., is a classification score of 99% acceptable in case of detection of allergic ingredient classification?). Studies on pixel-wise semantic segmentation of the food images are still needed to better deal with ingredients identification. An in-depth analysis of the volume estimation methods from single food images, as well as from multiple images is still missing in literature. In the next Chapter a new food dataset will be introduced to face the aforementioned problems.

Chapter 4

A 3D food dataset for volume estimation[1]

Over the last years there have been a number of systems that use visual meal information to output nutrient content, mainly calories and carbohydrates [122, 123, 124, 125, 126], with only few of them being validated by end-users [127, 128]. Typically, once the visual information is available, a number of computer vision steps are executed: food detection, semantic segmentation and volume estimation. By knowing the food type and its volume and by using food composition databases the contained nutrients are estimated. A key element in the development and technical validation of the computer vision steps is the data availability. However, the currently available food image datasets addresses only needs related to the food recognition step. In this chapter, I introduce a database that contains annotated and labelled RGB and RGB-D images from 80 different central-European meals served on a round dish accompanied by accelerometer data. Each meal consists of two to four different food items (e.g. vegetables, meat) of know weight, volume and nutrient composition. The newly introduced database offers resources to improve the current methods, compare among different approaches and hopefully progress the field of automatic diet assessment.

[1] The work presented in this chapter has been done while I was a visiting Scholar at the ARTORG Center - University of Bern, under the supervision of PD Dr. Stavroula Mougiakakou.

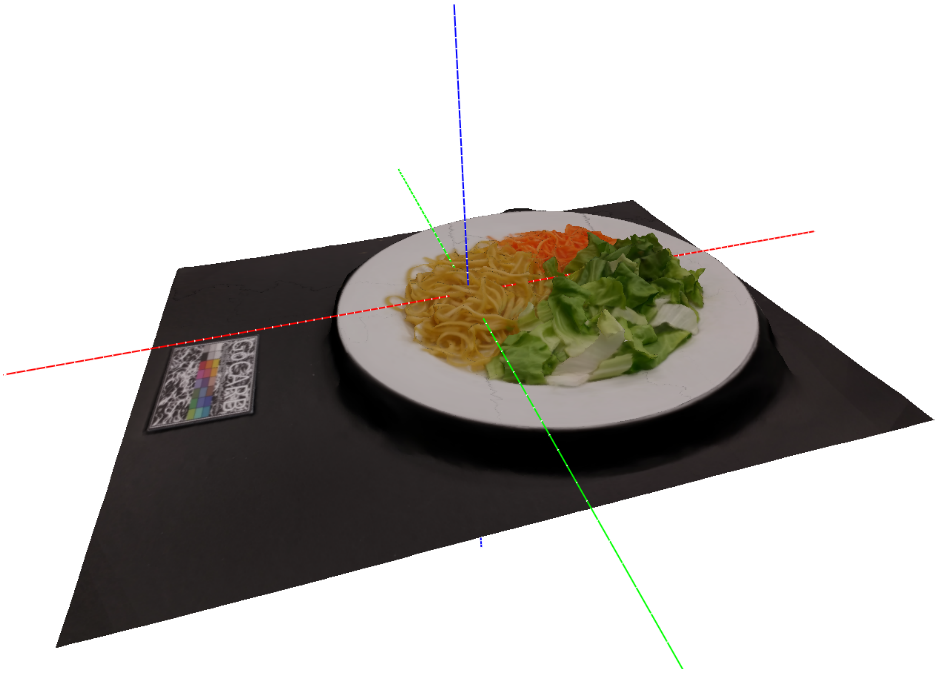


Figure 4.1: A sample of 3D model in the proposed dataset.

4.1 Proposed Dataset

Each of the 80 meals was placed on a table with a fully visible reference card next to it for color and geometric calibration. The acquisition procedure was conducted in the environment of a laboratory following two setups: i) constrained and ii) unconstrained. For each setup, the following systems were used: Intel® RealSense$^{\text{TM}}$ Camera SR300 and GoCARB App [128] installed on a Samsung Galaxy S4. Finally, the LG Nexus 5X was used to get a 3D multiview reconstruction used as ground truth for computing the food items' volume. An example of dish 3D model is reported in Fig. 4.1.

4.1.1 Constrained Setup

The dish was placed in a small table with a rotating bracket mount with limited degrees of freedom, in order to control distance and angle. The acquisition device was attached at the top of the bracket. Data were acquired at two different distances

(40*cm* and 60*cm*) and four angles (0°, 30°, 60°, 90°). Thus, for each dish a total of eight captures have been acquired.

- Intel® RealSense™ Camera SR300: From the depth sensor, we got four different types of images per capture:
  1. A 24-bit RGB image at 1920 × 1080;
  2. A 16-bit depth 640 × 480 image, where the pixel values is the distance from the sensor in tenth of millimetres;
  3. A depth image aligned with the RGB one;
  4. An RGB image masked with the related aligned depth map.
- GoCARB App installed in a Samsung Galaxy S4: From the GoCARB system for each capture I get a 4128 × 3096 RGB image and the information about calibration, as well as the gravity vector.

4.1.2 Unconstrained Setup

In the unconstrained setup, the device was placed freely in front of the dish and data have been acquired at a randomly chosen distance and angle in the range of 40*cm* to 60*cm* and 45° to 90° respectively.

- Intel® RealSense™ Camera SR300: From the depth sensor, 200 consecutive RGBD frames at 10 fps have been captured.
- GoCARB App installed in a Samsung Galaxy S4: Three image pairs have been captured, each of them with the characteristics mentioned in the constrained setup.

Finally, approximately 50 images with resolution 4032 × 3024 were captured from all possible angles above the table (360° view) using the LG Nexus 5X. These images were used to build the ground truth 3D model of each meal. The acquisition information is summarized in Fig. 4.2.

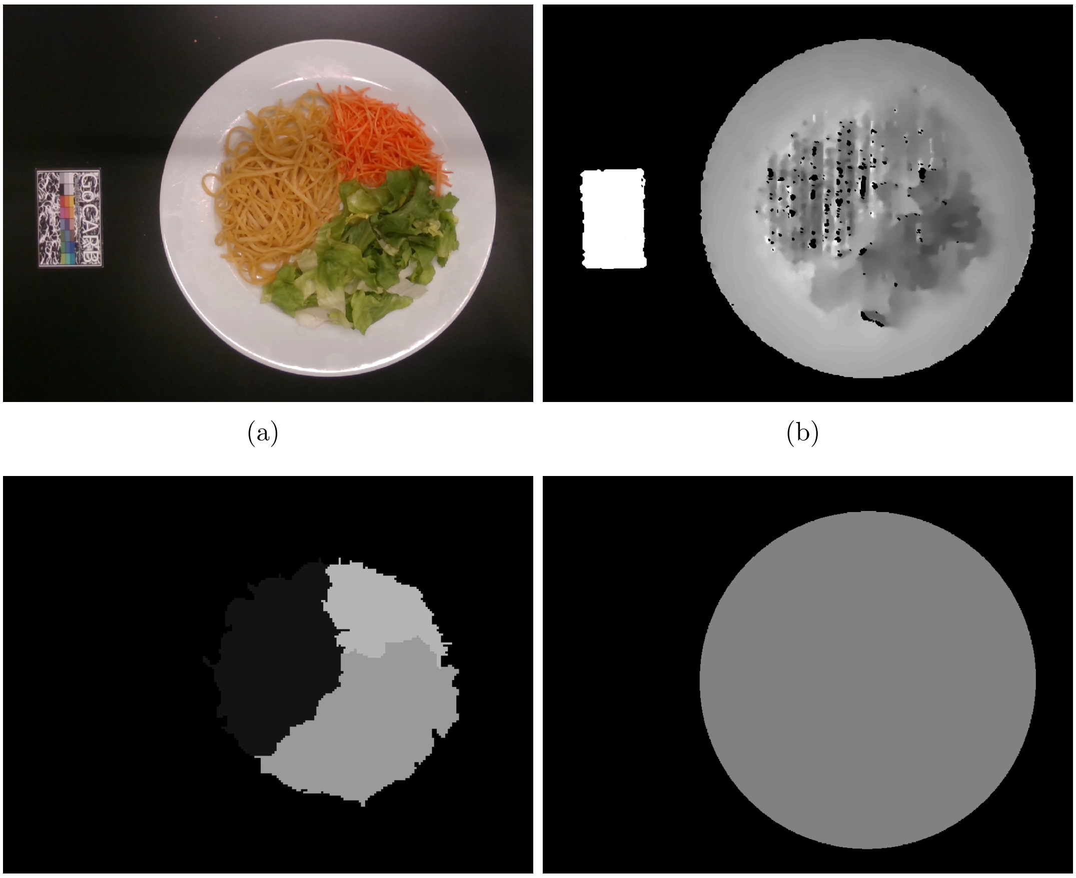
Sensors	**Setup**	**Images**	**Distance (cm)**	**Angle**	**Maps**	
					Recognition	**Segmentation**
Intel® RealSense™ Camera SR300	Constrained	8 RGB-D	40	0°	-	
				30°	-	
				60°	×	1
				90°	×	1
			60	0°	-	
				30°	-	
				60°	×	1
				90°	×	1
	Unconstrained	200 RGB-D	[40-60]	[45°-90°]	×	2
Samsung Galaxy S4 (using GoCARB)	Constrained	8 RGB	40	0°	-	
				30°	-	
				60°	×	1
				90°	×	1
			60	0°	-	
				30°	-	
				60°	×	1
				90°	×	1
	Unconstrained	6 RGB	[40-60]	[45° -90°]	×	1
LG Nexus	Unconstrained	~50 RGB	[40-60]	360° view	×	1
Total/ dish		**~272 (208 RGB-D; ~64 RGB)**			×	**12**
Total for the 80 dishes		**21807 (16640 RGB-D; 5167 RGB)**			×	**960**

Figure 4.2: A summary of the acquired RGB and RGB-D data, along with the provided maps. For each meal served on a round dish the weight and volume of each food items is available, as well as information from smartphone's accelerometer.

4.1.3 Data Processing

Image labelling and annotation: For a subset of the acquired RGB and RGB-D images, segmentation and recognition maps are provided after manual manipulation. Details are presented in Fig. 4.2, while a sample of the proposed database in given in Fig. 4.3.

Ground truth estimation: The set of photos obtained with the LG Nexus 5X has been used to create a 3D reconstruction of the dish, through the online Autodesk Recap 360 service. The resulting 3D models were manually cleaned, rotated to



(a)

(b)

(c)

(d)

Figure 4.3: (a) the RGB acquisition performed with Intel$^{\circledR}$ RealSense$^{\text{TM}}$ at $40cm$ and $90°$; (b) the depth map to the picture in (a); (c) the segmentation map for the image in (a); (d) the plate map for the image in (a). Intensities have been adjusted for visualization.

a horizontal alignment, and rescaled by using the real size as obtained from the calibration card. In this clean model, I manually separated the food items. Hence, I have computed the individual volumes, in order to use it as ground truth for volume estimation algorithms.

4.2 Baseline

The images of the proposed database were used to benchmark some state of the art methods for food segmentation, depth and volume estimation.

4.2.1 Segmentation

Food segmentation is a challenging task due to the great variability in food types, shapes and colours. Here, I investigate whether the use of depth-map information could improve the segmentation result. To this end, I applied a method similar to [129] and compared the results with and without considering the depth as input. The method consists of two main steps: border maps extraction by a convolutional neural network (CNN) and region growing segmentation. In these experiments, I altered the first step by utilizing different CNN architectures (SegNet [130] and U-Net [131]) and using the depth map as additional input. Specifically, the top view ($40cm$ and $90°$) acquisition performed with Intel® RealSense™ Camera SR300 was used, after being inverted to represent the distance from the table. Finally, all the data have been normalized and automatically cropped to 256×256 by employing the plate map. I used 60 images for training, 10 for validation and 10 for testing. To increase the image variability,I augmented the dataset by considering two flips and four rotations.

The results, confirm that depth-map information can reduce the error for borders extraction step and consequently for segmentation. Online augmentation (column AUG RGB-D in Table 4.1, has been performed by randomly modifying for each image at each iteration of the training data. Specifically, we add a random number from a normal distribution with mean 0 and standard deviation 0.01, to the colour channels, while we multiply the depth map with a random number with mean 1 and standard deviation 0.1. The metrics used to assess the performance are the same used in [129]. I tested different architectures (SegNet, Unet), loss functions (mean square error - MSE and mean absolute error - MAE) and batch normalization strategy (per feature-map, mode 0; per batch, mode 2). Best results have been achieved with Unet, with no batch normalization e by training the CNN with online augmentation. As expected the best result has been obtained with depth information, moreover data augmentation with the smallest standard deviation has increased the training generalization.

Table 4.1: Segmentation results (MSE: Mean square error; MAE: Mean absolute error)

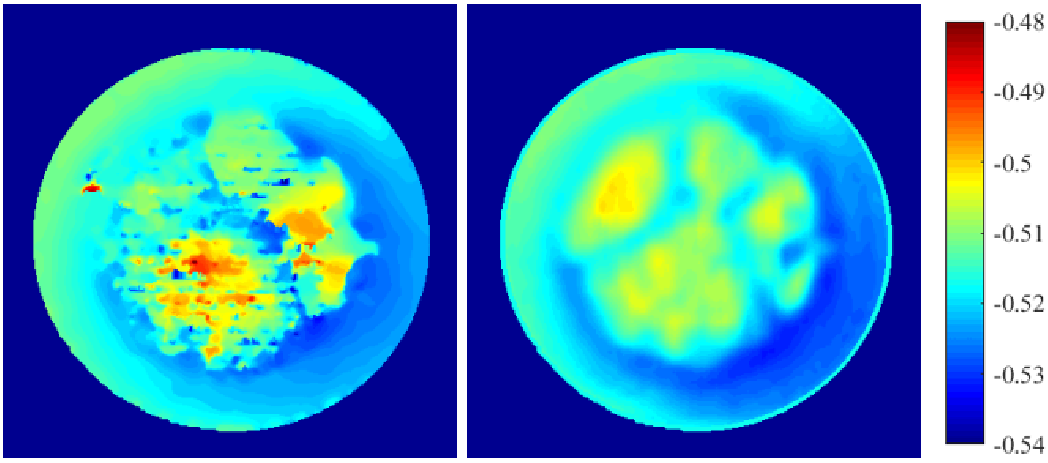
CNN	Loss	AUG RGB-D	RGB		RGB-D	
			Min Fscore	Total Fscore	Min Fscore	Total Fscore
Segnet	MSE	No	0.7329	0.9326	0.7059	0.9288
Unet	MAE	No	0.6889	0.9268	0.7351	0.9342
Unet	MSE	No	0.6875	0.9247	0.7332	0.9328
Unet	MAE	Yes	0.6893	0.9281	**0.7426**	**0.9369**

4.2.2 Depth Estimation

Calculating the depth of a food image is a significant component in understanding the 3D geometry of a meal, which is essential for food volume estimation. However, depth prediction is usually performed by stereo images or motion as in [132]. Here, I present a method that performs depth estimation by using just one RGB image, as input to a CNN. The method is inspired by [133], although some modifications have been made to adapt it to the problem. The considered task focuses on the estimation of the depth of near distance objects (within 1 meter), whereas the scenarios in [133] range within several meters. The dataset is composed by two images per dish acquired by the Intel® RealSense™ Camera SR300 in unconstrained setup: 60 dishes used for training, 10 for validation and 10 for testing, resulting in 120, 20 and 20 images, respectively. The training data were augmented by flipping the images left-to-right. The chosen network architecture is similar to Segnet-Basic [130], which consists of four encoding and four decoding convolutional layers. Each convolutional layer has 64 kernels and is followed by batch normalization and a ReLU activation, while the last layer uses the sigmoid activation function as a loss function, we use the mean absolute difference (MAD) between the estimated depth map and the ground truth from the depth sensor. For optimization, we use Adam with learning rate of 0.0005. Table 4.2 reports the quantitative comparison of the depth prediction on the proposed dataset, where only the pixels inside the plate are evaluated. Apart from MAD value, the absolute relative difference (ARD) with respect to ground truth is also provided for the sake of clarity. As expected, in food depth prediction scenario, the result obtained using standard algorithm of [133] shows relatively poor performance. However, by using the proposed method, the performance is significantly improved. For further demonstration, the prediction result performed by the proposed method is shown in Fig. 4.4(b), revealing a good

Table 4.2: Comparison on the proposed dataset (MAD: Mean absolute difference; ARD: Absolute relative difference).

Method	**MAD (mm)**	**ARD (%)**
NYU [133]	37.09	7.53
Proposed	**8.64**	**1.76**



(a) (b) (c)

Figure 4.4: (a) depth map captured by the Intel® RealSense™ Camera SR300; (b): depth map predicted by the proposed network; (c) Colour bar is in meter.

agreement with the ground truth depicted in Fig. 4.4(a).

4.2.3 Volume Estimation

Knowing the food volume is critical to estimate its nutritional value. In this experiment, I compare the performance of the GoCARB system in two different scenarios. As first I estimate the volume of each food item by reconstructing a 3D model as in [132]. The second experiment is aimed to assess the importance of depth map in volume estimation. I replace the depth estimation step as calculated in [132] with the depth obtained from the RGB-D images captured by depth sensor. In this case, I have to estimate the vertical direction and the table plane from the depth map to calculate the volume. To do so, I modify the table plane estimation method of [132]. First, I detect the plate through RGB channels, then sample the depth map at its border, and fit a plane to the selected points to find the ellipse plane. To find the table plane, I select all the points outside of the plate, and shift the ellipse plane to their modal height. To measure the performance the mean absolute percentage error, as defined in [132], was used. In these conditions, the average error using stereo reconstruction was 13.8%, and 14% using RGB-D images. The two methods provide comparable results, however it has to be noted that the RGB-D sensor baseline (distance between the two elements of the stereo reconstruction module) is quite small, reducing its accuracy, while already developed algorithms were employed without any prior optimization the specific problem. However, these results indicate that a monocular RGB-D image can replace stereo pairs for volume estimation without performances drop.

4.2.4 Discussion

In this chapter I have introduced a new multimedia food database that contains images, depth maps, weight/volume measurements of served meals, nutrient content together with the corresponding annotations, labels and accelerometer data. To benchmark this dataset the results of some baseline methods on food segmentation and depth/volume estimation have been presented. As expected, the segmentation's results confirm that depth-map information can decrease the error for borders extraction step and consequently for the regions identification. In food depth prediction task, the result achieved using state-of-art algorithm of [133] have shown poor performance. However, by using the proposed method, the error has been significantly reduced. Finally, for volume estimation problem, the results indicated that a RGB-D image could replace a stereo pairs of RGB ones without error increasing. The proposed dataset, will be publicly available for the researcher in the field, to test new food understanding algorithms about semantic segmentation, depth and volume estimation.

Chapter 5

Conclusion

The core of this thesis is the investigation of food understanding by using Computer Vision and Machine Learning approaches. The three main problems addressed are: food vs non-food discrimination; food images retrieval and classification; segmentation and volume estimation. The aforementioned three vision tasks are mentioned in order with respect to the complexity of the related challenges.

Chapter 2, addresses the discrimination between food and non-food images. It can be considered the first step to be employed in a food understanding engine: before to classify the food it is essential understand if an image depicts food. The proposed approach exploits the One-Class Classification paradigm. According to this paradigm a classifier is trained by using samples from a single class (usually the positive one).

The results pointed out two main facts:

- The best performance are achieved by combining Bag of Textons representation and SIFT one. As shown in [92], texture information are highly discriminant for food images;
- Achieved True negative rate (non-food images correctly classified) is 94.44% while true positive rate food (food images correctly classified) and 65.43%. This suggests that, although the proposed method provide encouraging results, the employed classifier (One-Class SVM) tends to penalize food images. This depends on the tolerance parameters of OSVM, that can be tuned. However, the employed tolerance is the one that maximize the accuracy with respect to the considered food dataset.

Moreover, when images contain food but also background and other objects, the classification become more difficult. This is because the whole image is considered during food vs non-food classification.

The second problem addressed in the thesis is the food retrieval and classification, described in Chapter 3. A new dataset, named UNICT-FD1200, has been introduced for the study of food image representation and different tests have been done to compare state of the art representation approach.

The main contribution is the introduction of a new computational strategy to encode Anti-Texton. The proposed algorithm, allows to exploit spatial information of standard Bag of Textons representaiton. It have proved that the proposed Anti-Textons representation outperform the results based on the standard Bag of Textons paradigm.

The last food understanding problem considered in this thesis is segmentation and volume estimation. It is discussed in Chapter 4. Food volume estimation is probably the last step of a food understanding engine. In fact, if one knows the kind of food item and its volume, will be able to compute the nutritional values (i.e. proteins, carbohydrates, etc.).

Because of the absence of a publicly available dataset to face the problem in a proper way, we introduced a new one that contains images, depth maps, weight/volume measurements of served meals, nutrient content together with the corresponding annotations, labels and accelerometer data. Three baseline methods on food segmentation and depth/volume estimation have been described.

The segmentation's results confirm that depth-map information can decrease the error. For volume estimation problem, the results indicated that a RGB-D image can replace a stereo pairs of RGB ones without error increasing.

5.1 Future directions

Food understanding is a challenging problem because of food variability in appearance and its intrinsic deformability. One of the main issues is the availability of proper labelled datasets to test methods, baseline algorithms and common evaluation procedure. At moment, there are several public dataset of food. Nonetheless, most of these datasets are suitable for retrieval and classification task and contains

RGB images only. In future works we consider to extend the dataset described in Chapter 4. The use of depth map for food understanding tasks have to be further investigated, since the provided benchmarks shown positive results. Currently most of the consumer mobile system don't mount a depth sensor; despite this, the cost reductions and the recent quality stereo vision method, are allowing to perform depth estimation with most of the smartphones. Finally, we are considering to develop a food understanding engine that is able to work in real-time on a video stream. Temporal dimension could surely improve volume estimation task; nevertheless, problem like blurring, shaking, etc. have to be properly handled.

Appendix A

Cultural Heritage Preservation and Exploitation

In this appendix, I report a set of works where computer science and modern technologies are employed for Cultural Heritage preservation and exploitation.

Firstly, the problem of virtual unrolling of ancient papyrus is described. The term "Virtual unrolling" refers to the method and strategy to digitally unroll a papyrus scroll. The aim is read the ancient document by avoiding to damage it. The proposed approach is based on the use of X-ray computer tomography to get a sliced version of the papyrus scroll. Then, the slices "stack" is processed by using morphological operator to perform the digital unrolling.

Among the new technologies currently proposed for the application to Cultural Heritage, the potentialities of the 3D scanning technique represent a significant example of how originally far apart fields, such as the one of conservation, that of research and that of advanced industry, can find a common interest ground. Non-invasive experimental use of methodologies and innovative tools have been developed for analysis procedures of geometric dimensional data, restoration and monitoring. 3D scan is the core of the next works presented in this Appendix.

The study of Morgantina Silver Treasure is addressed in Section A.2. The main contribution of this work is the creation of a semantic segmentation platform. It is intended to present the results of chemical and physical analysis conducted on the Morgantina Treasure.

The next work focus on the low-cost hand-held 3D scanners. The aim, is to understand the weakness and the advantage of these kind of scanners. Hence, I conducted a study on a XVIII century doorway placed in the monastery of Benedettini in Catania, by performing a comparison between a low-cost hand-held 3D scanner and a Time of Flight one.

Finally, it is presented a digital anastylosis of a Greek Archaic statue from ancient Sicily and the development of a new public outreach protocol for those with visual impairment or cognitive disabilities through the application of 3D printing and haptic technology. Specifically, it is presented an analysis two assess if two archaeological pieces (a head and a torso) are parts of the same kouros.

A.1 Virtual Unrolling

Papyrus is a thin material made from the pith of the papyrus plant Cyperus Papirus, an aquatic plant that was once abundant in southern Egypt and in southern Sudan. It is currently cultivated in the Nile Delta. Examples of typical use are related to the classical period of Egyptian civilization, where they were used to record historical events, business transactions or even trivial events (Fig. A.1).

During the excavation of archaeological sites are often found many ancient artifacts of this type, however it is not always possible to physically open the scrolls and read their contents (Fig. A.3(b)). This happens for various reasons, such as the parchment may be too fragile and opening it would risk to introduce tears or creases, it could be now totally stuck or even afflicted by parasites and worms that have destroyed the material. The common practice is to manually carefully remove one by one each layer of papyrus and recompose this kind of puzzle over a flat plane. This is an invasive practice and it is not more allowed by the recent trend of the restoration theory.

Therefore, it is necessary to investigate new techniques for analysing the content of parchments and papyri scrolls without the need to physically unroll them: this technique is called *Virtual Unrolling*. In this study I performed a virtual enrolling of a papyrus scroll using a X-ray computed tomography scanner (GE Optima 660). Papyrus has been sectioned in several thin axial slices, in each of which it appears as a spiral (Fig. A.2); interestingly, in correspondence of the Egyptian inscriptions, I detected some areas characterized by a higher density than the remaining part of the sheet.

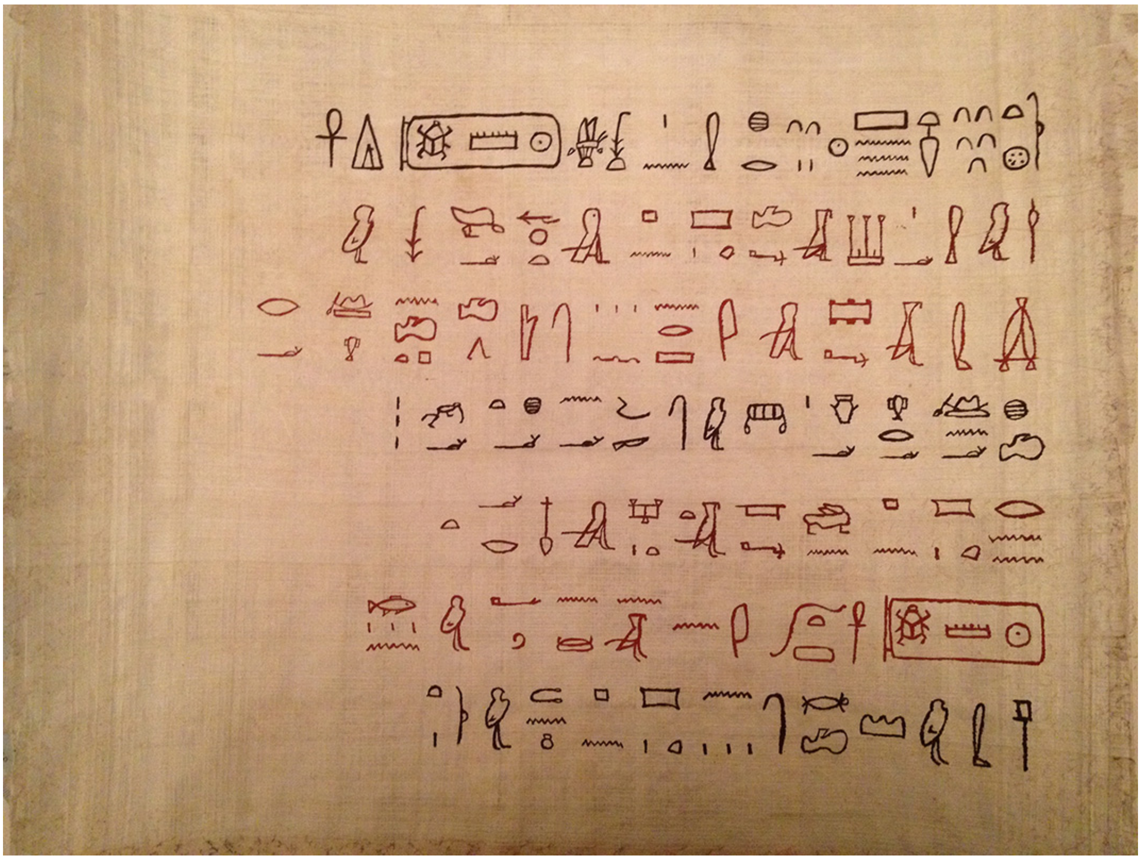


Figure A.1: The papyrus sample used to test the proposed virtual unrolling procedure.

In order to make this study-case realistic, it has been realized a papyrus substrate made by the original method described by Plinius the Elder. It has been painted with hieroglyph inscription of Thutmosis III using pigments and binders compatible with the Egyptian use (ochres with natural glue). In particular, the papyrus substrate was made by carefully cutting a papyrus (Cyperus Papyrus) into pieces of desired sheet length (about 40 cm), peeling the rind off these, and slicing the pith into thin layers from the narrowest of the three angles of the triangular stem to the middle of the opposite side. The strips were leached in water until translucent, and pounded. Then, while still moist, they were laid horizontally on a board, one above the other overlapping by 1-2 mm. A second layer was then laid at right angles vertically covering the first layer, and both were subsequently pressed together until dry. The strips were used without any glue to steak the layers each other,

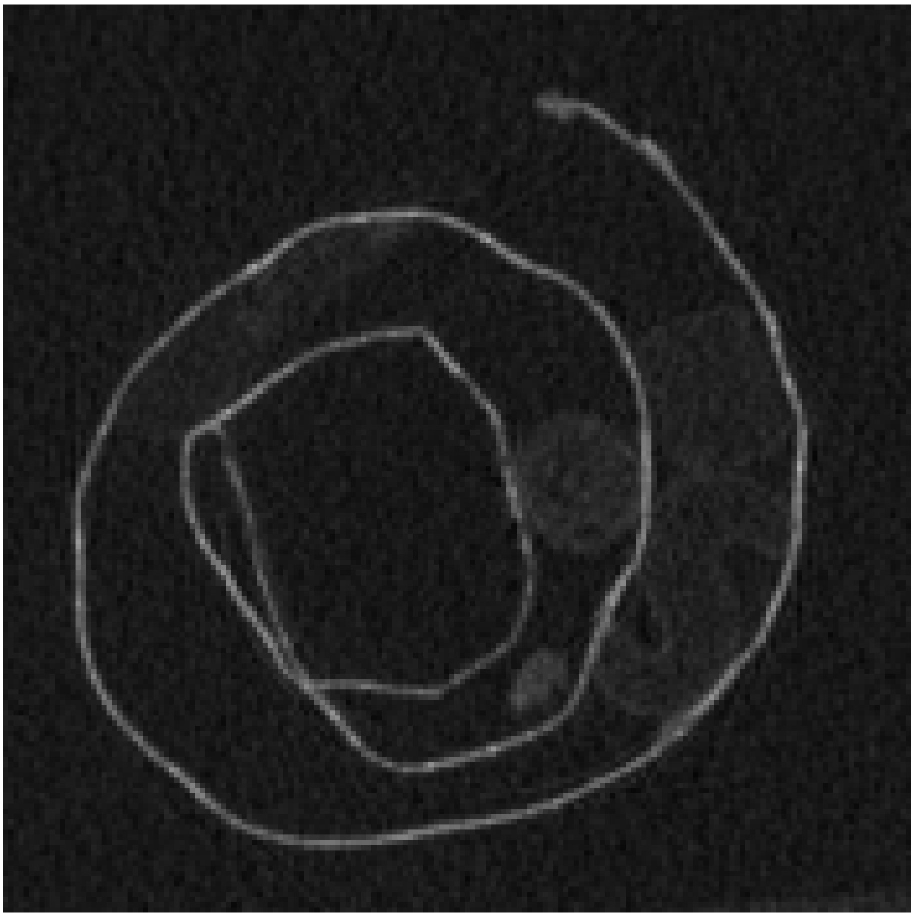


Figure A.2: CT image showing a thin axial section of the papyrus.

being the adherence due to both physical and chemical reasons because the natural polysaccharides contained in the fibres The inks were prepared by mixing different ochres and carbon powder with Arabic gum dissolved in water. Red ochre mainly contained Fe_2O_3 and the brown one Fe_2O_3+MnO_2+carbon. The black was obtained with carbon powder. If metal oxides shows very different X ray radio-densities with respect to the organic substrate, the carbon based ink of course are not so different. In case of a pure carbon inks, the use of ^{13}C Nuclear Magnetic Resonance Tomography (NMRCT) would be used instead of XCT: from the software point of view the origin of the images stack is absolutely not a problem.

Subsequently, after CT acquisition of the papyrus, I have developed an algorithm to process the slices acquired. Main aim of this semi-automated procedure is to identify the path that should be followed to completely unroll papyrus. Moreover, I have also treated particular difficult cases, such as interrupted path or multiple path. Finally, once paths have been computed for (almost) all slices, then it is possible to properly exploit 3D volumetric data acquired using CT for virtual unrolling.

A.1.1 Related Works

Physically unrolling could be problematic, in particular when scrolls are very fragile. For instance, in the 1980s "Oslo method" was applied in an attempt to unroll two Herculaneum scrolls, but it results in partial destruction of them, and definitely in an irreversible loss for cultural heritage [134, 135]. For this reason, starting from that event all further attempts to physically unrolling papyri or parchments had been abandoned in favor of digital techniques that could analyse scrolls without the need of physically open them.

Some works stressed the possibility to use X-ray computed tomography (CT) to inner analysis of rolled parchments and papyri [136, 137, 138]. In particular, archaeological site of Hercolanum is rich of carbonized parchments impossible to unscroll, so many groups have focused their efforts in analysis of artifacts coming from this site [137].

More recent works, that can use better CT devices with respect on works of some decades ago, continue to afford the issue of virtually unrolling [139, 140]. Furthermore, in all of them the acquired CT slices could even present spirals almost well-spaced, where layers are quite distinguishable between each others, or totally ill-spaced. In the first case, the problem of overlapping sheet could be solved in practical way; for instance, in [139] authors applied an algorithm of Graph Cut for this purpose and used a 3D scanned version of papyrus. Besides, in the second one, the problem of entire parchment virtually unscrolling could be too complex, but some small parts of the sheet can be restored anyway; for example, in [138] authors shown how problematic could be an ill-spaced spiral when ink has low contrast; they had selected some regions of the spiral and manually made them well-spaced; finally, they performed virtual unrolling just on it.

In this Appendix, a novel semi-automated method to solve virtual unrolling issue is presented. Despite the aforementioned approaches, the proposed algorithm does not use 3D scans, but images only. Moreover, the path is automatically detected, in spite of manual approaches cited. The unique user-guided step is to select the start point and the final point position for a single papyrus slice. The proposed solution exploits morphological operator to detect slice path and to solve interrupted path issue.

A.1.2 Computed Tomography

Computed Tomography (CT) is one of the most accurate diagnostic techniques in Medicine, because it allows obtaining images of singular layers of patient's body using X-rays. The structures contained in each layer are represented in the CT images according to their density, detected through the measurement of the attenuation that every single element of the X-ray beam undergoes through the different tissues. The attenuation value of each pixel of the matrix is then converted into a specific grey level in accordance with the Hounsfield scale. So, radiologists can distinguish several anatomic components of the human organs (blood, fat, air, soft tissue, bone, etc.).

In this study I applied the principles of CT to a sheet model of papyrus with some Egyptian inscriptions, with the aim to obtain its correspondent digital image. It has been used a 64-multi-detector "Optima CT660" scan (GE Healthcare). Firstly, I performed a scan proof with the papyrus rolled out; then the rolled papyrus sheet was lying on the patient bed (Fig. A.3(a)), put inside the gantry and centred with laser lights (Fig. A.3(b)).

Then images acquisition was programmed in the axial plane adopting the following parameters:

- Scan type: helical;
- FOV: 50;
- Slice thickness: $0.625mm$ (the least possible X-ray collimation);
- Rotation time: $0.8s$;
- Pitch: 0.98
- Tilt: 0.0

- Matrix 512 × 512
- kV: 80
- mA: 50

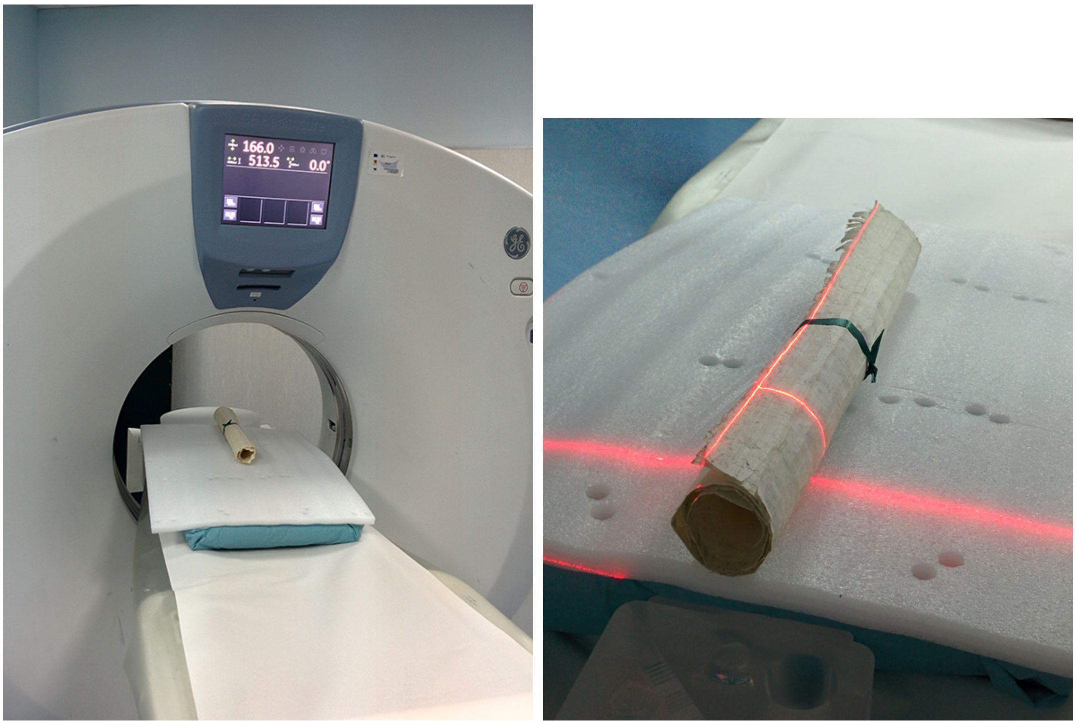
Images were transferred to the GE Advantage Window 4.6 console for post-processing with both Multiplanar Projection Reconstructions (MRP) and Maximum Intensity Projection (MIP) in the coronal oblique plane. In particular, the images obtained with the "curve" mode, selecting the layers of the papyrus manually point-by-point, allowed the detection of certain symbols on papyrus (Fig. A.4). The examinations confirmed that, although the papyrus sheet has a very small thickness (about $1mm$), CT allows a satisfactory digital acquisition. Moreover, CT can discriminate the different density of the material with which the symbols were plotted than the paper of the papyrus, allowing the identification of the areas with the Egyptian inscriptions as points of higher density in both native and, above all, reconstructed images. However, virtual unrolling by CT presented some limits due to the overlapping of the layers of the sheet that caused an incomplete reconstruction with missing parts of the papyrus.

A.1.3 Proposed Method

The images we get from XCT represent the papyrus section, so these look like as spirals. For each papyrus section, the start and final point of the spiral should be identify in order to build an array of ordered pixel. Unfortunately, it is not easy to sort the points which compose the slice, because of the low resolution and some papyrus overlapping sheet. For this reason, I decided to use a single section profile and follow it for all the slices, assuming that there are few differences between a couple of different sections.

The first step of the proposed algorithm is the selection of a good slice using the following criteria:

- Low number of overlapping sheets into the slice. In the best case there is no overlaps;
- The spiral path should be almost entirely detected through a raw segmentation.



(a)

(b)

Figure A.3: GE Optima 660 scanner: papyrus was lying on the patient bed inside the gantry (a) and centred with laser lights (b).

Step 1: Slice selection

To satisfy first criteria the skeleton branch-points are detected using morphological operator [141]. Specifically, the following image segmentation is performed:

- Gamma transformation with $\gamma = 2$ and contrast stretching in order to highlighted the papyrus section;
- Otsu thresholding to get a raw segmentation;
- Morphological skeletonization to detect a first raw path;
- Morphological branch-points detection to identify overlaps;

If a sheet overlap exists, then it produces a branch-point into the skeleton. Of course, not all the branch-points indicate a sheet overlap, since they could be image

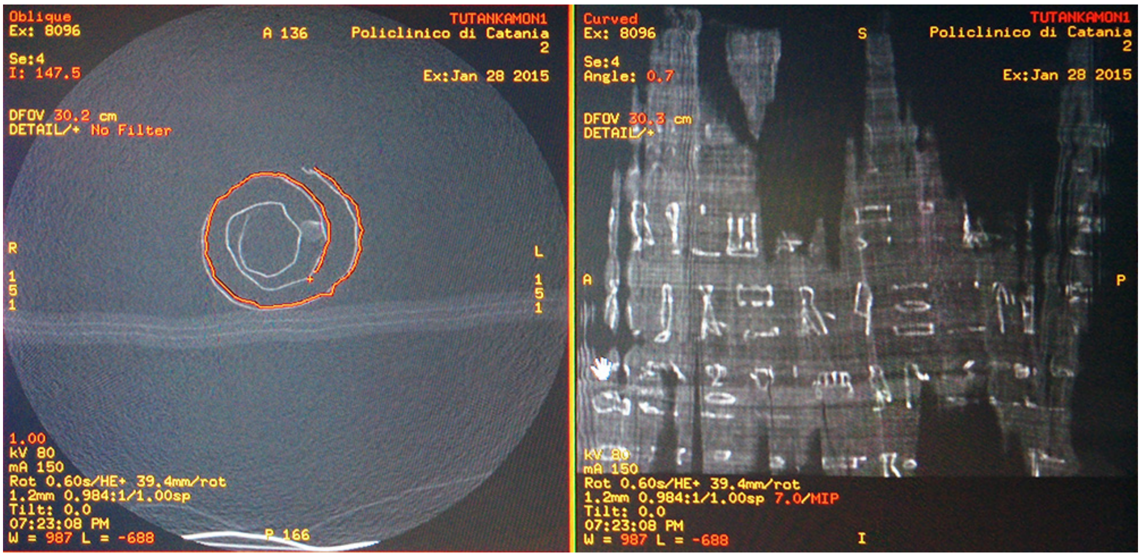


Figure A.4: Utility function of application software provided with GE Optima 660. On the left the papyrus section is manually drawing. On the right the result shows missing parts.

artifacts or papyrus creases. However we choose the slices with the lowest number of branch-points in order to minimize the probability that an overlap occurs.

To satisfy the second criteria the number of skeleton points is counted for each papyrus slice, then the average number is computed. Finally, a slice with a number of skeleton points nearest to the average value is chosen. Skeletons with too few points describe slices poorly definite, while skeletons with too many points could be affected by noise.

Step 2: Slice reconstruction

Once a good slice is detected, the spiral must be rebuilt. Firstly, we apply the preprocessing described above. Then the branch-points and their 3×3 neighbourhood are removed from the skeleton in order to disconnect the ramifications. Now the end-points are detected using the proper morphological operator. The end-points are pixels which locate a break on the skeleton.

So the user chooses the start point and the final point of the spiral. This is the unique user interaction of the algorithm. Starting from the initial point, the

3×3 neighbourhood is taken into account for each pixel and the following steps are performed:

- If into 3×3 neighbourhood there is a not visited skeleton point, which is not an end-point, then it is added to an array of visited point. You move on this new point and the algorithm continues.
- If into 3×3 neighbourhood there is a not visited skeleton point, which is an end-point, then a break has been reached. To rebuild the missing path, the intensity grey value of the contrast stretched image is used. Specifically, the pixel of maximum value is taken. However the 3×3 neighbourhood is weighted through a probability mask. There is a mask for each possible direction in 3×3 neighbourhood, so there are eight masks. These eight masks has built using a sampled derivative of gaussian filters and weighs more the pixel along the last direction of movement. When a new end-points is reached two skeleton parts are reconnected.

  The algorithm ends when the final point is reached.

Step 3: Virtual Unrolling

The last step is the papyrus virtual unrolling. For each slice we select the sequence of pixel whose coordinates are stored in the vector of visited points. For each coordinate the pixels of maximum intensity along the direction of the gradient is chosen. Indeed sheet and text or image have an higher intensity than background. In this way we obtain a string of pixels for each slice (e.g., in Fig. A.5). By stacking all this string we get the image of the papyrus rolled out.

A.1.4 Experimental results

To test the proposed approach a set of 259 slices of a single rolled papyrus has been used. The original size of the slice image is 512×512, but a crop to 175×175 has been performed in order to focus on the papyrus spiral profile. Using the aforementioned criteria, a good slice has been automatically selected: a section with 0 branch-points and a number of skeleton points near to average number. In Fig. A.6 the processing step for the chosen slice is shown.

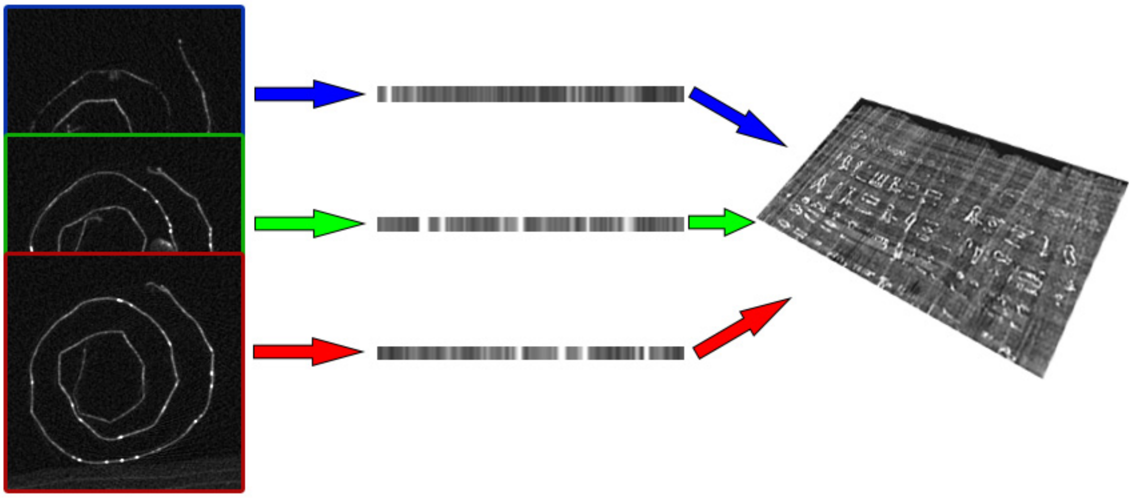


Figure A.5: A representation of the process pipeline.

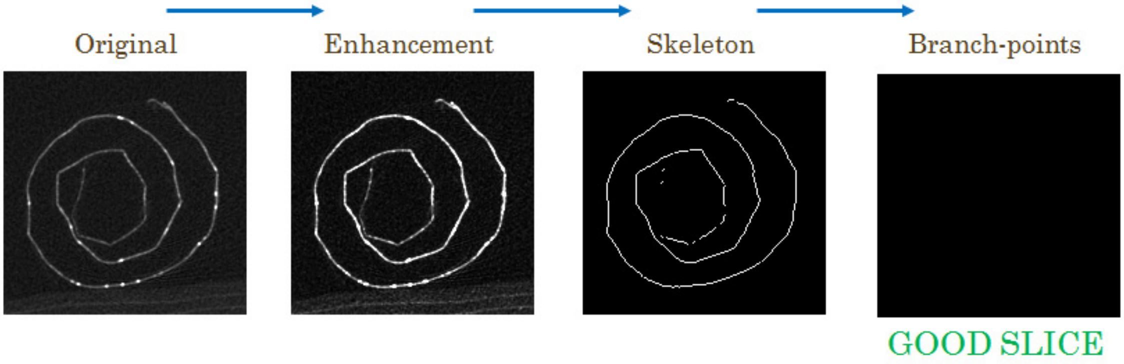


Figure A.6: An example of good slice processed. First the original cropped image is enhanced, then a morphological skeletonization is performed. Finally it can be seen the branch-points map is empty, that is no overlaps occur.

However, to show the overlap problem a bad slice example has been reported in Fig. A.7, where the branch-points pixel locate on the overlapping area can be seen. Of course, the image final image in Fig. A.6 is totally black because no branch-points have been detected.

In Fig. A.8 the reconstructed path after the step 2 is shown. Finally, using the array of visited points each of 259 spirals is unrolled. For each coordinate in the array, we take into account 25 pixel (according to resolution) along the gradient

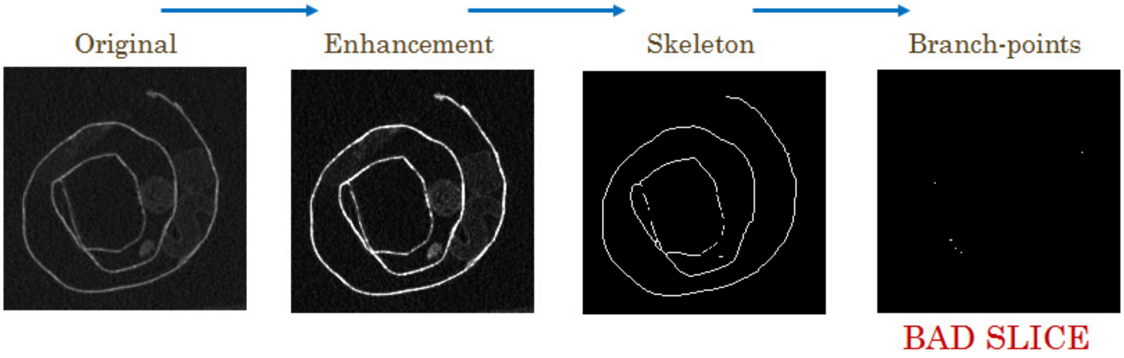


Figure A.7: An example of good slice processed. First the original cropped image is enhanced, then a morphological skeletonization is performed. Finally it can be seen the branch-points detected for the processed slice.

direction and choose the maximum value. This strategy is motivated by the fact that there is some slight difference between the prototype path and every other spiral path.

By merging all the processed slice the image of unrolled papyrus is built. This result can be seen in Figs. A.9(a) and A.9(b). In Fig. A.10 a comparison between the original version and the virtual unrolled one is shown and the common symbols are boxed.

A.1.5 Discussion

This study has been motivated by the criticality of a physical papyrus unrolling, because of the high risk to damage the cultural heritage materials. To this aim I proposed a method based on the use of a X-Ray CT scanner in order to obtain digital cross-section images of the papyrus. The input of the proposed algorithm is a set of slices from a single CT acquisition. Through mathematical morphology the spiral path of a good slice is rebuilt, in order to be used as prototype path for every other slices. The experimental results show that this approach is valid, since many symbols of the original papyrus become visible after the virtual unrolling. In the future works I consider to solve the overlapping sheets issue for each slice, so to use more than a single path. In this way a better and more accurate result could

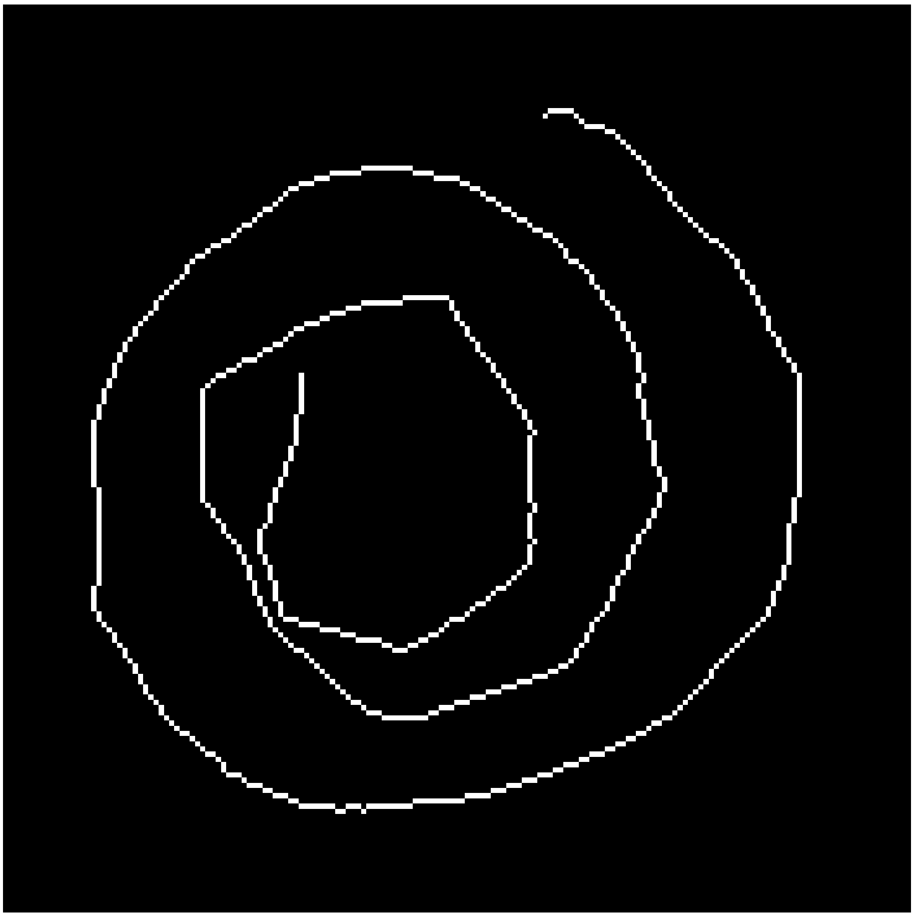
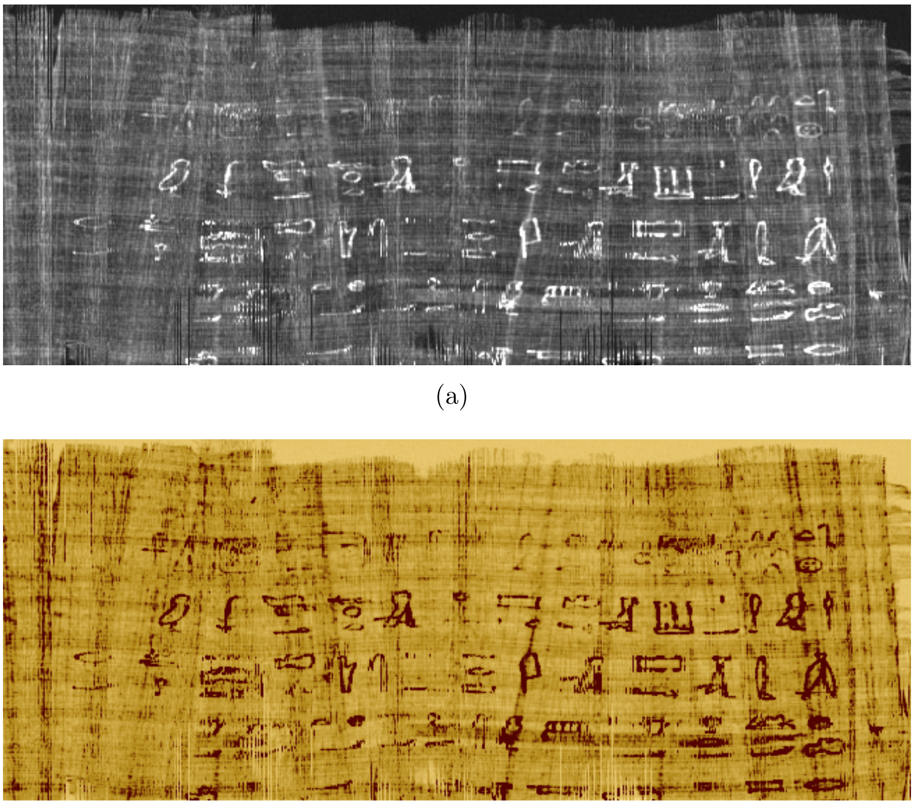


Figure A.8: The processed good slice after the reconstruction step..

be obtained. I am also considering to use an high-resolution device to reduce the probability of the overlays and the interruption in the spiral path.

A.2 Integrated 3D models of the Morgantina silver Treasure

The silver Treasure of Morgantina is one of the most valuable collections of the Museum of Aidone (Sicily). It consists of sixteen pieces of worked silver that were returned to Italy in 2010 following an agreement between the Ministero dei Beni e le Attività Culturali, the Regione Siciliana, and the Metropolitan Museum of Art of New York (Fig. A.11). Through police inquiry and data derived from direct



(a)

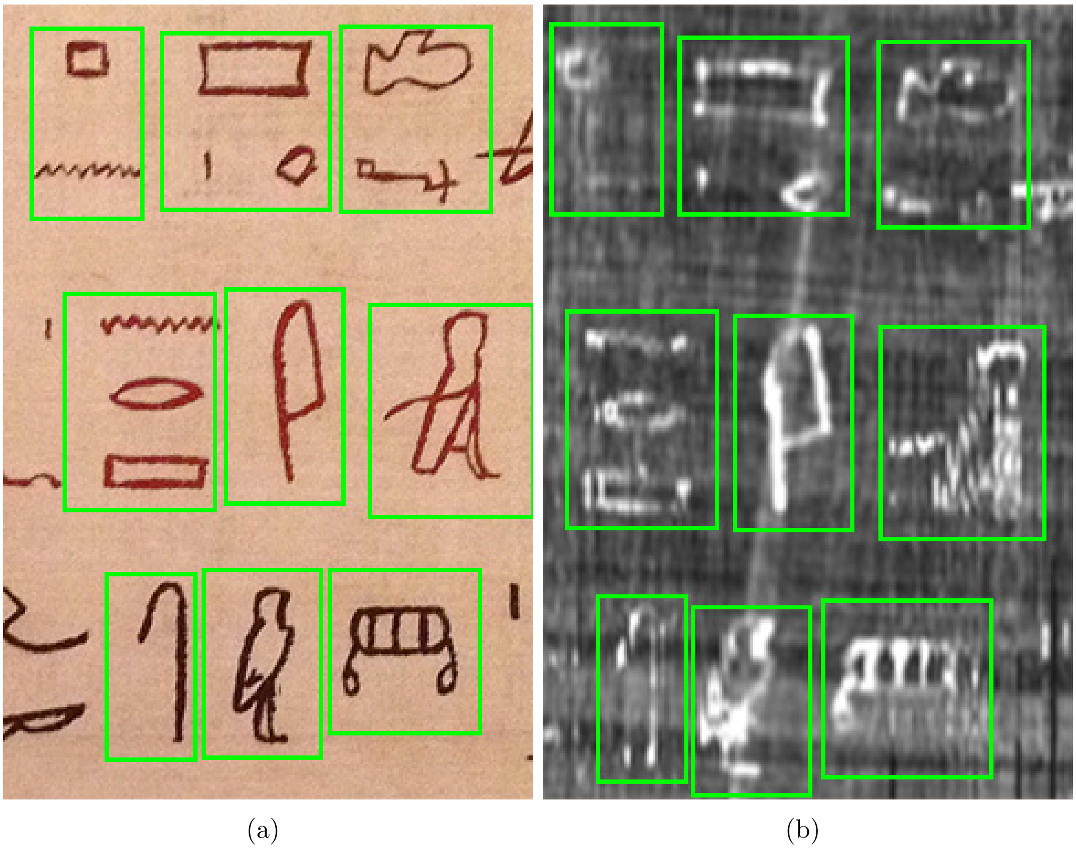
(b)

Figure A.9: (a) A virtual unrolled version of the papyrus through the proposed algorithm; (b) A false color version of the image (a).

archaeological excavation in a specific area of the ancient city of Morgantina, the origin of the finds was determined to be the so-called House of Eupòlemos, where the valuables were hidden probably during the Second Punic War [142].

Since the agreement signed in 2006 provides for the alternating temporary exhibition of the silver treasure for four years at the Museum of Aidone and then for four years at the Metropolitan Museum, a careful campaign of non-invasive analysis was prepared to document the conservation status and previous treatments of finds.

With the aim to realize a new tool to increase the existing archaeological knowledge and to obtain referenced information of the conservation state, 3D models and diagnostic data have been for the first time acquired and organized within a web-oriented platform.



(a) (b)

Figure A.10: A comparison between the original unrolled study-case papyrus (a) and the image of virtual unrolled papyrus (b). Common symbols are highlighted.

The non-invasive methods have provided complementary results, for a more comprehensive evaluation of the state of conservation and executive technique. The diagnostic study was directed to:

- distinguishing the original material from degradation and/or restoration materials;
- obtaining a deeper knowledge of the production technique;
- assessing the current state of conservation and acquiring useful data for scheduled monitoring.

One of the main purposes of this study is to produce significant scientific material for an innovative and interactive fruition to offer to visitors even during the period

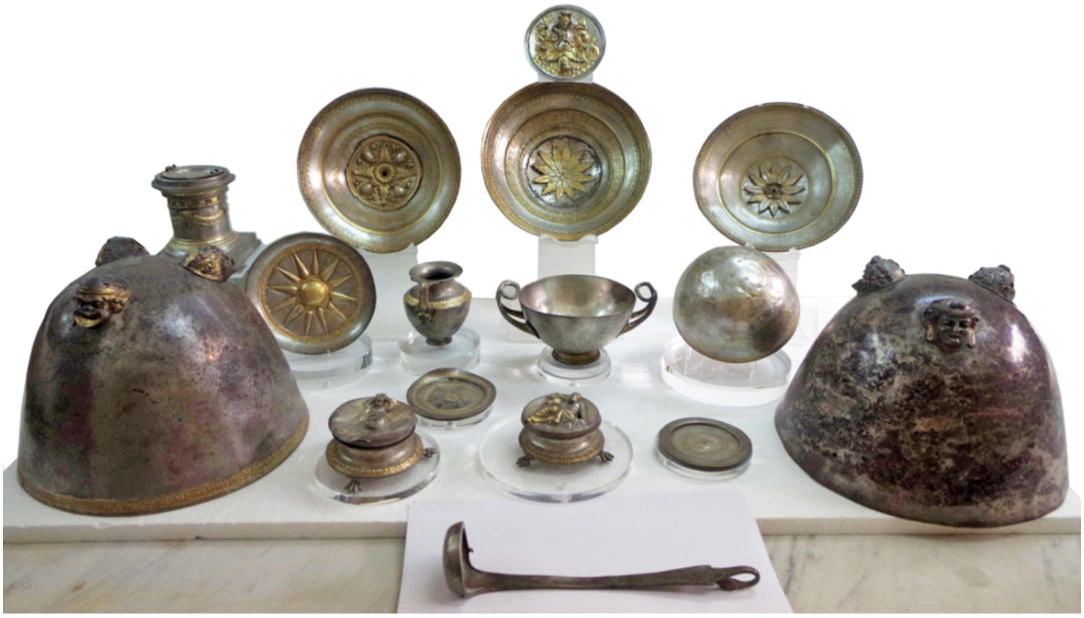


Figure A.11: The silver set from the Eupolemos's House - Archaeological Museum of Aidone (Sicily)

of absence of the silvers thank to virtual model available in the museum or in a dedicated website, for customized levels of visit of the Morgantina treasure.

A.2.1 Materials and Methods

In our case study, the innovative applied technologies had the purpose of creating a 3D collection data to assist the restoration and conservation of the Morgantina Treasure. Now, after the transfer of the collection, the 3D digitalization is bringing to restorers and archaeologists in documenting the process of investigations and presenting it to the public. The geometric survey helps us to evaluate the state of material preservation of the external and internal portions of the object and permits, each time the collection is moved to a new location, the registration of anomalies and stresses to which the object has been subjected through a systematic program of monitoring (Fig. A.12).

The process started on physical models is defined Reverse Modeling and the digital resolution up to 0.1 millimeters for each object was realized using a 3D portable

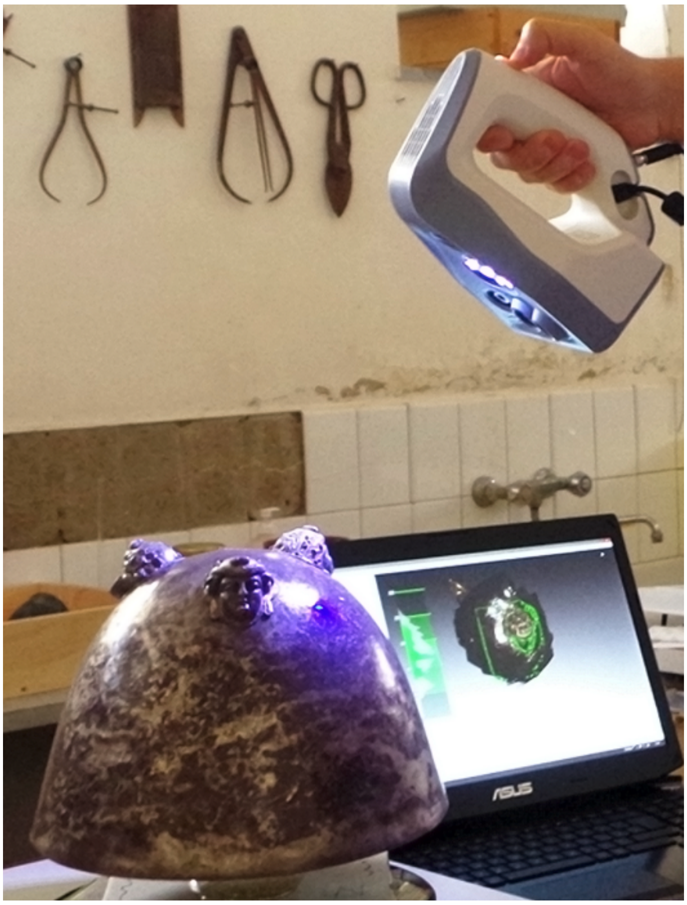


Figure A.12: Acquisition phases via 3D scanning of one piece of the collection.

scanning system with a structured light flash bulb (Artec 3D Scanner Spider), permitting highly detailed digital models to be produced. The choice of this technology was greatly determined by the physical characteristics of the 16 objects of collection to be scanned, including the size of pieces, the complexity of its outer surface, the light-reflecting properties of the surface of the metal object and the constraints on access/manipulation. The process with a high surface detail can be managed also to ensure enjoyment to various categories of users: cataloguing, restoration work, promotion, consumption and diffusion.

For each object in the collection the greatest difficulties were encountered in the alignment and registration phases of the front side and the back one, since their thickness were really tiny. It was necessary to set up some processing strategies to cope with specific problems of the objects. During the acquisition phase it has been necessary to employ specific markers and small coloured pellets modelling paste applied to the surface, after a careful evaluation with restorers. It has been acquired from a minimum of 5 to a maximum of 20 scans for each piece of collection. According to the complexity of the scanning object and the surface detail, the number of scans varies; a total of 180 scans were shot and 12GB of raw data were collected.

For each of the 3D models made the points or areas affected by the analyses have been referenced and linked to on the conservation status considerations as well as to diagnostic information obtained from spectrometric and imaging investigations. In this context, UV fluorescence data, X-Ray Fluorescence analysis (XRF) and digital X radiography, of all the silver objects were carried out directly in situ using portable equipment [143, 144].

UV-IF imaging can be used for localising and delimiting the specific areas, i.e. residuals of materials damaged and not visible to the naked eyes, where to perform further deepening spectroscopic and structural analyses. Energy Dispersive X-Ray Fluorescence (ED-XRF, or in short XRF) analysis is a non-invasive chemical technique, which allows, through the identification of chemical elements, to identify the constituting materials employed for the realization of the different investigated layers/surface (or different typology of works of art). The information coming from the X-radiography depend by absorption and scattering of the X-rays by the crossed volume and material type. They can vary from point to point, in relation to the composition and inhomogeneity thickness. The X-ray is used for the analysis of many types of artworks, to generally identify the structural features of the whole volume of the specimen. The method is completely non-destructive, both of the material that the information contained in it. Table A.1 summarize the main devices features and the acquisition parameters employed for each non-invasive techniques, optimized for the typology of analysed object and for diagnostic information to achieve.

A.2.2 Analysis

The high-quality 3D digital models are responsive to the complexity of the geometric-formal of the analyzed objects and the digital collection reproduces really well the decorations in organic form (Figs. A.13, A.14).

The diagnostic acquisitions carried out on the sixteen silver objects have produced 110 XRF spectra for the analysis of silver and gilded surfaces, and of the

Table A.1: Devices and acquisition parameters used for non-invasive diagnostic investigations.

Non-invasive technique	**Device characteristics**	**Acquisition parameters**
UV induced Visible Fluorescence (UV-IF)	Photo-camera CHROMA C4-DSP (C250ME – DTA srl); 6 Mpx CCD air cooling; KAF8300ME sensor; 8 interferential filter. Wood's lamps (UV, 365 nm) -160 W Sylvania, air fluxed- filtered by HeBO HU 01	450 nm; 540 nm; 600 nm interferential filter; 15 minutes exposition time (for each filter):
Digital X radiography	X-ray tube (Poskom X +, mod. PXP-100 CA, maximum voltage 110 kV)	70 kV voltage; 50 mA current
Energy Dispersive X-Ray Fluorescence (ED-XRF)	X-Ray tube (max voltage of 40 kV, max current of 0.2 mA, target Rh, collimator 1 or 2 mm); Silicon Drift Detector (SDD) with a 125-140 eV FWHM @ 5.9 keV Resolution; 1 keV to 40 keV Detection range of energy.	35 kV voltage; 80 microA current; 70 seconds acquisition time; 0.8 cm working distance.

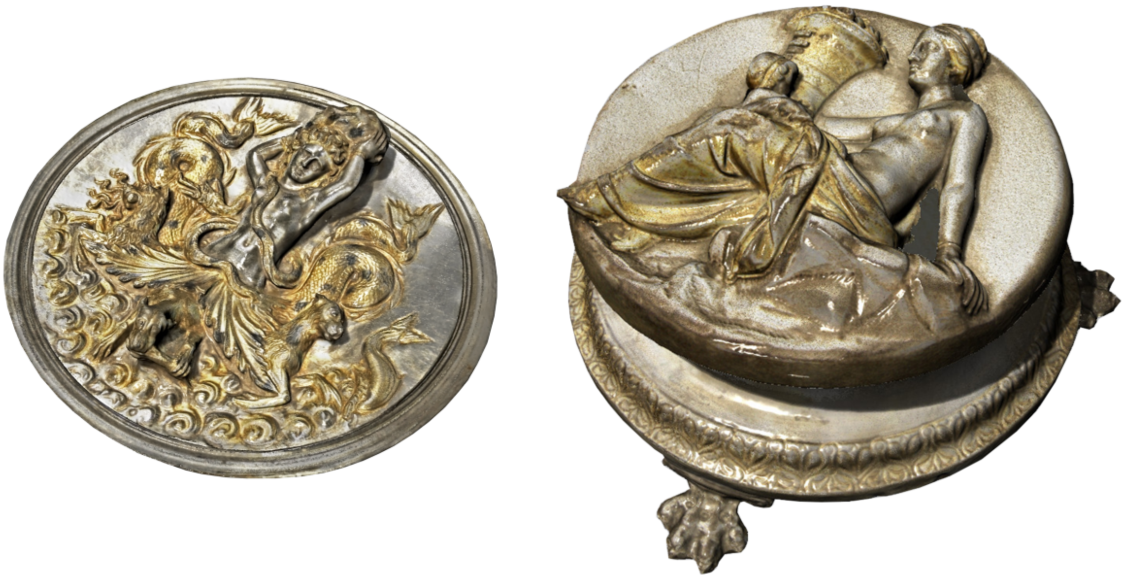
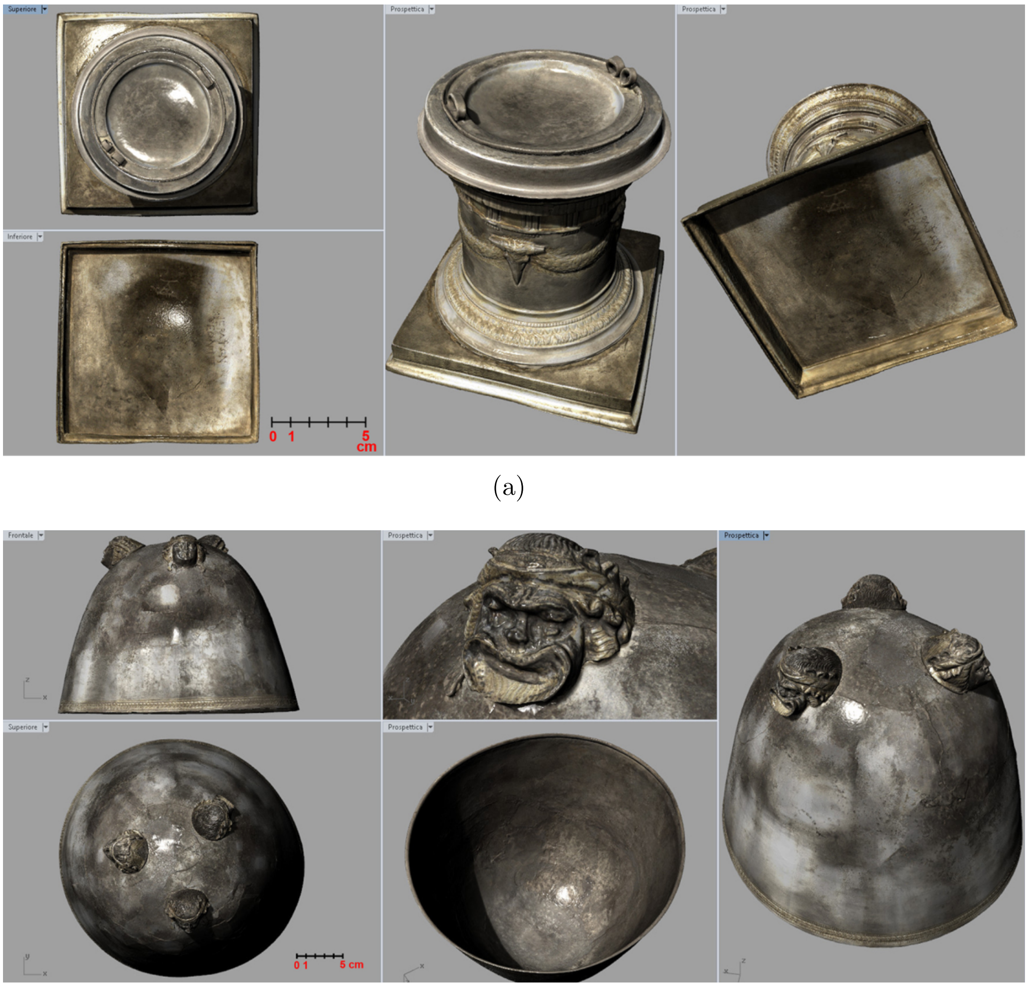


Figure A.13: 3D digital models: the Émblema with Scylla shown (on the left) and one of the two pyxides (on the right)

area affected by corrosion phenomena, that is, the formation of silver and or copper degradation products; 40 hours of UV fluorescence ($450nm$, $540nm$, $600nm$) acquisition for the identification of materials present on the surface, that is, integration, adhesives, protective materials; and 27 X-Ray exposures (2 projections for each object obtained placing more finds in the same plate in order optimize the number of acquisition) for structural analysis. The X-Ray imaging has allowed to document



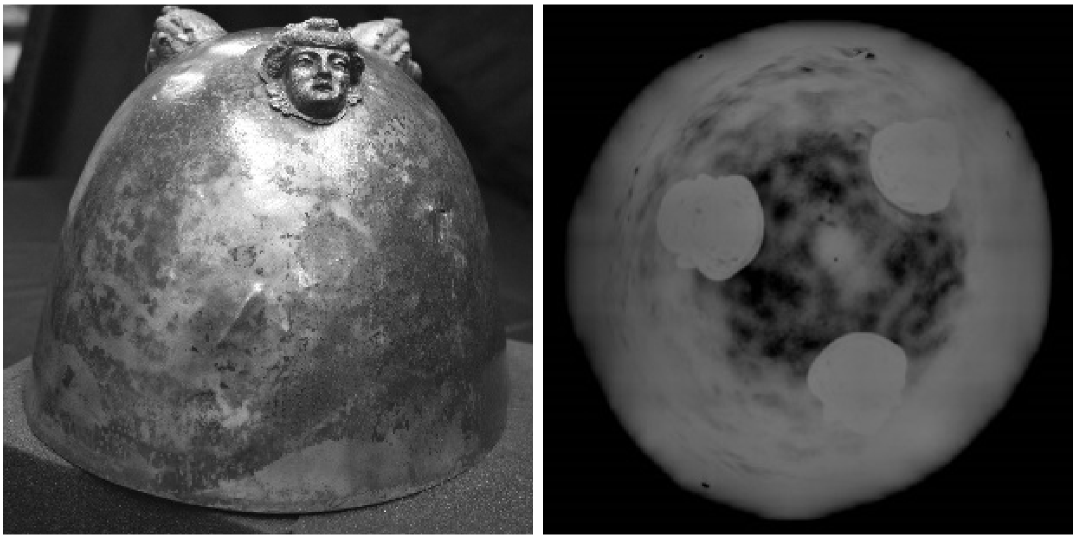
(a)

(b)

Figure A.14: Orthogonal and perspective projections of the arula (a) and the kyathos (b).

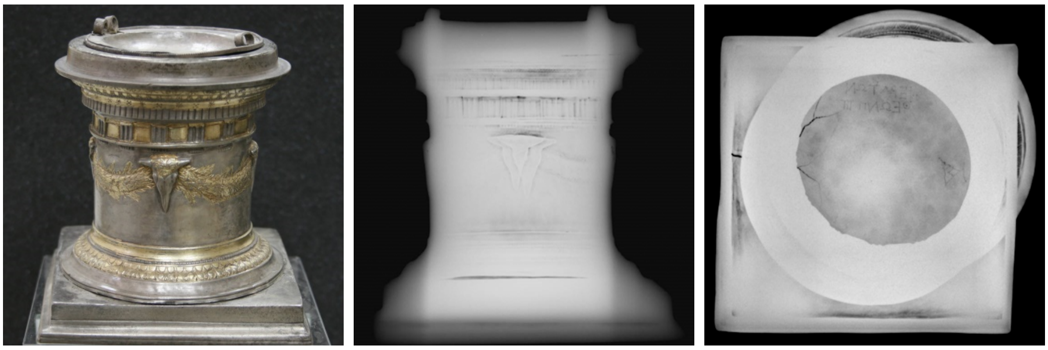
details related to the execution of embossing (Fig. A.15) and the technology of assembly (Fig. A.16).

The radiographic data, which analyses the internal structure of the object by comparing the varied absorption of X-rays, has provided information on the presence of fractures, which for the most part were subject to previous restoration (Fig. A.17), also highlighted by observations under Wood's light. Simultaneous observation of UV fluorescence image shows along the discontinuities the presence of organic material (adhesive) applied during prior restoration work carried out to solve fractures visible on X-ray. This deformation allows to suppose that the fractures are



(a) (b)

Figure A.15: X-Ray acquisition on Mastòs (a) and its upper-lower projection (b).



(a) (b) (c)

Figure A.16: X-Ray acquisition on Bomiskos (a), its lateral projection (b) and its upper-lower projection (c).

due at the time of the clandestine excavation.

For most of the analysed finds, UV fluorescence in the visible range acquisition has allowed us to map materials present on the surface, which were used for protection or integration during the past restorations. This technique highlights the use of different types of adhesives present in fractures already evident in the X-ray images.

In the case of the Eirene and Ploutos pyxide, the RX acquired in the lateral

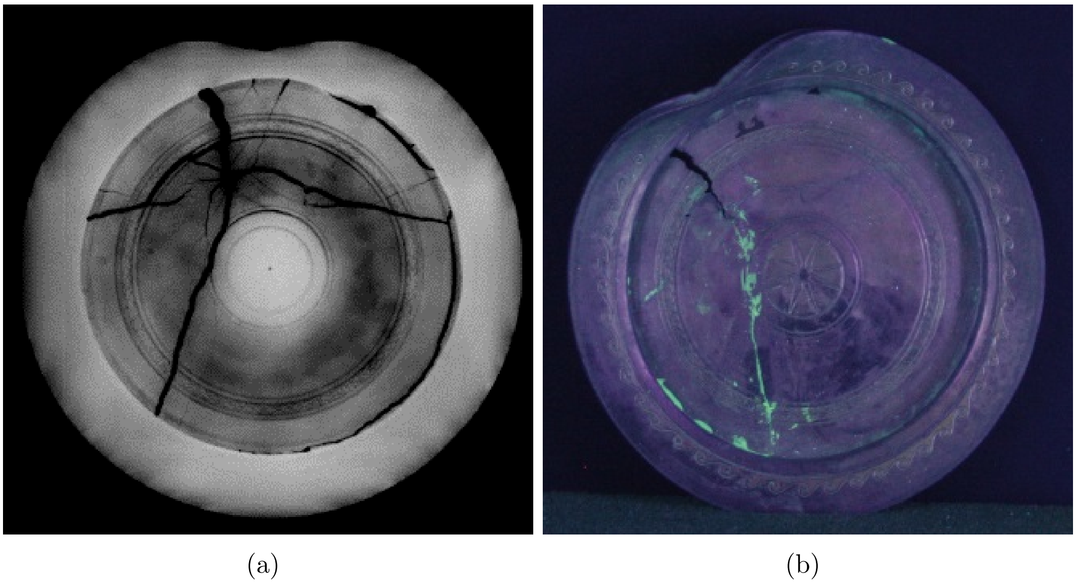

(a) (b)

Figure A.17: X-Ray (a) and UV fluorescence (b) acquisition on find NI 16a

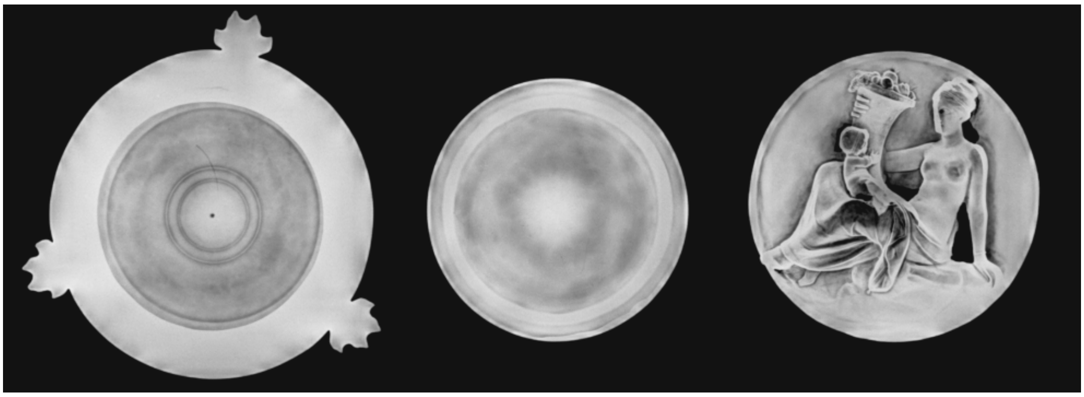


Figure A.18: X-Ray acquisition on pyxis

projection, showed the presence of a fracture along the external edge (Fig. A.18. The representation appears in radiography characterized by very variable thickness, up to thinning gradually further affecting the more relief surfaces (note the drapery of the female figure). These minimum thickness make the surface very resistant to mechanical stress. In some areas of junction between the surfaces in relief and those flat, mostly stressed in the execution process, real lacunae are evident.

For the study silver medallion with Scylla hurling a rock were acquired two

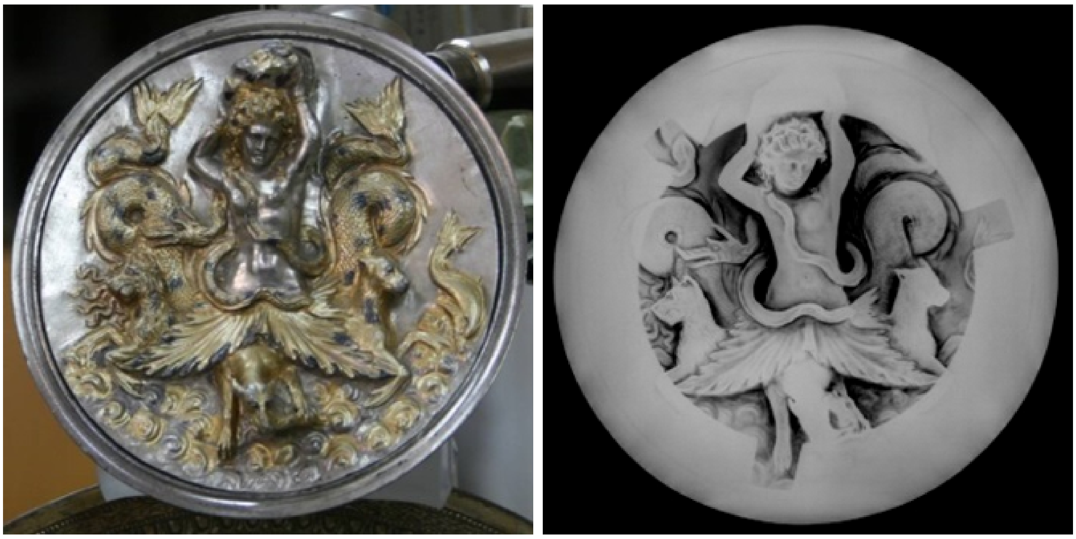


Figure A.19: Medallion with Scylla (a) and related X-Ray acquisition (b).

radiographic views (superior-inferior, and lateral projections), which showed the presence of fractures due to minimum thickness of silver foil, in the border between the flat and the relief surfaces (Fig. A.19).

Such mapping has not always been done in documenting previous conservation efforts. Finally, the analysis of the X Ray Fluorescence has enabled us to identify chemical elements, which provide information on both the silver alloy and the application of gold leaf decoration, as well restoration material localized by X- Ray (Figs. A.20 and A.21) and UV fluorescence imaging.

Among the constituent materials of precious finds, in addition to gold and silver in the silver matrix it was also found copper, but in variable ratio with respect to silver content. Starting from this analytical evidence, the ratios between the intensities of the characteristic XRF signals of copper and silver were calculated. These characteristic values, occurred constant within the metal matrix analysed for each of the findings, have allowed to obtain a significant distribution of the whole chemical data set. Indeed, on the basis of Cu/Ag ratio, three split clusters corresponding to the proposal by archaeologists based on stylistic criteria, were obtained (Fig. A.22).

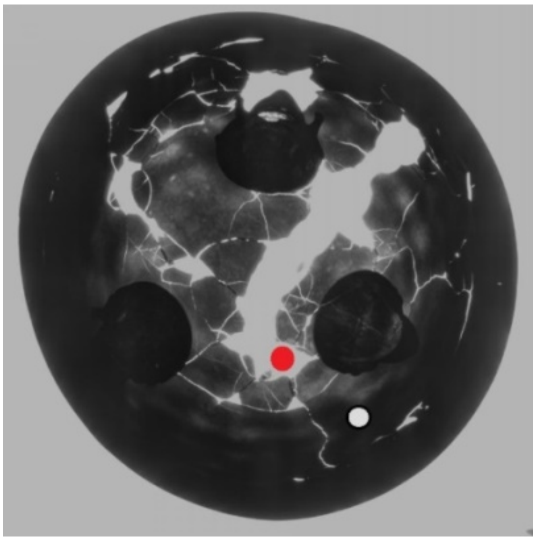


Figure A.20: X-Ray image on Mastòs inverted grey levels of the upper-lower projection.

The copper content was probably added voluntarily into alloy to modify properties rheological and mechanical properties of the melt, since the copper (above 3%) allow to increase the resistance of the silver and lowers the melting point. In correspondence with the gilded surfaces it has not been found the presence of mercury (attributable to the technique of gilding with amalgam) and consequently it is likely that the gold leaf has been applied to the silver surface by thermal treatment. Microscopic examinations have revealed in the centre of the cup NI 3 a small, convex protuberance in the middle (Fig. A.23). This feature has the same size as the garnets present in the other two similar cups, and it probably indicates that at the centre of the Inv. 3 cup once was a stone, presumably another garnet.

Each of these objects has a different technical history, as showed by the diagnostic results of analyses, consequently its behaviour with the environment can change in function of the chemical composition or technological processing [145]. The black tarnish of the silver, like in the Scylla medallion (Fig. A.24), can be caused by the presence of sulphide in the air, especially in unfavourable environmental conditions in which the relative humidity is over the 40%

It is more important to monitoring the environment and to recognize the causes of this decay as the pollution or the materials that constituted the showcase. The presence of the chloride salts in some silver items evinces a corrosion process that it

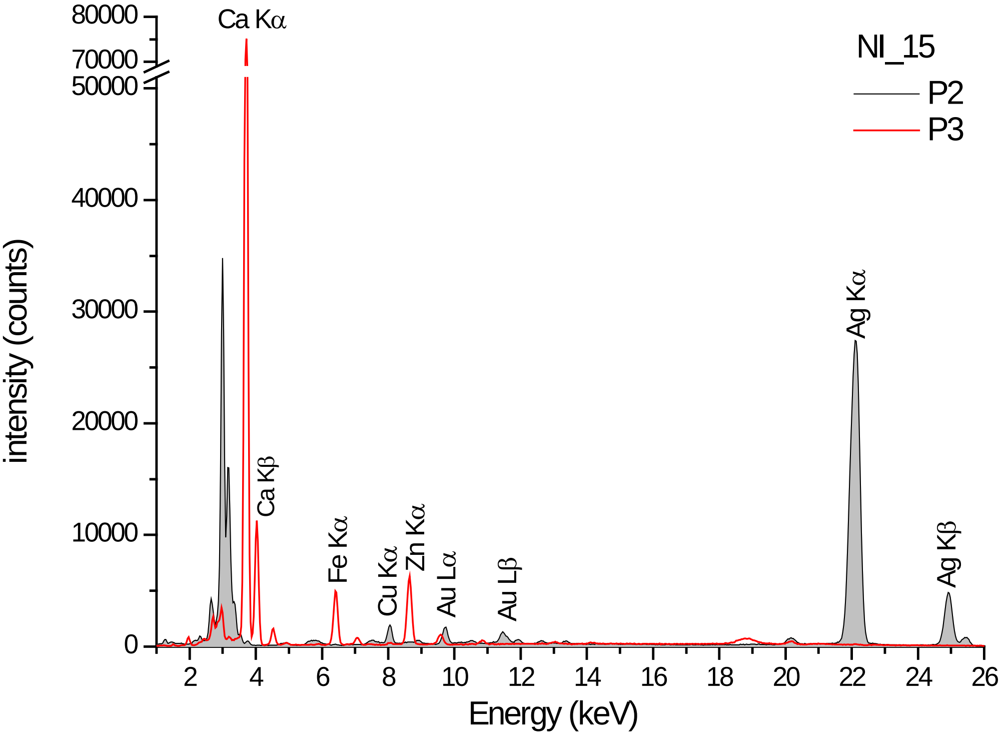


Figure A.21: XRF spectra acquired on the original surface (P2, grey) and on the integration (P3, red) shown in RX.

could be dangerous for the conservation of Morgantina Treasure.

A.2.3 Web Oriented and Android platform

In order to make the 3D models and the archaeometric data effectively available in a user friendly and integrated way, a web-oriented interface framework has been developed. Its main functionalities are the cataloguing of new 3D scans and the management of additional metadata, that can be implemented during the monitoring activities.

Through 3D scanning technologies the Morgantina silver gilt Treasure collection has been acquired to get 3D digital models, as described in the previous sections. In this work, the intent is also provide a user-friendly digital system to allow fruition and scientific analysis of the treasure pieces. In addition, I would make available

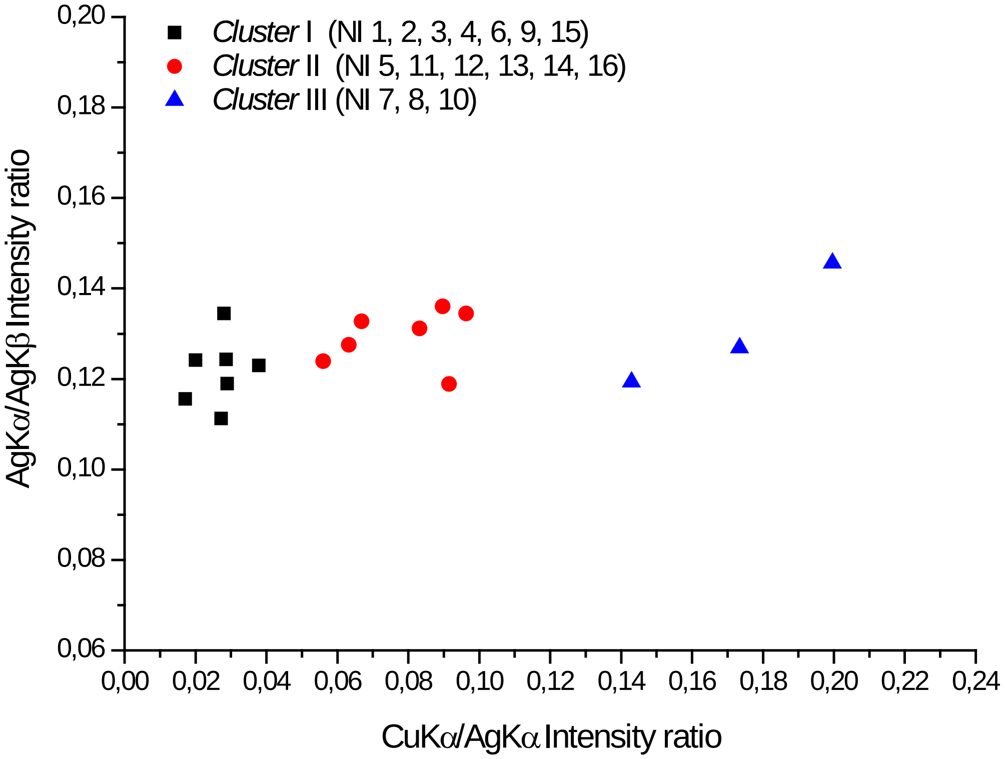


Figure A.22: Bi-plot of the whole set of Silver based on the XRF data relating to characteristic emission lines of copper ($K\alpha$) and Silver ($K\alpha$ and $K\beta$). Different colour highlights the three groups obtained on the basis of the Cu/Ag ratio.

results of aforementioned spectral analyses for research purpose. To this aim, I have developed a software for two kinds of platforms: Web application developed by Unity 5.0 and mobile Java application for Android. Although several 3D viewers already exist [146, 147], the aim is to realize a customized software which include functionalities specifically designed for cultural heritage scholars. Currently, the platforms described in this section implement elementary functionalities since they are prototypes thought to test the feasibility of a more long term project. Its main specs are the cataloguing of already existing or totally new 3D scans and the management of additional metadata. The digital version of the artifacts are augmented with semantic annotations about the history, measurements data, expert comments and so on. The meaning of term "semantic annotation" is the action and results

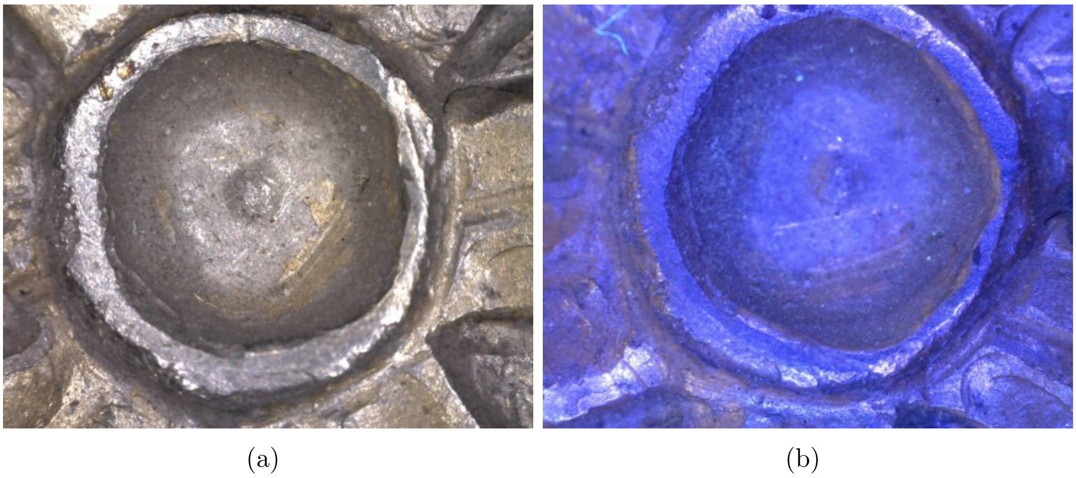


Figure A.23: Observation to the digital optical microscope (50×) of the central area of the emblem cup (NI 3), acquired both in the (a) visible and in (b) ultraviolet light

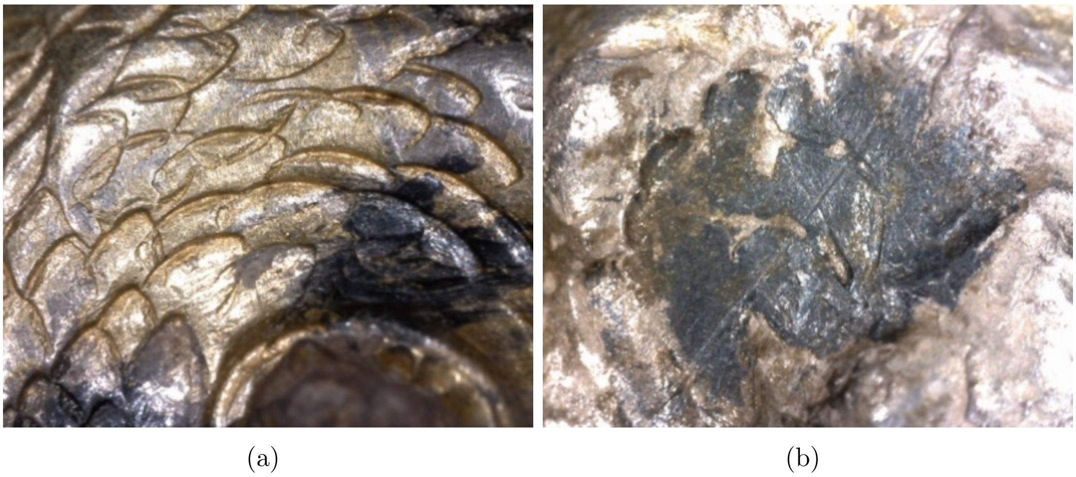


Figure A.24: Optical microscope visible observation (50×): details of black tarnish of the gilt silver on Scylla medallion.

of describing (part of) an electronic resource through metadata [148]. Firstly, a comprehensive description of software specification is given. Please note that this description focuses on the system functionalities, which are independent from the platforms employed (Web or Android).

Secondarily, a brief discussion about the technical details and the exclusive properties of the two platforms is reported. Opening graphical user interface gives a list of the available 3D models which are selectable for the investigation. Each of the 3D models is placed in a different 3D environment which offers conventional navigation functionalities, such as rotating and zooming. Hence, the surface and the finest details of each treasure artifact, can be examined from any point of view. Users are able to navigate around the 3D meshes through mouse, touchscreen or using the proper buttons on the GUI (it depends on the available input peripheral device). To properly analyse mesh surface, two visualization modalities are offered by the system: shaded mode and textured mode. Shaded mode is designed to permit an accurate geometric examination, because general shape and surface particulars look well marked with no texture data. This kind of surface analysis allows to visually detect alteration in the original form of the artifacts (e.g., deformation, missing parts). Instead, Texture mode shows material and colour information of the 3D models, which are very valuable to check the conservation state of the external surface. Chemical reaction (e.g., oxidation) or pigments scratches easily emerge if a texture analysis can be performed. Figs. A.25 and A.26 show Shade and Texture for Android platform respectively, while Figs. A.27 and A.28 show Shade and Texture modes for Unity platform.

The main feature of the proposed system is the semantic annotation which enriches the original 3D model with textual and visual data. Textual information gives a detailed description about some significant area of the artifact, while visual data (e.g., images and graphs) are useful to report analysis results and also for the comparison with the same artifacts in different time. Interactive parts of the meshes, are highlighted with well-noticeable markers, and when users select them, a tool-tip appears or a sided info-box shows the related info.

The proposed Unity Web system is mainly intended to assist experts to explore 3D models and consult analysis reports. The system is able to work by using a simple internet browser with no other specific client application. Moreover, the system can be accessible through internet to make available the 3D artifacts to researcher from all over the world. The prototype, has been developed by using Unity engine, version 5.0. It is an environment with an integrated game engine provided by Unity Technologies, which is typically employed to produce digital games for

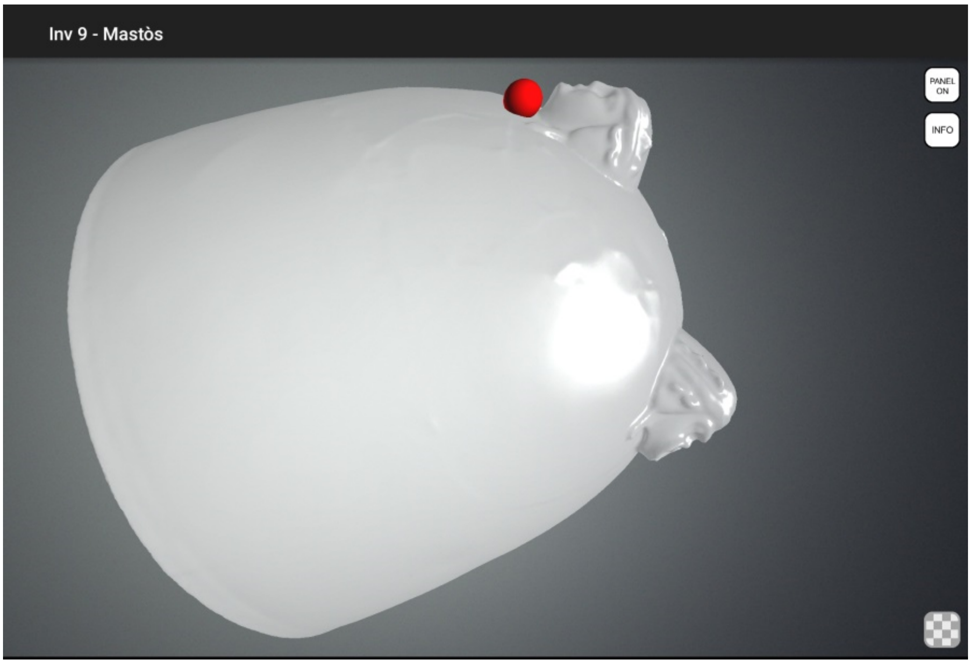


Figure A.25: Shaded version of a 3D model in Android platform. Red spheres are used as markers.

different platform, such as PC, consoles, mobile devices and websites. It allows to handle 3D model and other kinds of assets, such as material, light, image, and video. Unity 5.0 allows to encode the own algorithms in two different program languages: C# and JavaScript. In this work I employed C# and the Unity IDE called Mono Develop to implement the entire system. Although Unity is often used for digital game development, it is could be employed for generic purpose application related to 3D modelling. The main advantage of Unity is the simple way to manage multimedia resources and the user-friendly development GUI, as well as the multi-platform builder.

To give the possibility to test the proposed system, it is provided a demo version available at the following URL: http://iplab.dmi.unict.it/morgantinaJournal/.

The main aim of a mobile application (Android platform) is to follow user mobility. This lead us to develop a fruition system of Morgantina Treasure to enrich users experience during museum visits, which results especially useful when the original artifacts are lent to other institutions. Nevertheless, it has been decided to keep the semantic annotation feature in order to give the users historical information, as

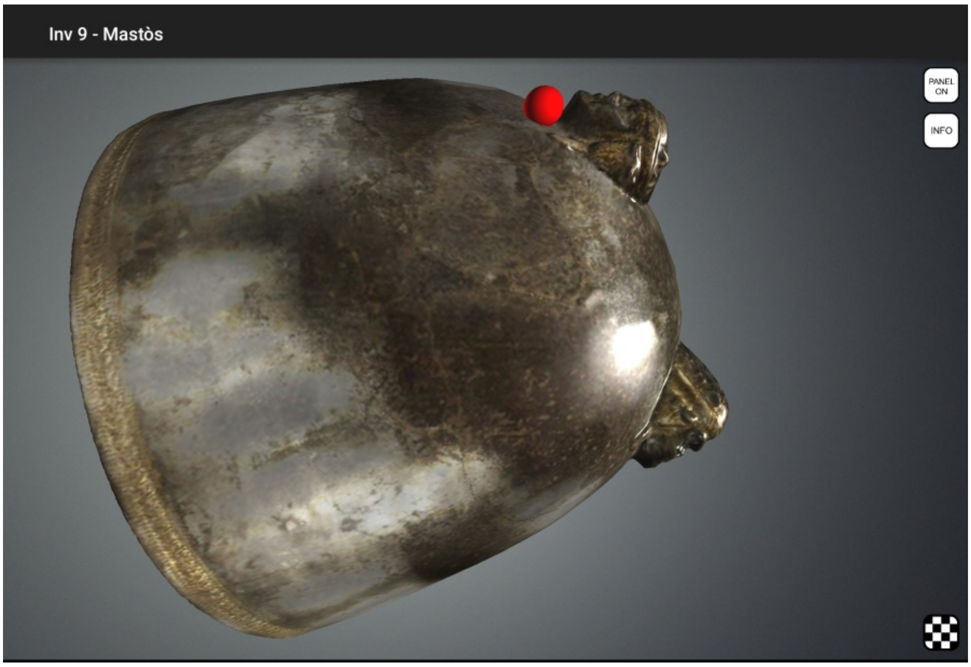


Figure A.26: Textured version of a 3D model in Android platform. Red spheres are used as markers.

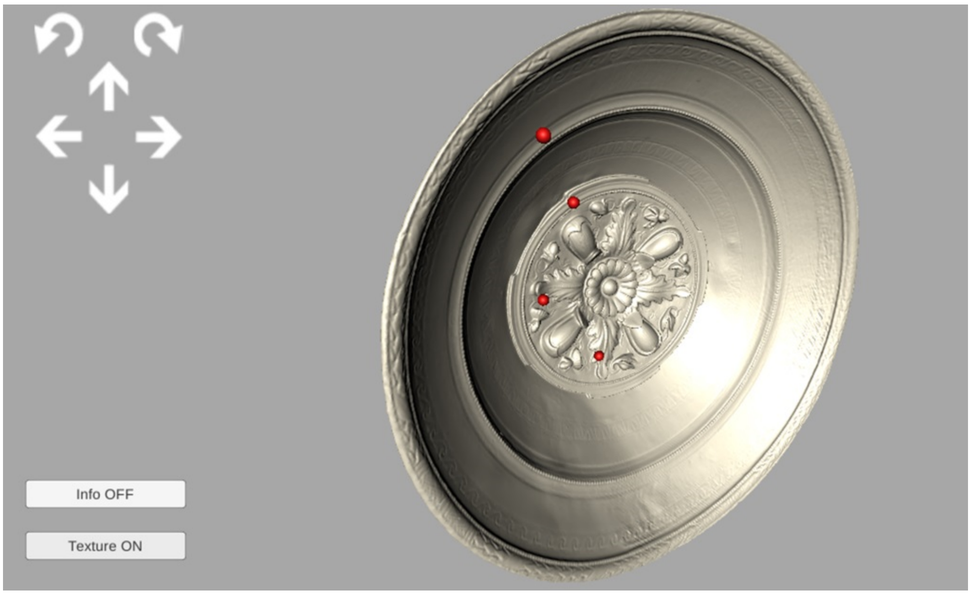


Figure A.27: Shaded version of a 3D model in Unity platform. Red spheres are used as markers.

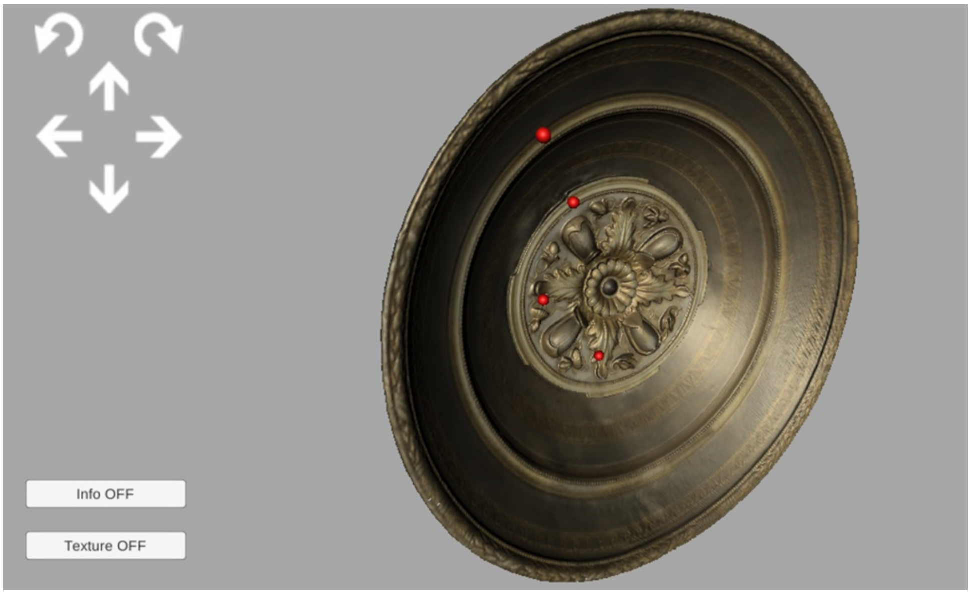


Figure A.28: Textured version of a 3D model in Unity platform. Red spheres are used as markers.

well as, a further platform for research purpose. The app has been developed by using Android Studio IDE and the build system Gradle, a plugin to assist project generation and maintenance. Java and the eXtensible Markup Language (XML) have been employed to encode the algorithms and the GUI of the proposed system. Specifically, Java was used to develop the function for handling 3D models, the user input and the interactions. On the other hand, XML provides a natty and standard tag scheme to define data structure e GUI. The 3D scene is drawn exploiting OpenGL ES (Open GL Embedded System), a subset of the standard OpenGL functions intended for embedded system, like smartphones, tablets and so on. Java interface for Open GL rendering calls is provided by Rajawali, a free library available under Apache License 2.0. Finally, to manage semantic annotation and related marker, it has been employed SQLite, the free Database Management System (DBMS) adopted by Android. The developed application can be found as demo version at the following link: http://iplab.dmi.unict.it/morgantinaJournal/. Tests have been performed on low-mid end device which mounts a CPU Intel Atom Z2520 Dual-core 1.2 GHz, a memory of 1GB, a GPU PowerVR SGX544 and the OS Android 5.0. Currently, despite the good application portability, it is not possible to ensure the correct functioning of all the devices and Android OS Version; a main requirement is an OS Android version 5.0 or higher.

A.2.4 Discussion

In this section I presented the results of a campaign of non-invasive diagnostic analysis (X-Ray, UV, XRF) and a 3D survey on the Morgantina Silver Treasure, in order to collect useful data for a twofold aim: monitoring the conservation state over time (to check after four years) and guaranteeing the virtual visit of the item during their absence.

The acquired 3D models and diagnostic data have been for the first time organized, in an integrated way, within a web-oriented platform and an Android application to increase the existing archaeological knowledge and to obtain referenced information of the conservation state. They were also used for the development of holograms now on display at the Museum of Aidone.

The ongoing web-oriented platform and the Android app consist of an active tool to management of metadata, which will gradually be implemented through knowledge acquired by specialists and at the same time contribute to the valorisation of these archaeological findings to the wide public. All the findings of the archaeometric analyses have been included in these digital platforms.

As future works, I am planning to conduct further analysis on the piece of Morgantina Treasure and to improve the functionalities of both the described platforms.

A.3 Low Cost Handheld 3D Scanning for Architectural Elements Acquisition

In the last years some new low cost emerging technologies have been released on the market delivering a long term dream of the practitioner of cultural heritage: fast, accurate, low cost 3D scanning with a handheld device. Envisioning the massive use of these cheap and easy to use devices in the next years, it is crucial to explore the possible fields of application thus testing their effectiveness in terms of easiness of 3D data collection, processing, mesh resolution and metric accuracy against the size and features of the objects. In this study I focus the attention on one emerging technology, the Structure Sensor device [149], in order to verify a 3D pipeline acquisition on an architectural element and its details. As case study we choose the XVIII century doorway placed in the monastery of Benedettini in Catania, in UNESCO's world heritage list. The doorway presents both planar, complex (mouldings) and sculpted surfaces and allow us to carry out several tests on different geometries. The goal is to outline a 3D pipeline following as much as possible, a low cost and open source workflow from 3D data collecting to the digital replica.

The methodological approach foresees the assessment of the 3D acquisition procedure in comparison with data obtained by a Time of Flight device in order to point out weaknesses and advantages of the hand held scanning approach in relation to other well assessed technology. Then 3D modeling issues are explored and discussed to obtain a digital replica in an open source environment suitable for architectural representation and communication purposes.

A.3.1 Handheld 3D scanning

The 3D scanners are devices which are able to collect geometry information about a real-world object or environment. Then these information are processed in order to build a digital 3D models of the scanned elements. Nowadays, 3D scanning devices play a key role in many research field and applications such as industrial, prosthetics and medicine prototyping, cultural heritage preservation and documentation, etc. [150, 151, 152, 153]. Since these devices works by employing many different technologies and their cost change in a wide price range, it is important to select the best solution for your own applications.

The most common technologies employed for 3D scanning are triangulation (e.g., laser triangulation or structured light) and Time of Flight (ToF). The first technology consists in a laser emitter and a sensor which receives the reflected beam with a certain angle. Distance between the surface point and the scanner, can be computed by starting from the emission and reception angles plus the distance between sensor and emitter. The Time of Flight scanner finds the surface distance by measuring the round-trip time of a pulse of light. A electromagnetic wave is emitted by the scanner, and the time before the reflected wave is received by a sensor is measured. Since the speed of light is known, the round-trip time allows to compute the travel distance of the wave. Finally, the structured light technologies are based on the emission of a light pattern (e.g., a grid, a set of strip), that is altered when it hits a surface. Hence, the geometry data of the hit surface is inferred by the extent of the alteration.

Sensors that exploit these technologies belong to the class of the so-called active sensors. Indeed, these devices "emit" electromagnetic waves on the objects to estimates their geometrical properties. On the other hand, sensors which do not introduce waves in the environment are called passive sensors. In this latter case, the 3D acquisition could be achieved for instance by stereo vision or structure from motion [154].

From another perspective the 3D scanning devices can be categorized in respect to their portability. In recent years, thanks to the miniaturization and integration of the electronic and optical sensors, has been possible to produce small and compact high performance 3D scanners [155, 156]. Hence, we may further distinguish two kinds of devices: handheld scanners and not portable ones. Today, the emerging handheld scanners are a remarkable resource for affordable price and good performance and the convenience ensured by the portability. For the relative low-cost and usability, most of these devices became consumer electronics product, while other are still used in professional context. However, they represent a great resource in the field of Cultural heritage. Below, I report a list of the main handheld 3D scanners and their knows specs in Table A.2:

Microsoft Kinect is mainly used in home videogames entertainment. However, some examples of applications of these devices to cultural heritage could be found in the works of Cappelletto [157] and Remondino [153]. **Scanify Fuel 3D** is a handheld device, which exploits combination of photometric and stereography techniques to acquire depth information, so it can reach a high accuracy. **Google Project Tango** is a Google device with exploits motion tracking to understand position and orientation of the device user. It is particularly suitable for augmented reality application. **Artec Eva and Artec Spider** are two semi-professional active scanners produced by Artec 3D company. The first one has a high resolution and it is suitable for small and detailed object, while Artec Eva is though for architectural elements such as doors, statue etc. **Structure Sensor** is a small active scanner produced by Occipital. It exploits structured light technology to guarantee a good quality scan with a low expense. This device has been employed in the study

Table A.2: Specs of the described handheld scanners

Sensor	Accuracy	Resolution	Acquisition Speed	Texture
Kinect V1	n.a.	n.a.	30 fps	Yes
Kinect V2	n.a.	n.a.	30 fps	Yes
Asus Xtion PRO Live	n.a.	n.a.	n.a.	Yes
Scanify Fuel 3D	0.35 mm	n.a.	10 fps	Yes
Google Project Tango	n.a.	n.a.	n.a.	Yes
Artec Eva	0.1 mm	0.1 mm	2,000,000 per second	Yes (standard ver.)
Artec Spider	0.05 mm	0.1 mm	1,000,000 per second	Yes
Structure Sensor	0.5 mm	1.0 mm	30/60 fps	Yes (with iPad)

conducted on this section, hence more details are provided in the following sections.

A.3.2 Structure sensor scanning for architectural elements

In most cases the use of handheld scanners is limited to small objects (approximately a volume of $1m^3$), if there is the need to acquire bigger objects, then it is necessary to carry out several scans and then align them in an unique model. Thus, in architectural heritage field the use of this kind of scanner should be recommended only for architectural details (basis, capitals, pedestals). Nevertheless, in this study I explore the possibility of using Structure Sensor also for bigger architectural element such us a doorway. The goal is to provide a full low cost and open source 3D pipeline highlighting potentialities and weakness. The study, is conducted on an eighteen century doorway in Benedettini monumental complex in Catania (UNESCO heritage) located in the gallery at the first floor of the monastery and it provides access to one of the cells of the friars, nowadays used as offices for the Department of Humanities of Catania University. This doorway, realized with limestone, is made by the plane surfaces of the jambs and architrave, the complex surfaces of the moldings (bed cornice, cymatium and tympanum), the sculpted decorations of the frieze and the capital. So I tested the performances of this sensor both on the details and on the overall shape of the doorway. The study is completed by a metric accuracy test that uses as ground truth a ToF scan [158].

Employed device

In this case study I employed the Structure Sensor (Fig. A.29). Similarly to Microsoft Kinect, this device has an operative range capability from $0.4m$ to $12m$.

Indeed, in the closer range from the sensor the device reaches a declared 3D point accuracy of $0.5mm$. The accuracy become smaller if the scanned object is placed over $3.5m$ or if the scanning volume is increased. Structure sensor is an infrared structured light 3D device, hence several issues are related to it: it does not work well in outdoor environment, since sunlight is a too strong source of infrared interference. However, the case study is located in an indoor environment not affected by direct sunlight interferences. Another critical issue is related to the material of the surface of the scanned objects. Infrared waves can be reflected, absorbed or distorted respectively by not opaque, black, or transparent surfaces, as glassy, plastic or polished objects. The case study is composed in the majority by opaque materials such as the limestone in the door jamb and decorations. The handle and the label of the door are in a polish metal but they have been still acquired with just some light distortion (Figure A.30). A possible third issue related to the Structure Sensor could be related to object with “poor geometry”: in the acquisition phase the sensor needs a minimum amount of geometrical details of the object to be scanned. This is required for an optimal frame-by-frame mesh reconstruction. If not enough geometry is provided then the sensor will prompt an error message and the acquisition will fail. This problem occurs in case of particularly flat object. The case study present a geometry complex enough to enable a good acquisition with the employed device.

Structure sensor can only acquire depth information by itself. In order to add some texture information an external RGB camera is needed. Structure sensor also needs an external computation unity to process acquired data. Usually Structure is attached or connected to an Apple iPad exploiting a wired connection and in this way textures can be acquired exploiting the standard RGB camera of the tablet. Although this is the most common way to use the sensor due to its practical aspects, this acquisition method is discarded, since the final model is decimated before the exportation from the tablet resulting in a too low quality mesh. Exploiting proper software like Skanect [159], it is also possible to connect the Structure and the tablet to a computer through a wireless network, or just the Structure with a wired one. In the latter case we do not acquire any color information reaching a real-time acquisition of the case study. Note that texture is not really needed to estimate the mesh (e.g. the geometry) of the scanned object. Latency during acquisition is an issue that must be taken into account: sensor could lose or estimate a wrong

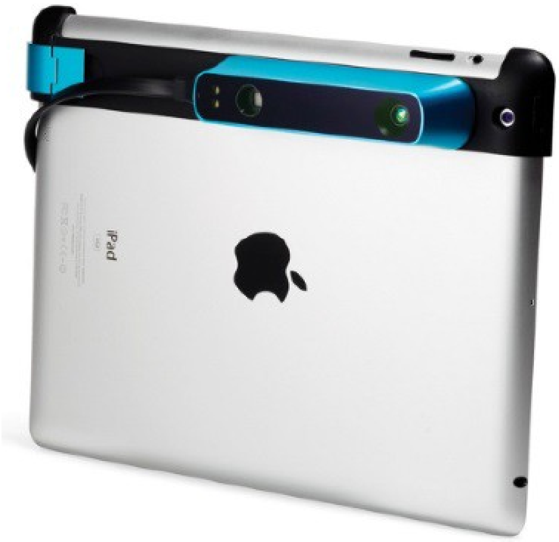


Figure A.29: Structure Sensor onto an iPad

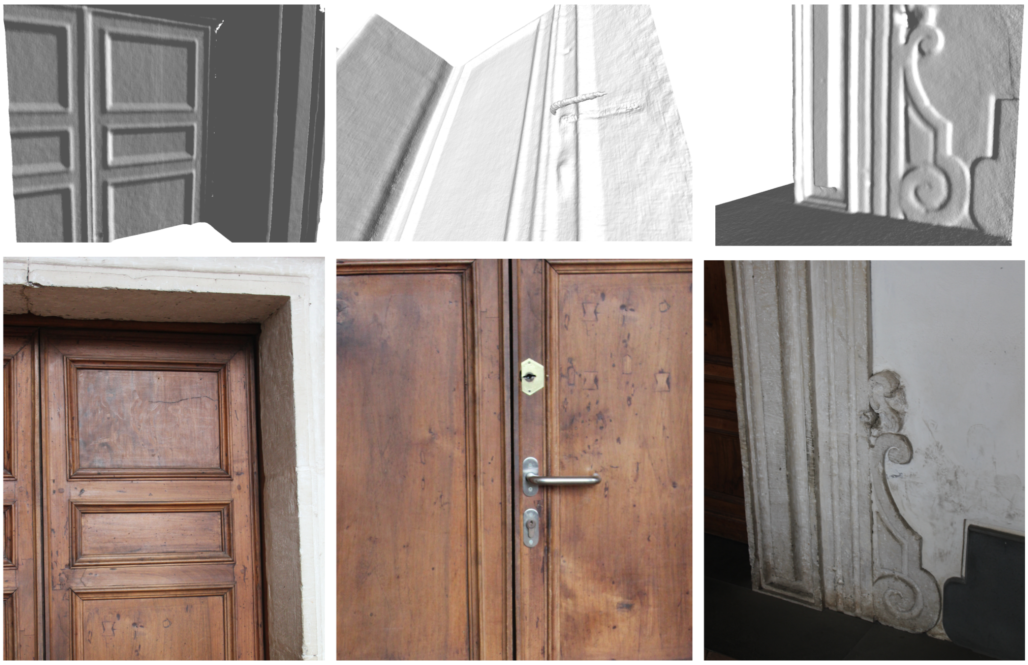


Figure A.30: View of the Structure Sensor behaviour on three different materials (from left to right): wood, metal, limestone.

alignment through consecutive acquisition instants, introducing noise or, in the worst case, requiring to restart the whole acquisition.

Acquisition

As said, the resolution of the final mesh is strictly related to the scanning volume. The case study is a door with an height of almost $3m$ and a width of almost $2.5m$ resulting in a scanning volume too large in order to obtain a quality sufficiently good. For this reason, it has been decided to set a scanning volume of $1m^3$ and to subdivide the acquisition of the door into several single acquisitions. I acquired a total of 23 parts, starting from the bottom left position until the top right. Note that the acquisition range depends also on the sensor, so the scanning volume could be set larger than the reported $1m^3$, but with limited precision. The 23 parts have been carefully acquired with at least the 30% of overlapping between each other. This redundant information is required to correctly perform the alignment process of the subparts into the full model. The meshes are processed and aligned by exploiting the software Meshlab [160]. We acquired highly detailed meshes, with an average number of $600K$ vertices and $1M$ faces. We perform a preprocess phase to reduce the noise, as some isolated face or vertex, using the Quadric Edge Collapse Decimation of Meshlab. We discard the 80% of the points in each mesh without any visual-perceptible loss of details. Then, using the Point Glue tool of Meshlab we perform all the required alignment and saved the final model of the case study in the common OBJ format.

Comparison with Time of Flight 3D scanning

In this subsection are reported the results obtained during the visual and metric accuracy tests. As ground truth I use a ToF mesh model. The pipeline followed is by the time used in literature [153, 161] and foresees the alignment of the different models in the same reference system and the calculation of the distance between the meshes by means of Hausdorff distance algorithm application [162]. Considering the performances of the handheld scanner and the purposes of this work I consider both two details (a capital and frames and mouldings of the jams and entablature) and the overall doorway.

During ToF laser scanner acquisition (using a HDS 3000 by Leica Geosystem) it has been decided to carry out three scans: one frontal and two lateral and I choose

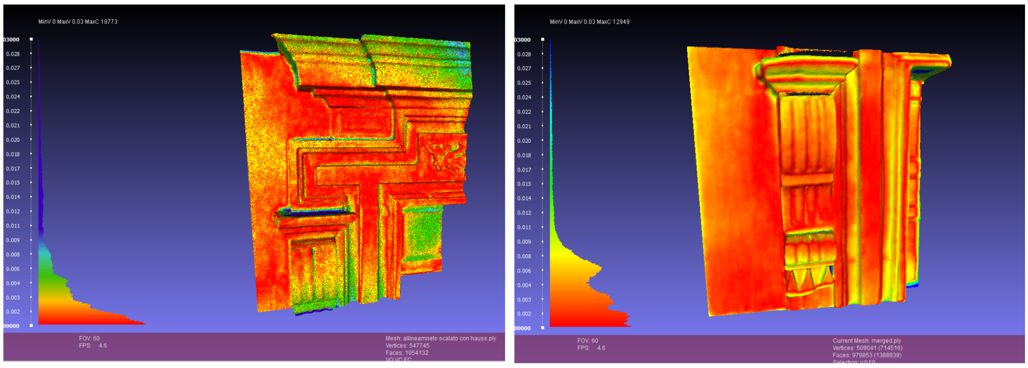
Table A.3: Experimental results. All values are expressed in mm.

Model	**Hausdorff Range**	**Mean**	**RMS**
Capital	$0 - 30$	4.344	6.879
Entablature	$0 - 30$	4.775	6.797
Overall Model	$0 - 50$	9.619	14.104

a scan step very dense (about $2mm$) to have a very detailed point cloud. In these cases, as reported in previous literature works [163], the size of the noise exceeds the sampling rate so that it hides most of the details: in the following meshing phase it is mandatory to apply a specific combination of surface reconstruction and smoothing algorithms in order to avoid spikes meshes. In Meshlab I carry out the merging of the scans into a unique model, then I apply the pipeline employed in Ref. [163] by testing and choosing the parameters that better smoothed the surfaces without losing details.

A first consideration that can be done, in terms of visual accuracy of the 3D reconstructions, is that the Structure Sensor single scan models are more detailed and less noisy with respect to ToF reconstructions. This is in line with the kind of used sensor. The comparison between the three models (two details and the overall doorway) and their corresponding ToF scans was carried out in Meshlab. As for the two details, the alignment between Sensor Structure model/ToF model involved an alignment error of $3mm$. The range calculation interval for Hausdorff distance is $0 - 30mm$. The second test involves the overall model of the doorway. A detailed visual analysis of the Structure Sensor model reveals some mismatches in the overlapping areas. These alignment errors could be interpreted as fallacies of the alignment step probably due to boundary geometric inconsistencies of the single scans. In order to take into account these mismatches, I calculate Hausdorff distance with the following range values $0 - 50mm$.

The experimental results are shown in Table A.3. Furthermore, it is very interesting to read the trend of the histogram and observe the distribution of the distances between the two meshes directly on the 3D model (Figures A.31 - A.32), where the red color means the minimum distance between the two meshes and the blue means the maximum one.



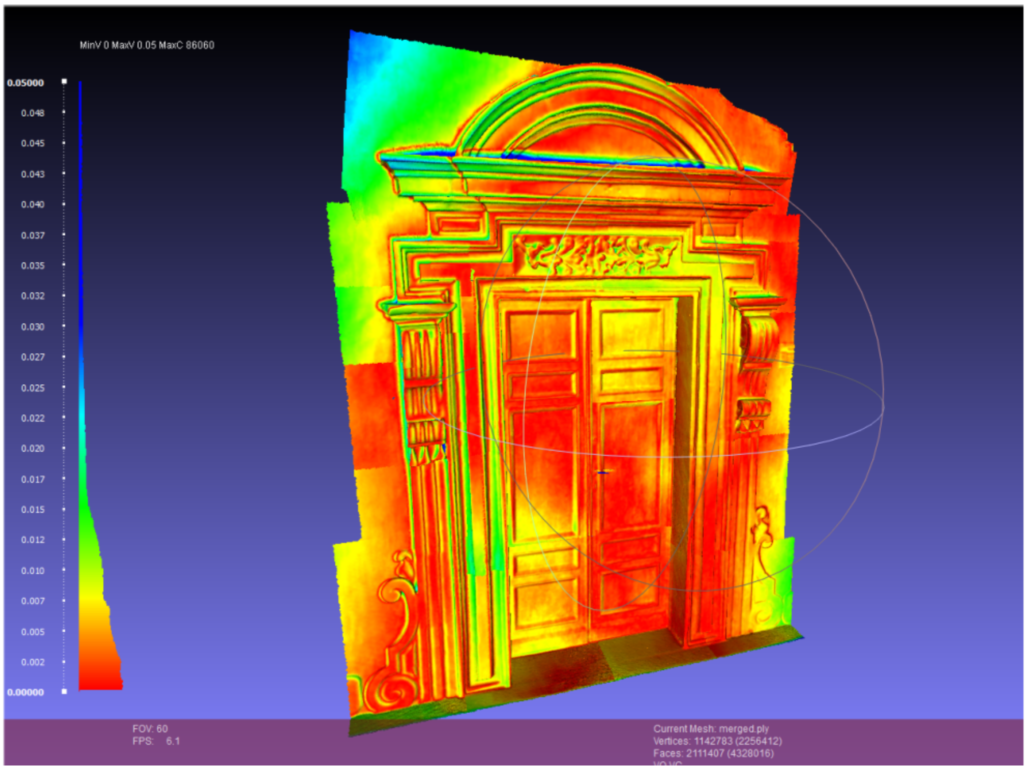
FOV: 60
FPS: 4.6

Mesh: allineamneto scalato con hauss.ply
Vertices: 547745
Faces: 1054132

FOV: 60
FPS: 4.6

Current Mesh: merged.ply
Vertices: 509041 (714516)
Faces: 979853 (1388939)

Figure A.31: Hausdorff distance and subsequent quality histogram between TOF model and Structure Sensor model of two chosen details.



FOV: 60
FPS: 6.1

Current Mesh: merged.ply
Vertices: 1142783 (2256412)
Faces: 2111407 (4328016)

Figure A.32: Hausdorff distance and subsequent quality histogram between TOF and Structure Sensor models.

A.3.3 Discussion

In this section it was defined a low cost indoor procedure facing the criticalities of the handheld 3D scanner Structure Sensor for architectural elements acquisition. Furthermore, the metric accuracy test highlighted the reliability of this sensor for the details acquisition. Indeed, as shown in Table A.3, the Mean distance computed on details is lower than $5mm$, comparable with the usual ToF accuracy. On the other hand, the Mean distance computed on the whole model is $9.6mm$, due to the severe amount of noise introduced by alignment process. These results demonstrate that this sensor can obtain high quality 3D models of architectural details, useful to integrate ToF scannings and make the digitalization of the cultural heritage easier and faster with affordable economical efforts.

A.4 Virtual anastylosis of Greek sculpture

This section deals with a virtual anastylosis of a Greek Archaic statue from ancient Sicily and the development of a new public outreach protocol for those with visual impairment or cognitive disabilities through the application of 3D printing and haptic technology. The case study consists of the marble head from Leontinoi in southeastern Sicily, acquired in the $XVIII$ century and later kept in the collection of the Museum of Castello Ursino in Catania, and a marble torso, retrieved in 1904 and since then displayed in the Archaeological Museum of Siracusa. Due to similar stylistic features, the two pieces can be dated to the end of the VI century BC. Their association has been an open problem, largely debated by scholars, who have based their hypotheses on comparisons between pictures, but the reassembly of the two artefacts was never attempted. As a result the importance of such an artefact, which could be the only intact Archaic statue of a kouros ever found in Greek Sicily, has not fully been grasped by the public. Consequently, the curatorial dissemination of the knowledge related with such artefacts is purely based on photographic material. As a response to this scenario, the two objects have been 3D scanned and virtually reassembled. The result has been shared digitally with the public via a web platform and, in order to include increase accessibility for the public with physical or cognitive disabilities, copies of the reassembled statue have been 3D printed and an interactive test with the 3D model has been carried out with a haptic device.

A.4.1 The case study: an Archaic kouros from Leontinoi?

The problematic case study is represented by two matching pieces of a statue kept in two different museums, the reputation of which can be restored via an exercise of virtual anastylosis. The research is developed through five main steps:

- 3D scanning of the two objects;
- virtual anastylosis;
- use of a web platform for public sharing;
- 3d printing of the reassembled statue;
- learning experience via haptic devices.

Greek Archaic sculpture is dominated by the production of statues of young naked boys, known as kouroi (plural of kouros meaning 'boy' in Greek), and young girls with long dresses, named korai (plural of kore meaning 'girl' in Greek), which have religious or funerary significance and for this reason are generally offered as ex-voto in sanctuaries or placed above or by tombs in cemeteries [164]. The statues were the symbolic representation of the worshippers consecrating their lives to t deities or idealized portraits of the dead. In Greek Sicily, there are several remarkable examples of kouroi and korai imported from Greece or locally produced, and some of them can certainly be considered as masterpieces of Greek statuary [165].

However, very few life-size statues were found intact. After the Classical period it became customary to detach the heads of Greek statues in order to create head-portraits. In fact, with a few exceptions of smaller scale statues found intact, this class of Greek statues in Sicily is represented just by heads without matching bodies and headless bodies. A unique case is the one of the "Biscari head" kept at the Museo Civico "Castello Ursino" di Catania and of the torso from Leontini in display at the Regional Archaeological Museum "Paolo Orsi" of Siracusa, both made of marble, dated between the end of VI - beginning of V century BC and almost unanimously believed to be part of the same life-size kouros.

The head (Fig. A.33), also known as the 'Biscari head', recovered in the site of the Greek city of Leontinoi, was exhibited for a long time in the Hall of Marbles of the Museum of Palazzo Biscari alla Marina before being incorporated in the main

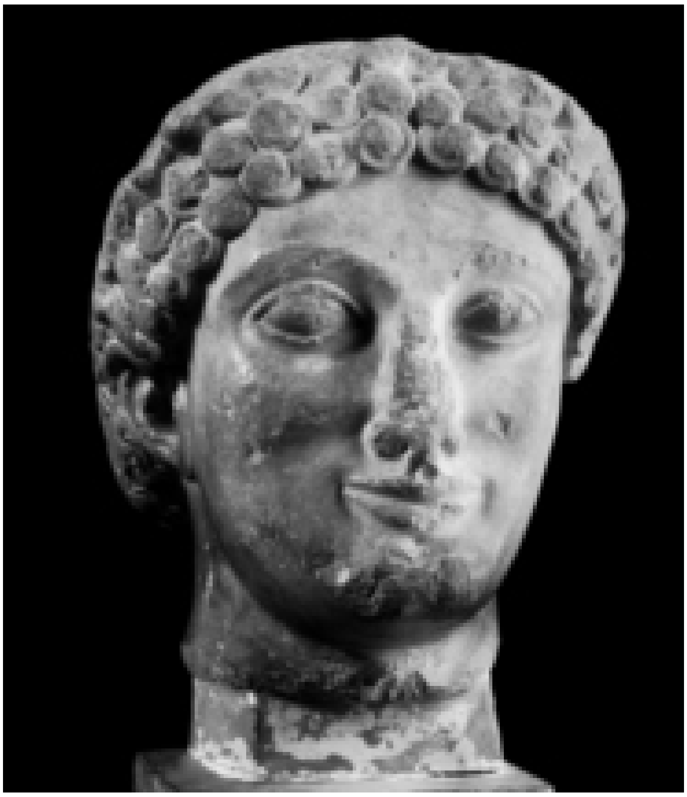


Figure A.33: The Biscari head (ca. 1938).

collection of the Museo Civico "Castello Ursino" of Catania [166, 167]. The torso (Fig. A.34) was accidentally found in the country right outside the area of the ancient colony of Leontinoi and purchased in 1904 for 1000 liras by Paolo Orsi from the Marquis of Castelluccio, who was another famous collector of antiquities. As separated artefacts, the two pieces were subject of several studies aimed to define their style, chronology and eventually, their provenance.

The first scholar who suggested a possible association between the head and the torso was Guido Libertini in the 30's. He produced a gypsum cast of the head in order to compare it with the torso to verify his hypothesis. Although a missing part of the neck did not allow for a perfect match, the volumetric correspondence together with the stylistic analogies were enough to support the idea that the two pieces were once a life-size kouros from Leontini. Unfortunately, no documentation has been recovered regarding this experiment. Many decades after, Gino Vinicio Gentili reappraised the problem of the association of the two pieces using a photofit

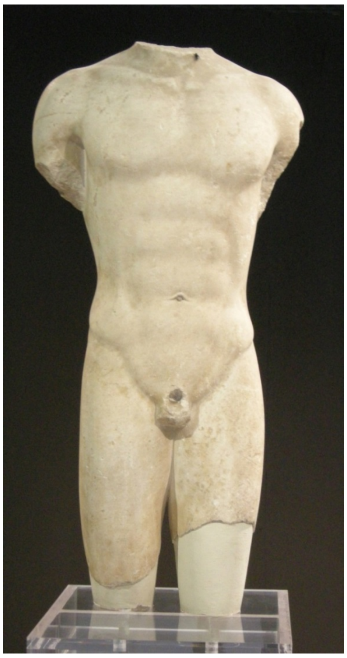


Figure A.34: The torso from Leontinoi.

(Fig. A.35), in which he matched the photographs of the head and the torso [168]. This further confirmation of Libertini's hypothesis was published in a scientific paper with a very limited distribution. Again, the general public missed the remarkable discovery of the first intact Sicilian kouros.

In order to go beyond the exercises of Libertini and Gentile and to provide the final proof of the compatibility of the two pieces as part of the same statue, a reconstructive study has been carried out based on the 3D scanning and virtual anastylosis of the kouros of Leontinoi.

A.4.2 Acquisition and Data Processing

The acquisition was carried out with extreme care in order to properly capture the many anatomical details of the two pieces (Fig. A.36). The scanning was performed using the Structure Sensor (see Section:A.3) connected through Wi-Fi to Skanect. The scan volume was set to $0.6m^3$ for the head and to $1.2m^3$ for the torso.

Both artefacts were placed on a pedestal; in particular the head was placed steadily on a metal support. After the 3D capturing, 3D models were manipulated

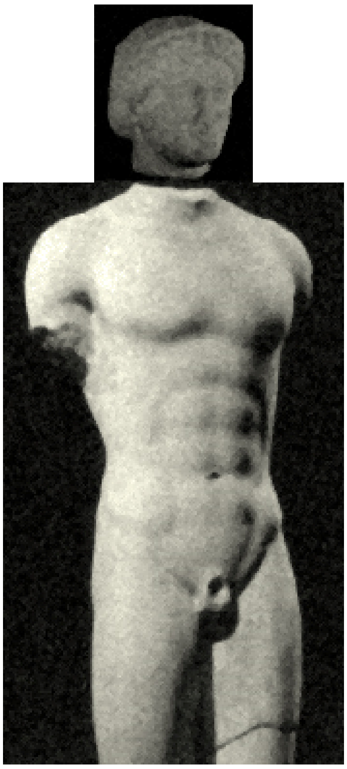


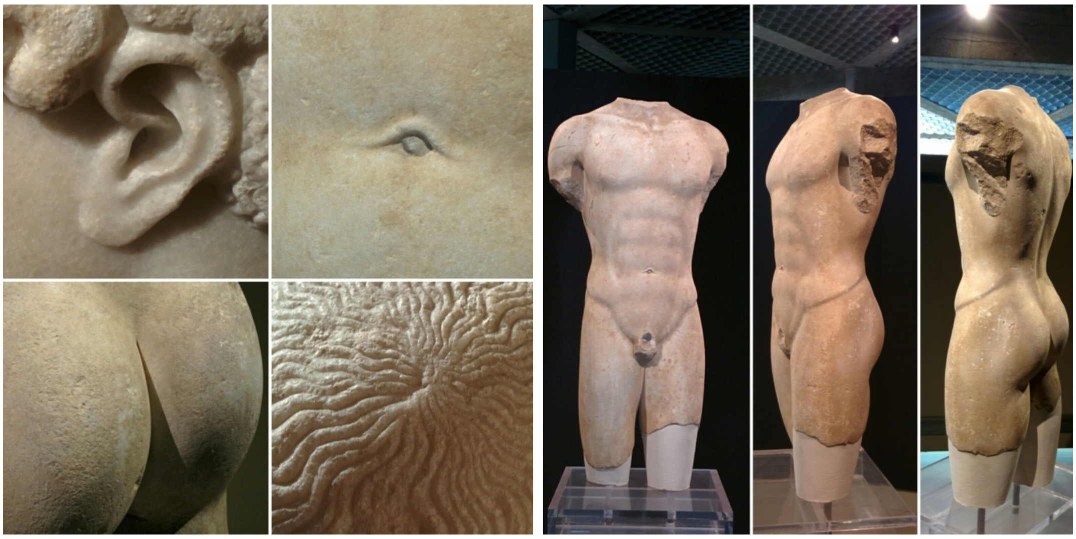
Figure A.35: Possible photofit of the head and the torso.

with two popular software among archaeologists: Meshlab and Blender. Meshlab [160] was employed in order to refine the models in a pre-processing phase: after digital acquisition the vertices extraneous to the artefacts were deleted (Fig. A.37).

In Meshlab, it was possible to take digital measures of the head and the neck is order to verify an eventual dimensional compatibility. As shown in Fig.A.38, dimensions of the lower part of the neck of the head are $12.67 \times 13.67cm$, while those of the upper part of the neck of the torso are $16.50 \times 13.27cm$. Such dimensions, considering the possibility of physical decay of the edges, makes dimensional compatibility between the two pieces likely.

Furthermore, when comparing the height of the head with that of the preserved torso (A.39) it is clear that they are proportional to one another.

Subsequently the models were imported into Blender [169]; in that virtual environment, the head and torso of the kouros were manually aligned. Technical and archaeological analysis have shown that the statue is missing part of the neck, which



(a) (b)

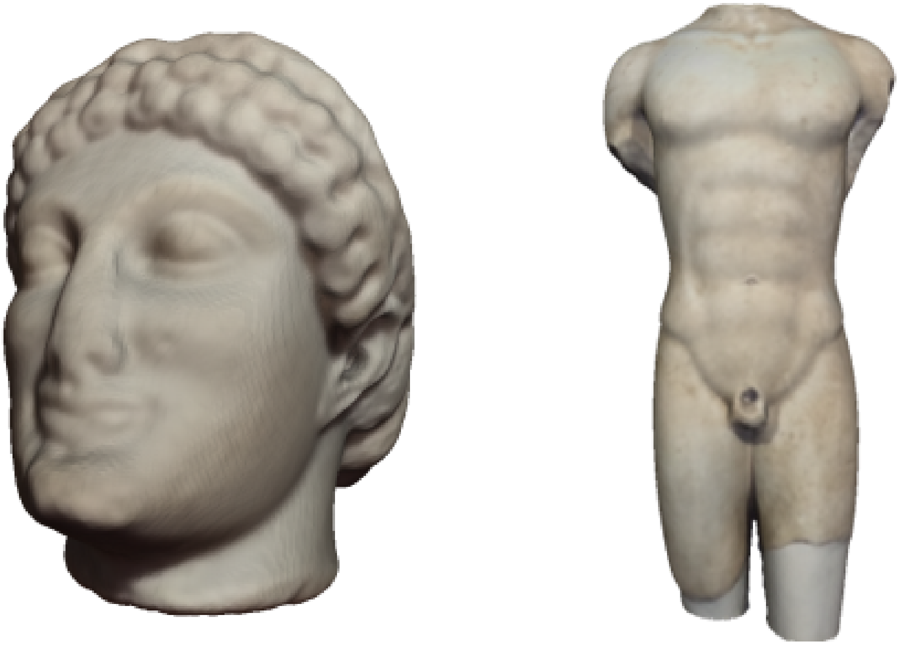
Figure A.36: (a) Details of anatomical features of the head and torso; (b) Views of anatomical features of the torso.

can be reflected in the model, obtaining the results shown in Fig. A.40.

A.4.3 Sharing the virtual kouros of Leontini with the public

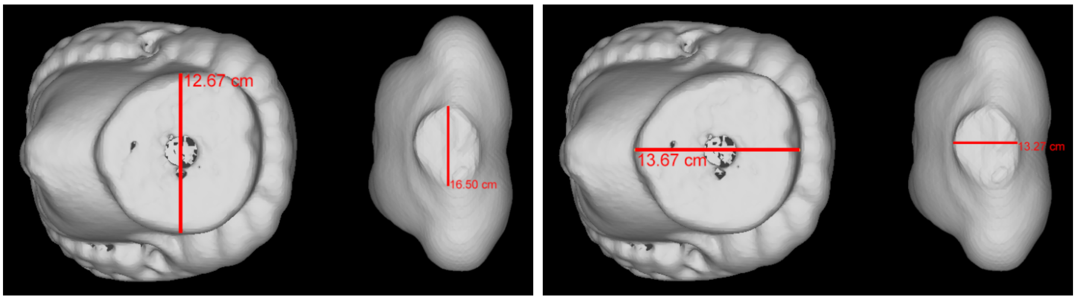
The research presented in this section has clearly demonstrated that the hypothesis suggested in the first place by Libertini was correct. The two pieces are certainly part of the same statue, as they did not just share the same stylistic features, but they are also compatible in terms of geometry. The virtual anastylosis has, in fact, added a further level of information not present previously (A.41). The statue seems very proportionate and the head, even in absence of a perfect match due to the lack of a segment of the neck, perfectly fits to the body.

A simple exercise of virtual anastylosis has given back to the community of scholars the first realistic representation of the kouros of Leontinoi, the first life-size statue of an Archaic kouros from Greek Sicily. How would it be possible then to share with the public this remarkable discovery? How will the reputation of the two artefacts be improved by such a discovery? Due to strict management policies, none of the two museums will surrender one of the pieces to the other in order to



(a) (b)

Figure A.37: (a) Textured 3D model of the head; (b) Textured 3D model of the torso.



(a) (b)

Figure A.38: (a) Phases of digital measuring with Meshlab, diameters of head and torso; (b) Phases of digital measuring with Meshlab, other diameters of head and torso.

recombine the pieces and allow just one of the two institutions have it in display. This suggests that the general public will never know about the kouros of Leontinoi and will never have the chance to see it in full.

In response to this scenario, the web platform developed for Morgantina Silver

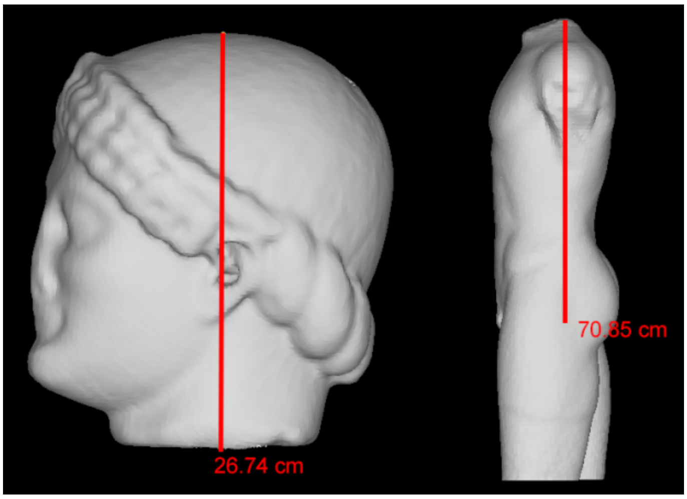


Figure A.39: Phases of digital measuring with Meshlab, heights of the pieces.

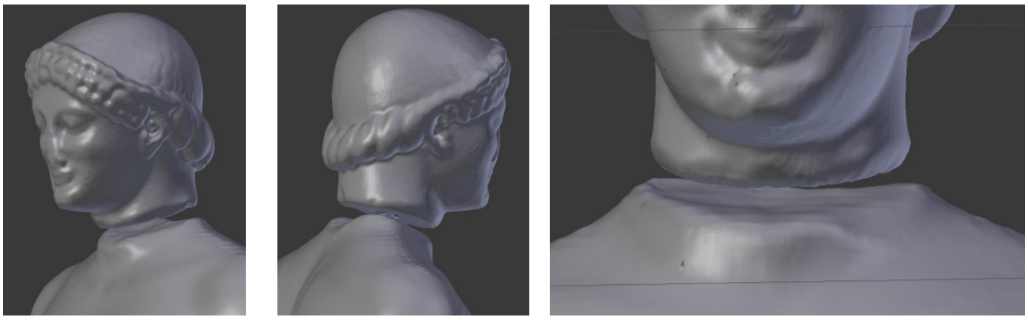


Figure A.40: Manual alignment of the 3D models of the head and the torso in Blender.

Treasure (see Section A.2) has been employed in order to share in a simple and effective way the results of this research (`http://iplab.dmi.unict.it/kouros/`).

A.4.4 3D printing

The next step of this research effort was to create a physical copy of the statue in scale 1:10 through 3D printing (Fig. A.42).

After final processing and digital corrections, the 3D model was converted to .STL format and sent to the printer after slicing. The model of the statue was fabricated on a highly customized Delta robot-type FDM (Fused Deposition Modeling) 3D printer at the University of South Florida labs. For enhanced part accuracy, the

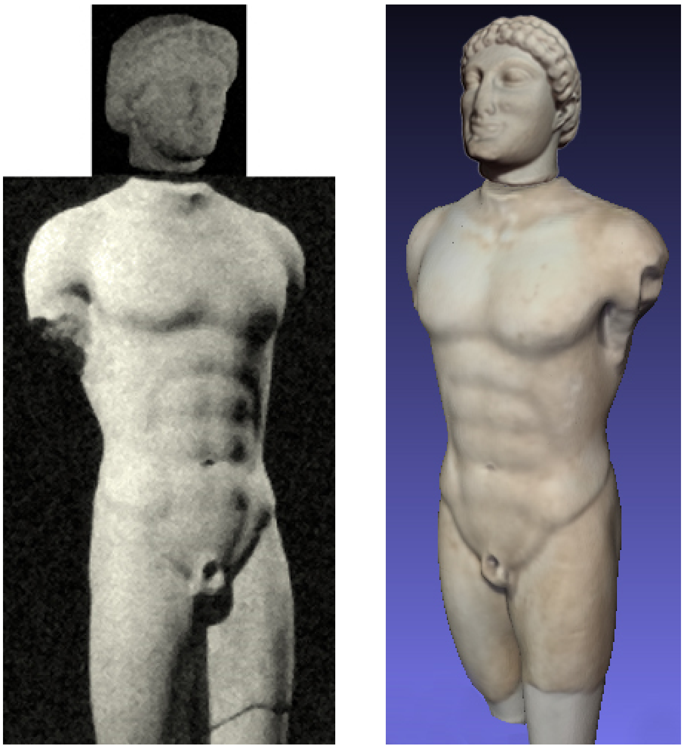


Figure A.41: Comparison between the photfit and the virtual anastylosis of the kouros of Leontinoi.

effector of this machine is held in place using a low friction magnetic suspension system. The positioning accuracy of this delta robot is better than $50\mu m$ in the x, y, and z directions. A low-force optically triggered z-probe was used to calibrate the build plate surface prior to printing to enhance print reliability and adhesion. The printing material selected was white PLA (polylactic acid) which was extruded at a temperature of $205°C$. This particular polymer was selected due to its ability to resist warpage and shrinkage which might cause layer delamination on an object of this size. To further minimize warpage, the build plate, made of glass with water-based acrylic glue adhesion promotor, was heated to $55°C$. Ambient conditions during printing were $26°C$, humidity $52 - 60\%$. Slicing layer height was set to $0.15mm$ ($150\mu m$) with a relatively low 12% part fill density. An extrusion nozzle of 0.4mm in diameter was used. Mechanical supports were enabled to ensure printing of overhanging and highly sloped geometry would be successful. Total print time was $\sim$ 24 hours and consumed $\sim 170g$ of polymer. Post print work-up was kept to

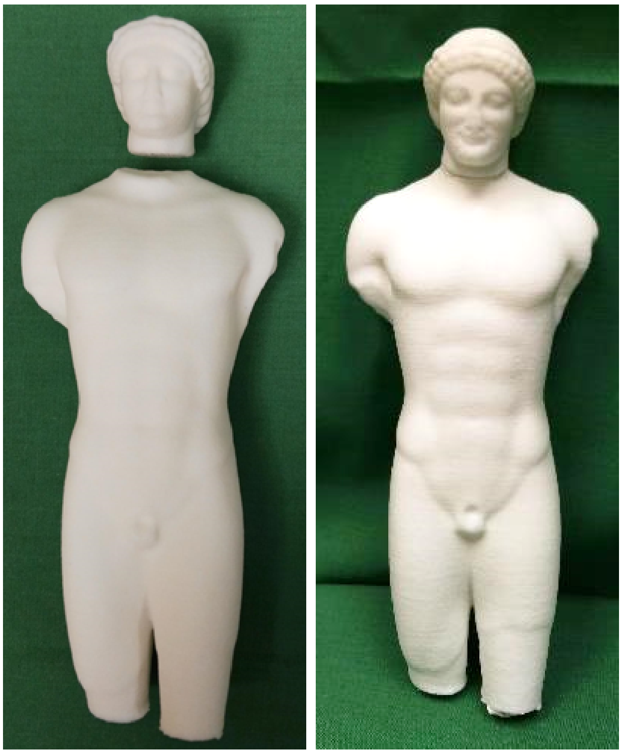


Figure A.42: (a) Phases of digital measuring with Meshlab, diameters of head and torso; (b) Phases of digital measuring with Meshlab, other diameters of head and torso.

a minimum and included the mechanical removal of the support structures and spot smoothing with a hot air rework tool. The physical model is not hollow, but fully solid in order to increase its weight for a more accurate and realistic final result.

A.4.5 Haptic technology

The choice to 3D print in scale and physically reassembly the statue would certainly be a good way for the curators of both the museums of Catania and Siracusa to showcase how this unique example of Greek sculpture looked like. Furthermore in the case of the archaeological museum of Siracusa, where there is already a tactile collection of artefacts ranging from Prehistory to the Greek period, the replica of the kouros of Leontinoi will represent another example of enhanced realization for the public with visual impairments. However, the process of 3D printing is still rather

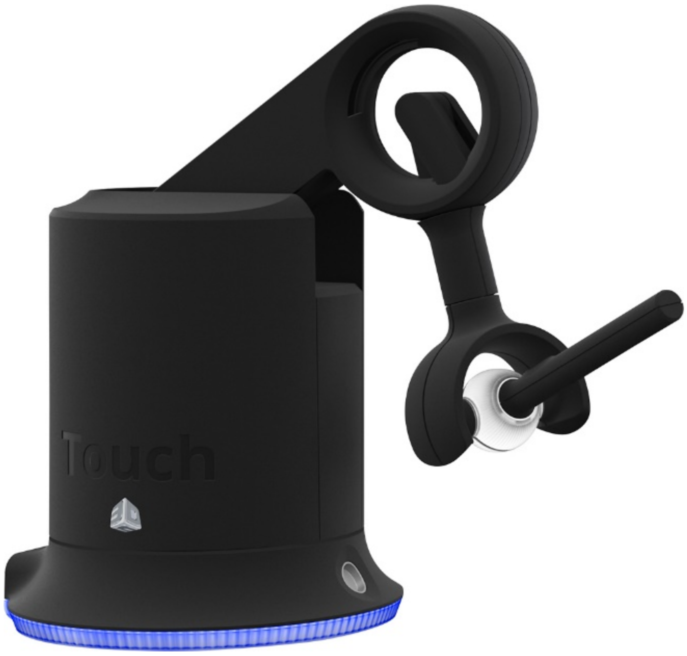


Figure A.43: 3D Systems Touch 3D Stylus.

time consuming and expensive, especially for models of medium to larger sizes and with other materials than simple polymers. In this respect, at this stage it cannot be the only solution to make archaeological objects immediately more accessible and to let visitors with or without cognitive deficits to learn from touching the subject of their interest.

In order to validate the sensorial experience of interacting with the 3D model of the reassembled statue and to compare it with the direct touch interaction with 3D print of the statue in 1 : 10 scale, an experimental test has been undertaken at the Center for Virtualization and Applied Spatial Technologies – CVAST of the University of South Florida. Using the haptic device 3D Systems Touch 3D Stylus (Fig. A.43) paired up with the proprietary software Geomagic Sculpt, a group of students were asked to interact with the digital model (Fig. A.44), and then to interact with the 3D print, and finally describe the feedback in a questionnaire.

The results achieved with a preliminary test employing a very limited sample of students clearly highlight the importance of any kind of touch interaction as a crucial step towards a more in-depth learning process. The other significant outcome is how the haptic device makes the interaction with the digital models more genuine

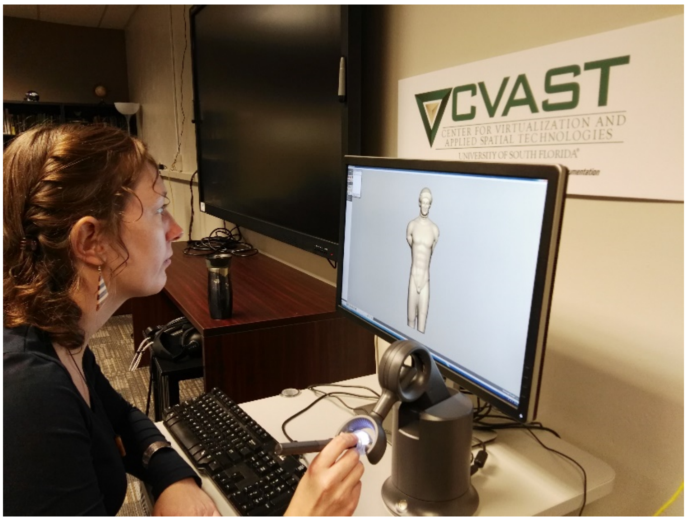

CVAST
CENTER FOR VIRTUALIZATION AND
APPLIED SPATIAL TECHNOLOGIES
UNIVERSITY OF SOUTH FLORIDA

Figure A.44: Touch interaction with the digital model of the statue through haptic technology.

and intense. Unfortunately, at this stage of the research it has not been possible to extend the experiment to a larger sample including students with visual impairments and cognitive disabilities, leaving room for a further step in future works.

Appendix B

Other Publications

In this Appendix it is reported a list of works published during my Ph.D. but not directly related to this thesis.

International Conferences:

- F. L. M. Milotta, D. Allegra, F. Stanco, G. M. Farinella, "An Electronic Travel Aid to Assist Blind and Visually Impaired People to Avoid Obstacles". Lecture Notes in Computer Science, 2015, Vol. 9257, pp. 604-615. DOI: 10.1007/978-3-319-23117-4_52.

- D. Allegra, F. Stanco, G. Valenti. "A Semi-automatic Algorithm for Applying the Ken Burns Effect". Smart Tools and Apps in computer Graphics, 2015. DOI: 10.2312/stag.20151296.

- G. Gallo, D. Allegra, Y. G. Atani, F. L. M. Milotta, F. Stanco, G. Catanuto. "Breast Shape Parametrization through Planar Projections". Lecture Notes in Computer Science, 2016, Vol. 10016, pp. 135-146. DOI: 10.1007/978-3-319-48680-2_13.

- D. Allegra, F. L. M. Milotta, D. Sinitò, F. Stanco, G. Gallo, Wafa Taher, G. Catanuto. "Description of Breast Morphology through Bag of Normals Representation". Lecture Notes in Computer Science, 2017, Vol. 10485, pp. 511-521. DOI: 10.1007/978-3-319-68548-9_47

Bibliography

[1] *Diet, Nutrition and the Prevention of Chronic Diseases*. Tech. rep. WHO Techinical Report Series - 916. Report of a Joint WHO/FAO Expert Consultation, Jan. 2002, p. 160. URL: http://apps.who.int/iris/bitstream/10665/42665/1/WHO_TRS_916.pdf?ua=1.

[2] R. A. Hammond and R. Levine. "The economic impact of obesity in the United States". In: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* 3 (Aug. 2010), 285–295. DOI: 10.2147/DMSOTT.S738.

[3] N. Suthumchai, S. Thongsukh, P. Yusuksataporn, and S. Tangsripairoj. "FoodForCare: An Android application for self-care with healthy food". In: *International Student Project Conference (ICT-ISPC)*. May 2016, pp. 89–92. DOI: 10.1109/ICT-ISPC.2016.7519243.

[4] W. D. Killgore and D. A. Yurgelun-Todd. "Body mass predicts orbitofrontal activity during visual presentations of high-calorie foods". In: *Neuroreport* 16 (8 May 2005), pp. 859–863. DOI: 10.1097/00001756-200505310-00016.

[5] M. Rosenbaum, M. Sy, K. Pavlovich, R. L. Leibel, and J. Hirsch. "Leptin reverses weight loss–induced changes in regional neural activity responses to visual food stimuli". In: *The Journal of Clinical Investigation* 118 (7 July 2008), pp. 2583–2591. DOI: 10.1172/JCI35055.

[6] N. Medic, H. Ziauddeen, S. E. Forwood, K. M. Davies, A. L. Ahern, S. A. Jebb, T. M. Marteau, and P. C. Fletcher. "The Presence of Real Food Usurps Hypothetical Health Value Judgment in Overweight People". In: *eNeuro* 3 (2 Apr. 2016), pp. 0025–16. DOI: 10.1523/ENEURO.0025-16.2016.

[7] J. F. Delwiche. "You eat with your eyes first". In: *Physiology & Behavior* 107 (4 Nov. 2012), pp. 502–504. DOI: 10.1016/j.physbeh.2012.07.007.

[8] K. McCrickerd and C. G. Forde. “Sensory influences on food intake control: moving beyond palatability”. In: *Obesity Reviews* 17 (1 Dec. 2016), pp. 18–29. DOI: 10.1111/obr.12340.

[9] O. Petit, A. D. Cheok, and O. Oullier. “Can Food Porn Make Us Slim? How Brains of Consumers React to Food in Digital Environments”. In: *Integrative Food, Nutrition and Metabolism* 3 (1 Jan. 2016), pp. 251–255. DOI: 10.15761/IFNM.1000138.

[10] E. A. Parrish and K. A. Goksel. “Pictorial pattern recognition applied to fruit harvesting”. In: *Transactions of the American Society of Agricultural and Biological Engineers* 20 (5 1977), pp. 822–827. DOI: 10.13031/2013.35657.

[11] P. Levi, A. Falla, and R. Pappalardo. “Image controlled robotics applied to citrus fruit harvesting”. In: *International Conference on Robot Vision and Sensory Controls*. Jan. 1988, pp. 2–4.

[12] D. C. Slaughter and R. C. Harrell. “Color Vision in Robotic Fruit Harvesting”. In: *Transactions of the American Society of Agricultural and Biological Engineers* 30 (4 1987), pp. 1144–1148. DOI: 10.13031/2013.30534.

[13] D. C. Slaughter and R. C. Harrell. “Discriminating fruit for robotic harvest using color in natural outdoor scenes”. In: *Transactions of the American Society of Agricultural and Biological Engineers* 32 (2 1989), pp. 757–763. DOI: 10.13031/2013.31066.

[14] M. Cardenas-Weber, A. Hetzroni, and G. E. Miles. “Machine Vision to Locate Melons and Guide Robotic Harvesting”. In: *Paper - American Society of Agricultural Engineers*. 1991, p. 21.

[15] F. Buemi, M. Massa, and G. Sandini. “Agrobot: a robotic system for greenhouse operations”. In: *Workshop on Robotics in Agriculture and the Food Industry*. Oct. 1995, pp. 172–184.

[16] A. R. Jiménez, A. K. Jain, R. Ceres, and J. L. Pons. “Automatic fruit recognition: a survey and new results using Range/Attenuation images”. In: *Pattern Recognition* 32 (10 Oct. 1999), pp. 1719–1736. DOI: 10.1016/S0031-3203(98)00170-8.

[17] P. Munkevik, T. Duckett, and G. Hall. “Vision System Learning for Ready Meal Characterisation”. In: *International Conference on Engineering and Food*. Mar. 2004.

[18] T. Kohonen. “The self-organizing map”. In: *Neurocomputing* 21 (1-3 Nov. 1998), pp. 1–6. DOI: 10.1016/S0925-2312(98)00030-7.

[19] P. Munkevik, T. Duckett, and G. Hall. “A computer vision system for appearance-based descriptive sensory evaluation of meals”. In: *Journal of Food Engineering* 78 (1 Jan. 2007), pp. 246–256. DOI: 10.1016/j.jfoodeng.2005.09.033.

[20] K. Kiliç, I. H. Boyaci, H. Köksel, and I. Küsmenoglu. “A classification system for beans using computer vision system and artificial neural networks”. In: *Journal of Food Engineering* 78 (3 Feb. 2007), pp. 897–904. DOI: 10.1016/j.jfoodeng.2005.11.030.

[21] D. W. Sun. “Inspecting pizza topping percentage and distribution by a computer vision method”. In: *Journal of Food Engineering* 44 (4 June 2000), pp. 245–249. DOI: 10.1016/S0260-8774(00)00024-8.

[22] C. J. Du and D. W. Sun. “Multi-classification of pizza using computer vision and support vector machine”. In: *Journal of Food Engineering* 86 (2 May 2008), pp. 232–242. DOI: 10.1016/j.jfoodeng.2007.10.001.

[23] S. Gunasekaran. “Computer vision technology for food quality assurance”. In: *Trends in Food Science & Technology* 7 (8 Aug. 1996), pp. 245–256. DOI: 10.1016/0924-2244(96)10028-5.

[24] T. Brosnan and D. W. Sun. “Improving quality inspection of food products by computer vision - A review”. In: *Journal of Food Engineering* 61 (1 Jan. 2004), pp. 3–16. DOI: 10.1016/S0260-8774(03)00183-3.

[25] C. J. Du and D. W. Sun. “Learning techniques used in computer vision for food quality evaluation: a review”. In: *Journal of Food Engineering* 72 (1 Jan. 2006), pp. 39–55. DOI: 10.1016/j.jfoodeng.2004.11.017.

[26] A. J. Rich. “A programmable calculator system for the estimation of nutritional intake of hospital patients”. In: *The American Journal of Clinical Nutrition* 34 (10 Oct. 1981), pp. 2276–2279.

[27] P. D. Wright, G. Shearing, A. J. Rich, and I. Johnston. “The Role of a Computer in the Management of Clinical Parenteral Nutrition”. In: *Journal of Parenteral and Enteral Nutrition* 2 (5 Nov. 1978), pp. 652–657. DOI: 10.1177/014860717800200506.

[28] K. Kitamura, T. Yamasaki, and K. Aizawa. “Food log by analyzing food images”. In: *International Conference on Multimedia*. Oct. 2008, pp. 999–1000. DOI: 10.1145/1459359.1459548.

[29] K. Aizawa, G. C. De Silva, M. Ogawa, and Y. Sato. “Food Log by snapping and processing images”. In: *Virtual Systems and Multimedia*. Oct. 2010, pp. 71–74. DOI: 10.1109/VSMM.2010.5665963.

[30] K. Kitamura, T. Yamasaki, and K. Aizawa. “FoodLog: Capture, Analysis and Retrieval of Personal Food Images via Web”. In: *Workshop on Multimedia for cooking and eating activities*. Oct. 2009, pp. 23–30. DOI: 10.1145/1630995.1631001.

[31] K. Kitamura, C. De Silva, T. Yamasaki, and K. Aizawa. “Image processing based approach to food balance analysis for personal food logging”. In: *International Conference on Multimedia and Expo*. July 2010, pp. 625–630. DOI: 10.1109/ICME.2010.5583021.

[32] Y. Maruyama, G. C. De Silva, T. Yamasaki, and K. Aizawa. “Personalization of food image analysis”. In: *Virtual Systems and Multimedia*. Oct. 2010, pp. 75–78. DOI: 10.1109/VSMM.2010.5665964.

[33] D. G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision* 60 (2 Nov. 2004), pp. 91–110. DOI: 10.1023/B:VISI.0000029664.99615.94.

[34] G. Shroff, A. Smailagic, and D. P. Siewiorek. “Wearable context-aware food recognition for calorie monitoring”. In: *International Symposium on Wearable Computers*. Oct. 2008, pp. 119–120. DOI: 10.1109/ISWC.2008.4911602.

[35] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney. “Recognition and volume estimation of food intake using a mobile device”. In: *Workshop on Applications of Computer Vision*. Dec. 2009. DOI: 10.1109/WACV.2009.5403087.

[36] M. a. Fischler and R. C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24 (6 June 1981), pp. 381–395. DOI: 10.1145/358669.358692.

[37] J. Dehais, S. Shevchik, P. Diem, and S. G. Mougiakakou. "Food Volume Computation for Self Dietary Assessment Applications". In: *International Conference on Bioinformatics and Bioengineering*. Nov. 2013. DOI: 10.1109/BIBE.2013.6701615.

[38] N. Chen, Y. Y. Lee, M. Rabb, and B. Schatz. "A computer vision system for appearance-based descriptive sensory evaluation of meals". In: *AMIA Annual Symposium*. Nov. 2010, pp. 106–110.

[39] J. Matas, O. Chum, M. Urban, and T. Pajdla. "Robust wide-baseline stereo from maximally stable extremal regions". In: *Image and Vision Computing* 22 (10 Sept. 2004), pp. 761–767. DOI: 10.1016/j.imavis.2004.02.006.

[40] H. Bay, T. Tuytelaars, and L. Van Gool. "SURF: Speeded Up Robust Features". In: *Lecture Notes in Computer Science*. Vol. 3951. May 2006, pp. 404–417. DOI: 10.1007/11744023_32.

[41] M. Agrawal, K. Konolige, and M. R. Blas. "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching". In: *Lecture Notes in Computer Science*. Vol. 5305. May 2008, pp. 102–115. DOI: 10.1007/978-3-540-88693-8_8.

[42] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN: 0-201-12227-8.

[43] 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". In: *International Conference on Wearable and Implantable Body Sensor Networks*. May 2012. DOI: 10.1109/BSN.2012.11.

[44] G. J. Burghouts and J.-M. Geusebroek. "Performance evaluation of local colour invariants". In: *Computer Vision and Image Understanding* 113 (1 Jan. 2009), pp. 48–62. DOI: 10.1016/j.cviu.2008.07.003.

[45] S. Battiato, G. M. Farinella, G. Gallo, and D. Ravì. “Exploiting Textons Distributions on Spatial Hierarchy for Scene Classification”. In: *Journal on Image and Video Processing* 2010 (1 Apr. 2010), p. 919367. DOI: 10.1155/2010/919367.

[46] S. Lazebnik, C. Schmid, and J. Ponce. “A sparse texture representation using local affine regions”. In: *Transactions on Pattern Analysis and Machine Intelligence* 27 (8 Aug. 2005), pp. 1265–1278. DOI: 10.1109/TPAMI.2005.151.

[47] A. Bhattacharyya. “On a Measure of Divergence between Two Multinomial Populations”. In: *Sankhyā: The Indian Journal of Statistics* 7 (4 July 1946), pp. 401–406.

[48] K. Yanai and T. Joutou. “SURF: Speeded Up Robust Features”. In: *International Conference on Image Processing*. Nov. 2009, pp. 285–288. DOI: 10.1109/ICIP.2009.5413400.

[49] S. Marčelja. “Mathematical description of the responses of simple cortical cells”. In: *Journal of the Optical Society of America* 70 (11 Nov. 1980), pp. 1297–1300. DOI: 10.1364/JOSA.70.001297.

[50] M. Varma and D. Ray. “Learning The Discriminative Power-Invariance Trade-Off”. In: *International Conference on Computer Vision*. Oct. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4408875.

[51] H. Hoashi, T. Joutou, and K. Yanai. “Image Recognition of 85 Food Categories by Feature Fusion”. In: *International Symposium on Multimedia*. Dec. 2010, pp. 296–301. DOI: 10.1109/ISM.2010.51.

[52] N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. In: *Computer Vision and Pattern Recognition*. June 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[53] Y. Matsuda, H. Hoashi, and K. Yanai. “Recognition of Multiple-Food Images by Detecting Candidate Regions”. In: *International Conference on Multimedia and Expo*. July 2012, pp. 25–30. DOI: 10.1109/ICME.2012.157.

[54] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan. “Object Detection with Discriminatively Trained Part-Based Models”. In: *Transactions on Pattern Analysis and Machine Intelligence* 32 (9 Sept. 2010), pp. 1627–1645. DOI: 10.1109/TPAMI.2009.167.

[55] Y. Deng and B. S. Manjunath. “Unsupervised segmentation of color-texture regions in images and video”. In: *Transactions on Pattern Analysis and Machine Intelligence* 23 (8 Aug. 2001), pp. 800–810. DOI: 10.1109/34.946985.

[56] A. E. Abdel-Hakim and A. A. Farag. “CSIFT: A SIFT Descriptor with Color Invariant Characteristics”. In: *Computer Vision and Pattern Recognition*. Vol. 2. June 2006, pp. 1978–1983. DOI: 10.1109/CVPR.2006.95.

[57] S. Lazebnik, C. Schmid, and J. Ponce. “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories”. In: *Computer Vision and Pattern Recognition*. Vol. 2. June 2006, pp. 2169–2178. DOI: 10.1109/CVPR.2006.68.

[58] Y. Matsuda, H. Hoashi, and K. Yanai. “Multiple-food recognition considering co-occurrence employing manifold ranking”. In: *International Conference on Pattern Recognition*. Nov. 2012, pp. 2017–2020.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Neural Information Processing Systems*. Dec. 2012, pp. 1097–1105.

[60] Y. Kawano and K. Yanai. “Food image recognition with deep convolutional features”. In: *International Joint Conference on Pervasive and Ubiquitous Computing*. Sept. 2014, pp. 589–593. DOI: 10.1145/2638728.2641339.

[61] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. “Image Classification with the Fisher Vector: Theory and Practice”. In: *International Journal of Computer Vision* 105 (3 Dec. 2013), pp. 222–245. DOI: 10.1007/s11263-013-0636-x.

[62] D. Ravì, B. Lo, and G.-Z. Yang. “Real-time Food Intake Classification and Energy Expenditure Estimation on a Mobile Device”. In: *International Conference on Wearable and Implantable Body Sensor Networks*. June 2015. DOI: 10.1109/BSN.2015.7299410.

[63] F. Perronnin and C. Dance. “Fisher Kernels on Visual Vocabularies for Image Categorization”. In: *Computer Vision and Pattern Recognition*. June 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383266.

[64] F. Perronnin, J. Sánchez, and T. Mensink. “Improving the Fisher Kernel for Large-Scale Image Classification”. In: *Lecture Notes in Computer Science*. Vol. 6314. 2010, pp. 143–156. DOI: 10.1007/978-3-642-15561-1_11.

[65] Y. Kawano and K. Yanai. “Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation”. In: *Lecture Notes in Computer Science*. Vol. 8927. Mar. 2014, pp. 3–17. DOI: 10.1007/978-3-319-16199-0_1.

[66] K. Yanai and Y. Kawano. “Food image recognition using deep convolutional network with pre-training and fine-tuning”. In: *International Conference on Multimedia & Expo Workshops*. June 2015, pp. 1–6. DOI: 10.1109/ICMEW.2015.7169816.

[67] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. “PFID: Pittsburgh fast-food image dataset”. In: *International Conference on Image Processing*. Nov. 2009, pp. 289–292. DOI: 10.1109/ICIP.2009.5413511.

[68] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. “Food Recognition Using Statistics of Pairwise Local Features”. In: *Computer Vision and Pattern Recognition*. June 2010, pp. 2249–2256. DOI: 10.1109/CVPR.2010.5539907.

[69] J. Shotton, M. Johnson, and R. Cipolla. “Semantic Texton Forests for Image Categorization and Segmentation”. In: Advances in Computer Vision and Pattern Recognition. Springer London, 2013, pp. 211–227. ISBN: 978-1-4471-4929-3. DOI: 10.1007/978-1-4471-4929-3_15.

[70] W. Wu and J. Yang. “Food Recognition Using Statistics of Pairwise Local Features”. In: *International Conference on Multimedia and Expo*. June 2009, pp. 1210–1213. DOI: 10.1109/ICME.2009.5202718.

[71] Z. Zong, D. T. Nguyen, P. Ogunbona, and W. Li. “On the Combination of Local Texture and Global Structure for Food Classification”. In: *International Symposium on Multimedia*. Dec. 2010, pp. 204–211. DOI: 10.1109/ISM.2010.37.

[72] T. Ahonen, A. Hadid, and M. Pietikäinen. “Face Description with Local Binary Patterns: Application to Face Recognition”. In: *Transactions on Pattern Analysis and Machine Intelligence* 28 (12 Oct. 2006), pp. 2037–2041. DOI: 10.1109/TPAMI.2006.244.

[73] S. Belongie, J. Malik, and J. Puzicha. “Shape matching and object recognition using shape contexts”. In: *Transactions on Pattern Analysis and Machine Intelligence* 24 (4 Aug. 2002), pp. 509–522. DOI: 10.1109/34.993558.

[74] D. T. Nguyen, Z. Zong, P. Ogunbona, and W. Li. “Object detection using Non-Redundant Local Binary Patterns”. In: *International Conference on Image Processing*. Sept. 2010, pp. 4609–4612. DOI: 10.1109/ICIP.2010.5651633.

[75] G. M. Farinella, M. Moltisanti, and S. Battiato. “Classifying Food Images Represented as Bag of Textons”. In: *International Conference on Image Processing*. Oct. 2014, pp. 5212–5216. DOI: 10.1109/ICIP.2014.7026055.

[76] G. M. Farinella, M. Moltisanti, and S. Battiato. “Food Recognition Using Consensus Vocabularies”. In: *Lecture Notes in Computer Science*. Vol. 9281. Aug. 2015, pp. 384–392. DOI: 10.1007/978-3-319-23222-5_47.

[77] A. Topchy, A. K. Jain, and W. Punch. “Clustering ensembles: models of consensus and weak partitions”. In: *Transactions on Pattern Analysis and Machine Intelligence* 27 (12 Dec. 2005), pp. 1866–1881. DOI: 10.1109/TPAMI.2005.237.

[78] V. Bettadapura, E. Thomaz, A. Parnami, G. D. Abowd, and I. Essa. “Leveraging Context to Support Automated Food Recognition in Restaurants”. In: *Winter Conference on Applications of Computer Vision*. 2015, pp. 580–587. DOI: 10.1109/WACV.2015.83.

[79] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp. “Combining global and local features for food identification in dietary assessment”. In: *International Conference on Image Processing*. Sept. 2011, pp. 1789–1792. DOI: 10.1109/ICIP.2011.6115809.

[80] E. Tola, V. Lepetit, and P. Fua. “DAISY: An efficient dense descriptor applied to wide-baseline stereo”. In: *Transactions on Pattern Analysis and Machine Intelligence* 32 (5 Apr. 2009), pp. 815–830. DOI: 10.1109/TPAMI.2009.77.

[81] M. Bosch, T. Schap, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp. "Integrated database system for mobile dietary assessment and analysis". In: *International Conference on Multimedia and Expo*. July 2011, pp. 1–6. DOI: 10.1109/ICME.2011.6012202.

[82] M. H. Rahmana, M. R. Pickering, D. Kerr, C. J. Boushey, and E. J. Delp. "A New Texture Feature for Improved Food Recognition Accuracy in a Mobile Phone Based Dietary Assessment System". In: *International Conference on Multimedia and Expo Workshops*. July 2012, pp. 418–423. DOI: 10.1109/ICMEW.2012.79.

[83] Y. Kawano and K. Yanai. "Real-Time Mobile Food Recognition System". In: *Computer Vision and Pattern Recognition Workshops*. June 2013, pp. 1–7. DOI: 10.1109/CVPRW.2013.5.

[84] A. Vedaldi and A. Zisserman. "Efficient Additive Kernels via Explicit Feature Maps". In: *Transactions on Pattern Analysis and Machine Intelligence* 34 (3 Mar. 2012), pp. 480–492. DOI: 10.1109/TPAMI.2011.153.

[85] M.-Y. Chen, Y.-H. Yang, C.-J. Ho, S.-H. Wang, S.-M. Liu, E. Chang, C.-H. Yeh, and M. Ouhyoung. "Automatic Chinese food identification and quantity estimation". In: *SIGGRAPH Asia 2012 Technical Briefs*. Nov. 2012, pp. 1–4. DOI: 10.1145/2407746.2407775.

[86] C. Pham, D. Jackson, J. Schöning, T. Bartindale, T. Plotz, and P. Olivier. "FoodBoard: Surface Contact Imaging for Food Recognition". In: *International Joint Conference on Pervasive and Ubiquitous Computing*. Sept. 2013, pp. 749–752. DOI: 10.1145/2493432.2493522.

[87] T. K. Ho. "Random decision forests". In: *International Conference on Document Analysis and Recognition*. Vol. 1. Aug. 1995, pp. 278–282. DOI: 10.1109/ICDAR.1995.598994.

[88] L. Bossard, M. Guillaumin, and L. Van Gool. "Food-101 – Mining Discriminative Components with Random Forests". In: *Lecture Notes in Computer Science*. Vol. 8694. Sept. 2014, pp. 446–461. DOI: 10.1007/978-3-319-10599-4_29.

[89] W. Xin, D. Kumar, N. Thome, M. Cord, and F. Precioso. "Recipe recognition with large multimodal food dataset". In: *International Conference on Multimedia Expo Workshops*. July 2015, pp. 1–6. DOI: 10.1109/ICMEW.2015.7169757.

[90] L. Herranz, X. Ruihan, and J. Shuqiang. "A probabilistic model for food image recognition in restaurants". In: *International Conference on Multimedia and Expo*. June 2015, pp. 1–6. DOI: 10.1109/ICME.2015.7177464.

[91] X. Ruihan, L. Herranz, J. Shuqiang, W. Shuang, S. Xinhang, and R. Jain. "Geolocalized Modeling for Dish Recognition". In: *Transactions on Multimedia* 17 (8 Aug. 2015), pp. 1187–1199. DOI: 10.1109/TMM.2015.2438717.

[92] G. M. Farinella, D. Allegra, and F. Stanco. "A Benchmark Dataset to Study the Representation of Food Images". In: *Lecture Notes in Computer Science*. Vol. 8927. Mar. 2015, pp. 584–599. DOI: 10.1007/978-3-319-16199-0_41.

[93] S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravì. "Aligning codebooks for near duplicate image detection". In: *Multimedia Tools and Applications* 72 (2 Aug. 2014), pp. 1483–1506. DOI: 10.1007/s11042-013-1470-4.

[94] Y. Hu, X. Cheng, L.-T. Chia, X. Xie, D. Rajan, and A.-H. Tan. "Coherent Phrase Model for Efficient Image Near-Duplicate Retrieval". In: *Transactions on Multimedia* 11 (8 Dec. 2009), pp. 1434–1445. DOI: 10.1109/TMM.2009.2032676.

[95] M. Varma and A. Zisserman. "A Statistical Approach to Texture Classification from Single Images". In: *International Journal of Computer Vision* 62 (1-2 Apr. 2005), pp. 61–81. DOI: 10.1023/B:VISI.0000046589.39864.ee.

[96] X. Qi, R. Xiao, C.-G. Li, Y. Qiao, J. Guo, and X. Tang. "Pairwise Rotation Invariant Co-occurrence Local Binary Pattern". In: *Transactions on Pattern Analysis and Machine Intelligence* 36 (11 Nov. 2014), pp. 2199–2213. DOI: 10.1109/TPAMI.2014.2316826.

[97] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp. "Analysis of food images: Features and classification". In: *International Conference on Image Processing*. Oct. 2014, pp. 2744–2748. DOI: 10.1109/ICIP.2014.7025555.

[98] D. Nistér and H. Stewénius. “Scalable Recognition with a Vocabulary Tree”. In: *Computer Vision and Pattern Recognition*. Vol. 2. June 2006, pp. 2161–2168. DOI: 10.1109/CVPR.2006.264.

[99] F. Foroni, G. Pergola, G. Argiris, and R. I. Rumiati. “The FoodCast research image database (FRIDa)”. In: *Frontiers in Human Neuroscience* 7 (Mar. 2013). DOI: 10.3389/fnhum.2013.00051.

[100] P. Pouladzadeh, A. Yassine, and S. Shirmohammadi. “FooDD: Food Detection Dataset for Calorie Measurement Using Food Images”. In: *Lecture Notes in Computer Science*. Vol. 9281. Aug. 2015, pp. 441–448. DOI: 10.1007/978-3-319-23222-5_54.

[101] H. Kagaya, K. Aizawa, and M. Ogawa. “Food Detection and Recognition Using Convolutional Neural Network”. In: *International Conference on Multimedia*. Nov. 2014, pp. 1085–1088. DOI: 10.1145/2647868.2654970.

[102] S. S. Khan and M. G. Madden. “A Survey of Recent Trends in One Class Classification”. In: *Lecture Notes in Computer Science*. Vol. 6206. Aug. 2010, pp. 188–197. DOI: 10.1007/978-3-642-17080-5_21.

[103] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. “Estimating the Support of a High-Dimensional Distribution”. In: *Neural Computation* 13 (7 July 2001), pp. 1443–1471. DOI: 10.1162/089976601750264965 7.

[104] D. G. Lowe. “Object Recognition from Local Scale-Invariant Features”. In: *International Conference on Computer Vision*. Vol. 2. Sept. 1999, pp. 1150–1157. DOI: 10.1109/ICCV.1999.790410.

[105] M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou. “A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model”. In: *Journal of Biomedical and Health Informatics* 18 (4 July 2014), pp. 1261–1271. DOI: 10.1109/JBHI.2014.2308928.

[106] E. Nowak, F. Jurie, and B. Triggs. “Sampling Strategies for Bag-of-Features Image Classification”. In: *Lecture Notes in Computer Science*. Vol. 3954. May 2006, pp. 490–503. DOI: 10.1007/11744085_38.

[107] A. Vedaldi and B. Fulkerson. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. http://www.vlfeat.org/. 2008.

[108] T. Ojala, M. Pietikäinen, and T. Mäenpää. "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns". In: *Pattern Analysis and Machine Intelligence* 24 (7 July 2002), pp. 971–987. DOI: 10.1109/TPAMI.2002.1017623.

[109] B. Julesz. "Textons, the elements of texture perception, and their interactions". In: *Nature* 290 (Mar. 1981), pp. 91–97. DOI: 10.1038/290091a0.

[110] T. Leung and J. Malik. "Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons". In: *International Journal of Computer Vision* 43 (1 Feb. 2001), pp. 29–44. DOI: 10.1023/A:1011126920638.

[111] D. Williams and B. Julesz. "Filters Versus Textons in Human and Machine Texture Discrimination". In: *Neural Networks for Perception - Human and Machine Perception*. Ed. by H. Wechsler. Academic Press, 1992, pp. 145–175. ISBN: 978-0-12-741251-1. DOI: 10.1016/B978-0-12-741251-1.50014-X.

[112] D. Williams and B. Julesz. "Perceptual asymmetry in texture perception". In: *Proceedings of the National Academy of Sciences of the United States of America* 89 (14 1992), pp. 6531–6534.

[113] G. J. Van Tonder and Y. Ejima. "From image segregation to anti-textons". In: *Perception* 29 (10 Oct. 2000), pp. 1231–1247. DOI: 10.1068/p2931.

[114] C. Schmid. "Constructing models for content-based image retrieval". In: *Computer Vision and Pattern Recognition*. Vol. 2. Dec. 2001, pp. 39–45. DOI: 10.1109/CVPR.2001.990922.

[115] G. J. Burghouts and J.-M. Geusebroek. "Material-specific adaptation of color invariant features". In: *Pattern Recognition Letters* 30 (3 Feb. 2009), pp. 306–313. DOI: 10.1016/j.patrec.2008.10.005.

[116] C. R. Maurer, Q. Rensheng, and R. Vijay. "A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions". In: *Pattern Analysis and Machine Intelligence* 25 (2 Feb. 2003), pp. 265–270. DOI: 10.1109/TPAMI.2003.1177156.

[117] A. Rosenfeld and J. L. Pfaltz. “Distance functions on digital pictures”. In: *Pattern Recognition* 1 (1 July 1968), pp. 33–61. DOI: 10.1016/0031-3203(68)90013-7.

[118] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. “Object retrieval with large vocabularies and fast spatial matching”. In: *Computer Vision and Pattern Recognition*. June 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383172.

[119] N. Martinel, C. Piciarelli, C. Micheloni, and G. L. Foresti. “On Filter Banks of Texture Features for Mobile Food Classification”. In: *International Conference on Distributed Smart Cameras*. Sept. 2015, pp. 14–19. DOI: 10.1145/2789116.2789132.

[120] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar. “Menu-Match: Restaurant-Specific Food Logging from Images”. In: *Winter Conference on Applications of Computer Vision*. Jan. 2015, pp. 844–851. DOI: 10.1109/WACV.2015.117.

[121] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going Deeper with Convolutions”. In: *Computer Vision and Pattern Recognition*. June 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[122] M. Merler, H. Wu, R. Uceda-Sosa, Q.-B. Nguyen, and J. R. Smith. “Snap, Eat, RepEat: A Food Recognition Engine for Dietary Logging”. In: *International Workshop on Multimedia Assisted Dietary Management*. Oct. 2016, pp. 31–40. DOI: 10.1145/2986035.2986036.

[123] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy. “Im2Calories: Towards an Automated Mobile Vision Food Diary”. In: *International Conference on Computer Vision*. Dec. 2015, pp. 1233–1241. DOI: 10.1109/ICCV.2015.146.

[124] M. Anthimopoulos, J. Dehais, S. Shevchik, B. H. Ransford, D. Duke, P. Diem, and S. Mougiakakou. “Computer Vision-Based Carbohydrate Estimation for Type 1 Patients With Diabetes Using Smartphones”. In: *Journal of Diabetes Science and Technology* 9 (3 Apr. 2015), pp. 507–515. DOI: 10.1177/1932296815580159.

[125] T. Miyazaki, G. C. de Silva, and K. Aizawa. "Image-based Calorie Content Estimation for Dietary Assessment". In: *International Symposium on Multimedia*. Dec. 2011, pp. 363–368. DOI: 10.1109/ISM.2011.66.

[126] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp. "The Use of Mobile Devices in Aiding Dietary Assessment and Evaluation". In: *European Journal of Clinical Nutrition* 4 (4 Aug. 2010), pp. 756–766. DOI: 10.1109/JSTSP.2010.2051471.

[127] L. Bally, J. Dehais, C. T. Nakas, M. Anthimopoulos, M. Laimer, D. Rhyner, G. Rosenberg, T. Zueger, P. Diem, S. Mougiakakou, and C. Stettler. "Carbohydrate Estimation Supported by the GoCARB System in Individuals With Type 1 Diabetes: A Randomized Prospective Pilot Study". In: *Diabetes Care* 40 (2 Apr. 2017), e6–e7. DOI: 10.2337/dc16-2173.

[128] D. Rhyner, H. Loher, J. Dehais, M. Anthimopoulos, S. Shevchik, H. R. Botwey, D. Duke, C. Stettler, P. Diem, and S. Mougiakakou. "Carbohydrate Estimation by a Mobile Phone-Based System Versus Self-Estimations of Individuals With Type 1 Diabetes Mellitus: A Comparative Study". In: *Journal of Medical Internet Research* 18 (5 May 2016). DOI: 10.2196/jmir.5567.

[129] F. Zhu, M. Bosch, T. Schap, N. Khanna, D. S. Ebert, C. J. Boushey, and E. J. Delp. "Segmentation Assisted Food Classification for Dietary Assessment". In: *Computational Imaging*. Feb. 2011. DOI: 10.1117/12.877036.

[130] V. Badrinarayanan, A. Kendall, and R. Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *Transactions on Pattern Analysis and Machine Intelligence* (2017).

[131] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science*. Vol. 9351. Nov. 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[132] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou. "Two-View 3D Reconstruction for Food Volume Estimation". In: *Transactions on Multimedia* 19 (5 May 2017), pp. 1090–1099. DOI: 10.1109/TMM.2016.2642792.

[133] D. Eigen, C. Puhrsch, and R. Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Neural Information Processing Systems*. Vol. 3. Jan. 2014, pp. 2366–2374.

[134] J. Leclant. “Le retour à la bibliothèque de l’Institut des fragments de deux papyri calcinés d’Herculanum déroulés par le Centre international des Papyri de Naples”. In: *Comptes rendus des séances de l’Académie des Inscriptions et Belles-Lettres* 146 (3 2002), pp. 841–843. DOI: 10.3406/crai.2002.22480.

[135] D. Delattre. “Le point sur les travaux relatifs au P. Herc.Paris.2”. In: *Comptes rendus des séances de l’Académie des Inscriptions et Belles-Lettres* 153 (2 2009), pp. 925–943. DOI: 10.3406/crai.2009.92558.

[136] B. W. Seales and L. Yun. “Digital Restoration using Volumetric Scanning”. In: *Joint ACM/IEEE Conference on Digital Libraries*. June 2004, pp. 117–124. DOI: 10.1109/JCDL.2004.240414.

[137] R. Baumann, D. Carr Porter, and B. W. Seales. “The Use of Micro-CT in the Study of Archaeological Artifacts”. In: *International Conference on NDT of Art*. 2008, pp. 1–9.

[138] B. W. Seales, J. Griffioen, R. Baumann, and M. Field. “Analysis of Herculaneum Papyri with X-Ray Computed Tomography”. In: *International Conference on non-destructive investigations and microanalysis for the diagnostics and conservation of cultural and environmental heritage*. 2011.

[139] O. Samko, Y.-K. Lai, D. Marshall, and P. L. Rosin. “Virtual unrolling and information recovery from scanned scrolled historical documents”. In: *Pattern Recognition* 47 (1 Jan. 2014), pp. 248–259. DOI: 10.1016/j.patcog.2013.06.015.

[140] V. Mocella, E. Brun, C. Ferrero, and D. Delattre. “Revealing letters in rolled Herculaneum papyri by X-ray phase-contrast imaging”. In: *Nature Communications* 6 (5895 Jan. 2015). DOI: 10.1038/ncomms6895.

[141] R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006. ISBN: 013168728X.

[142] P. G. Guzzo. “A Group of Hellenistic Silver Objects in the Metropolitan Museum”. In: *Metropolitan Museum Journal* 38 (2003), pp. 45–94.

[143] M. F. Alberghina, R. Barraco, M. Brai, T. Schillaci, and L. Tranchina. “Integrated analytical methodologies for the study of corrosion processes in archaeological bronzes”. In: *Spectrochimica Acta Part B: Atomic Spectroscopy* 66 (2 Feb. 2011), pp. 129–137. DOI: 10.1016/j.sab.2010.12.010.

[144] M. F. Alberghina, R. Barraco, M. Brai, L. Pellegrino, F. Prestileo, S. Schiavone, and L. Tranchina. “Gilding and pigments of Renaissance marble of Abatellis Palace: non-invasive investigation by XRF spectrometry”. In: *X-Ray Spectrometry* 42 (2 Mar. 2013), pp. 68–78. DOI: 10.1002/xrs.2435.

[145] R. J. H. Wanhill. “Case Histories of Ancient Silver Embrittlement”. In: *Journal of Failure Analysis and Prevention* 11 (3 June 2011), pp. 178–185. DOI: 10.1007/s11668-010-9429-5.

[146] *Smithsonian X 3D website:* http://3d.si.edu/. Last visit: September 2017.

[147] *Sketchfab website:* https://sketchfab.com/. Last visit: September 2017.

[148] Y. Liao, M. Lezoche, H. Panetto, and N. Boudjlida. “Semantic Annotation Model Definition for Systems Interoperability”. In: *Lecture Notes in Computer Science*. Vol. 7046. 2011, pp. 61–70. DOI: 10.1007/978-3-642-25126-9_14.

[149] *Structure Sensor Website:* http://structure.io/. Last visit: September 2017.

[150] L. Arcifa, D. Calì, A. Patanè, F. Stanco, D. Tanasi, and L. Truppia. “Laser scanning and 3D Modelling Techniques in Urban Archaeology: the Excavation of “St. Agata al Carcere” Church in Catania”. In: *Virtual Archaeology Review* 1 (2 May 2010), pp. 44–48.

[151] G. Gallo, F. Milanese, E. Sangregori, F. Stanco, D. Tanasi, and L. Truppia. “Coming back home. The virtual model of the Asclepius roman statue from the Museum of Syracuse (Italy)”. In: *Virtual Archaeology Review* 1 (2 May 2010), pp. 93–97.

[152] F. Stanco and D. Tanasi. “Beyond virtual replicas: 3D modeling and maltese prehistoric architecture”. In: *Journal of Electrical and Computer Engineering* (2013). DOI: 10.1155/2013/430905.

[153] F. Remondino. “Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning”. In: *Remote Sensing* 3 (6 May 2011), pp. 1104–1138. DOI: 10.3390/rs3061104.

[154] B. Curless. “From Range Scans to 3D Models”. In: *SIGGRAPH Computer Graphics* 33 (4 Nov. 1999), pp. 38–41. DOI: 10.1145/345370.345399.

[155] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys. “Turning Mobile Phones into 3D Scanners”. In: *Computer Vision and Pattern Recognition*. 2014, pp. 3946–3953. DOI: 10.1109/CVPR.2014.504.

[156] T. Schöps, T. Sattler, C. Häne, and M. Pollefeys. “3D Modeling on the Go: Interactive 3D Reconstruction of Large-Scale Scenes on Mobile Devices”. In: *3D Vision*. Oct. 2015, pp. 291–299. DOI: 10.1109/3DV.2015.40.

[157] E. Cappelletto, P. Zattunigh, and G. M. Cortelazzo. “3D Scanning of Cultural Heritage with Consumer Depth Cameras”. In: *Multimedia Tools and Applications* 75 (7 Apr. 2016), pp. 3631–3654. DOI: 10.1007/s11042-014-2065-4.

[158] I. Toschi, A. Capra, L. De Luca, J.-A. Beraldin, and L. Cournoyer. “On the evaluation of photogrammetric methods for dense 3D surface reconstruction in a metrological context”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. 2. May 2015, pp. 371–378. DOI: 10.5194/isprsannals-II-5-371-2014.

[159] *Skanect Website:* http://skanect.occipital.com/. Last visited September 2017.

[160] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. “Meshlab: an Open-Source Mesh Processing Tool”. In: *Eurographics Italian Chapter Conference*. 2008, pp. 129–136. DOI: 10.3390/rs8030178.

[161] M. Gaiani, F. Remondino, F. I. Apollonio, and A. Ballabeni. “An Advanced Pre-Processing Pipeline to Improve Automated Photogrammetric Reconstructions of Architectural Scenes”. In: *Remote Sensing* 8 (3 Feb. 2016). DOI: 10.3390/rs8030178.

[162] P. Cignoni, C. Rocchini, and R. Scopigno. "Metro: Measuring Error on Simplified Surfaces". In: *Computer Graphics Forum* 17 (2 June 1998), pp. 167–174. DOI: 10.1111/1467-8659.00236.

[163] M. Callieri, P. Cignoni, M. Dellepiane, and R. Scopigno. "Pushing Time-of-Flight Scanners to the Limit". In: *Virtual Reality, Archaeology and Intelligent Cultural Heritage*. 2009, pp. 85–92. DOI: 10.2312/VAST/VAST09/085-092.

[164] G. M. A. Richter. *Kouroi. Archaic Greek Youths. A Study of the Development of the Kouros Type in Greek Sculpture*. Phaidon Press, London, 1960.

[165] E. De Miro. *La scultura siceliota nell'età classica*. I Greci in Occidente, 1996.

[166] G. Libertini. *Il Museo Biscari*. Bestetti e Tumminelli, 1930.

[167] G. Libertini. *Il Castello Ursino e le raccolte artistiche comunali di Catania*. Zuccarello & Izzi, 1937.

[168] G. V. Gentili. *I due kouroi da Osimo e i tre kouroi del vecchio museo Archeologico di Siracusa nello studio e ricordo di Luigi Bernabò Brea*. M. Cavalier, M. Bernabò Brea.

[169] F. Niccolucci, M. Dellepiane, S. P. Serna, H. Rushmeier, and L. Van Gool. "A Blender Open Pipeline for a 3D Animated Historical Short Film". In: *Virtual Reality, Archaeology and Intelligent Cultural Heritage*. 2011, pp. 81–84. DOI: 10.2312/PE/VAST/VAST11S/081-084.