UNIVERSITÀ DEGLI STUDI DI CATANIA

Dipartimento di Matematica e Informatica

Dottorato di Ricerca Internazionale in Matematica e Informatica

XXX Ciclo

Filippo Luigi Maria Milotta

Multi-Device Media Analysis and Summarization for High Bandwidth Connected Environment

Tesi di Dottorato di Ricerca Internazionale

Tutor: Prof. Sebastiano Battiato

Anno Accademico 2016 - 2017

"I turisti diventano dunque involontari custodi di un patrimonio virtuale,
che non andrà mai più perduto,
dando così un senso anche ai nostri più inutili selfie."

"Hence, the tourists become unintended keepers of a virtual cultural heritage,
that will never be lost,
making meaningful even our most meaningless selfies."

SuperQuark, 21 June 2017.

Abstract

This dissertation collects all the research work done by the PhD candidate in the Joint Open Lab for Wireless Applications in multi-deVice Ecosystems (JOL WAVE CATANIA) of TIM Telecom Italia, which sponsored his doctoral fellowship. These applications, in which a big amount of multimedia data is analyzed and summarized, may be the enabling technology for LTE-based multimedial services. Three main categories of media have been treated in this dissertation: images, videos, and 3D data. For images and videos we realized two frameworks: The Social Picture and RECfusion. The Social Picture is a framework to collect and explore huge amount of crowd-sourced social images about public events, cultural sites and other customized private events. RECfusion is designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Through these two frameworks the following topics are discussed: Image Matching, Saliency Estimation, and Video and Scene Summarization. Particularly, we investigated advanced image matching techniques (e.g., Compact Descriptors for Visual Search - CDVS). We detailed Content Based Image Retrieval (CBIR) methods, and we described how to compute the heatmap, which represent a valuable tool for analysis and summarization of large images collections. Then, we reach a novel definition of saliency model that we named *Social Saliency*. This name has been chosen because the model of attention is obtained querying image databases of social media (e.g., Flickr, Instagram, Panoramio), that definitely represent a "social" environment. We describe how media collections can be employed for 3D reconstruction and social saliency estimation using images, scene and context tracking using videos, parametrization using 3D medical data, and preservation and restoration using 3D Cultural Heritage data. In all of these applications, we described analysis and summarization methods to be used in high bandwidth connected environments.

We present 4 real use-cases, co-authored and published in international journals and conferences: The Social Picture, RECfusion, 3D Data Analysis for Cultural Heritage, and 3D Data Analysis for Medical Research. Publications are listed in Section 1.2.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Prof. S. Battiato for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. His advice on both research as well as on my career have been priceless.

I would like to thank also professors of Image Processing Lab: F. Stanco, G.M. Farinella, and G. Gallo. Their professional suggestions in the research field of Multimedia, Computer Vision, and Pattern Recognition have been priceless for me in reaching the completion of this dissertation. I wish to say a special thank particularly to Prof. F. Stanco, my former advisor for bachelor and master degrees. While he will always be an irreplaceable mentor for me, I have also found in him a trusted friend.

A big thank goes to colleagues of the TIM Telecom Italia Joint Open Lab WAVE. I would like to say a special thank particularly to V. D'Amico, G. Torrisi and L. Addesso, who closely followed and guided my work within the JOL. I would like to express my gratitude to G. Torrisi for being my industrial tutor. His guidance helped me in understanding how a professional should act within a big enterprise like TIM Telecom Italia, which sponsored this doctoral fellowship.

My sincere thanks also goes to Prof. M. Decker, who invited me in the Center for Virtualization and Applied Spatial Technologies (CVAST) in the University of South Florida (USF), Florida, USA. He provided me an opportunity to join his team for 6 months as visiting scholar, and gave access to his laboratory and research facilities. Without his precious support, it would not be possible to conduct this doctoral fellowship as a *Doctorate International Fellowship.* The period I spent abroad was definitely a tremendous chance to increase my expertise. I take this opportunity to thank Prof. S. Sarkar, who generously gave me valuable advices and suggestions for improving this dissertation.

I would like to thank also Dr. G. Catanuto, who gave me a chance to conduct a very professional and interesting medical research in the field of breast parametrization. He gave me a clear evidence of how much research can change and improve the life of people.

Besides my advisors, I would like to thank the rest of my thesis committee: Prof. D. Tanasi, and Prof. S. Livatino, for their insightful comments and encouragement, but also for the hard questions which brought me to widen my research from various perspectives.

Contents

Dedicated to my family, who always believed in me,

to friends, who shared this journey with me,

and to God, who keep saying me
"do not be afraid"

Chapter 1

Introduction

In 2012 Telecom Italia, one of the major telco company in Italy, created the Joint Open Labs (JOLs). The JOLs can be seen within the scope of a project of valorization of the Open Innovation paradigm [1]. Indeed, the labs are set within campus of the most important Italian universities, from north to south of Italy. In the Open Innovation paradigm, the relationships and collaborations between companies and universities represent a core point of strength. When the Innovation process of a big company is opened to the university system and to local smaller enterprises, then many different actors are involved, from specialists to general audience. From this kind of contamination, the practical skills from the industrial and enterprise worlds are mixed with the theoretical ones from the academic world. In this way, new assets, products, services and ideas are developed employing the most advanced technologies, reducing time and cost of the research and development process (fast prototyping).

This dissertation collects all the research work done by the PhD candidate in the Joint Open Lab for Wireless Applications in multi-deVice Ecosystems (JOL WAVE) of TIM Telecom Italia, which is located within the University of Catania campus and sponsored this doctoral fellowship. In this lab, novel mobile applications based on versatile software are designed and implemented. The development process is performed employing highly connectable hardware and devices like smartphones, tablets, cameras, wearable devices, sensors, actuators, smart objects, and interactive screens. For each application, use scenarios and the most promising business models are defined and analyzed. The Quality of Service is another important issue, since high performance, low latency, ease of use, reliability, security, low response time (real-time), high integrity, and data-rate are advisable specifics of a good application.

All of the previous mentioned features are necessary in the context of high definition multimedia services. The ever-increasing spreading of mobile and wearable devices with embedded interconnected sensors allows the use of first-person vision (FPV) in order to take into account the point of view of the user himself (user-orient approach). The future networks will be optimized for handle high definition multimedia contents like real-time video based applications. In this scenario, the Long Term Evolution (LTE-4G) is a standard representing a valuable improvement in the mobile telecommunication [2]. LTE is the answer for the growing request of multimedial services allowing users to gain better performances.

With this dissertation we want to contribute to exponential growth of LTE-based multimedial services experienced in the last decade. Real use-cases are defined to prove that smart devices and ultra-broadband may be the enabling technology for LTE-based multimedial services. Three main categories of media are treated in this dissertation: images, videos, and 3D data. More in detail, for images and videos we realized two frameworks: The Social Picture and RECfusion, respectively.

In The Social Picture (TSP) an huge amount of crowdsourced social images can be collected and explored [3]. We distinguish three main kind of events: public, private, and cultural heritage related ones. The framework embeds a number of advanced Computer Vision algorithms, able to capture the visual content of images and organize them in a semantic way. We employed VisualSFM (VSFM) [4, 5] to compute a 3D sparse reconstruction of a collection within TSP [6]. VisualSFM creates a N-View Match (NVM) file as output. Starting from this NVM file, which characterizes the 3D sparse reconstruction, we are able to build two important relationships: the one between cameras and points, and the one between cameras themselves. Using these relationships, we implemented two advanced Image Analysis applications. In the first one, we considered the cameras as nodes in a fully connected graph in which the edges weights are equal to the number of matches between cameras. The spanning tree of this graph is used to explore images in a meaningful way, obtaining a Scene Summarization. In the second application, we defined three kinds of visual-features density maps: density map, weighted-density map and social-weighted-density map. We shown how the density-maps can be used together with the Structure from Motion (SfM) technique to highlight parts of the image with robust visual features. Among all the defined kind of density maps, we shown that the social-weighted-density map represent a good tool to stress the presence of visual features even when a strong occlusion is present in the image.

RECfusion is a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices [7, 8, 9]. The huge diffusion of mobile devices with embedded cameras has opened new challenges in the context of the automatic understanding of video streams acquired by multiple users during events, such as sport matches, expos, concerts. Among the other goals, there is the interpretation of which visual contents are the most relevant and popular (i.e., where users look). The popularity of a visual content is an important cue exploitable in several fields that include the estimation of the mood of the crowds attending to an event, the estimation of the interest of parts of a cultural heritage, etc. In live social events people capture and share videos which are related to the event. These data contain a certain amount of redundancy by regarding to the interesting scenes which have captured the attention of the crowd (e.g., fireworks during a folkloristic event). The popularity of a visual content can be obtained through the "visual consensus" among multiple video streams acquired by the different users devices. In this dissertation, we address the problem of detecting and summarizing the "popular scenes" captured by users with a mobile camera during events. For this purpose, we have developed RECfusion, in which the key popular scenes of multiple streams are identified over time. The framework is able to generate a video which captures the interests of the crowd starting from a set of videos by considering scene content popularity. The frames composing the final popular video are automatically selected from the different video streams by considering the scene recorded by the highest number of users' devices (i.e., the most popular scene).

Through these two frameworks, TSP and RECfusion, the following topics are discussed: Image Matching and Retrieval, Saliency Estimation, and Video and Scene Summarization. We show use-cases of multi-device images and videos analysis and summarization tasks. Then, focusing on the images, we treat in detail image matching employing advanced techniques (e.g., Compact Descriptors for Visual Search - CDVS [10, 11]). Finally, we trained a Convolutional Neural Network (CNN) in order to investigate the performances of a model trained to describe a novel kind of saliency that we named *Social Saliency*.

The Social Saliency is a saliency metric: it is used to measure how much the parts of an image are important (e.g., how much are viewed by the users, which acquired a higher amount of picture depicting these parts). In the 1985, Koch and Ullman [12] gave a first definition of model of attention and stated the concept of saliency map. More recent approaches to visual saliency have been reviewed through the last decades [13, 14]. We named our saliency metric as “social” because the model employed to learn and understand what are the salient parts of an image is based on a crowd-sourced consensus. The model of attention is obtained querying image databases of social media (e.g., Flickr, Instagram, Panoramio), that definitely represent a “social” environment.

The 3D data media are treated in the last part of the dissertation. We report in the appendices a benchmark of the 3D web viewers currently available. This kind of viewers is an enabling technology to allow online fruition of 3D contents for as many users and scholars as possible. We describe real use-cases in two different contexts: Cultural Heritage and Medical.

We address the issue of Cultural Heritage digitization through 3D scanning (Digital Archaeology [15]). This topic is important for preservation and conservation of Cultural Heritage sites. The destructive force of nature has demonstrated several times how an entire site can be annihilated in a short lapse of time causing irreparable damage, especially in those countries rich in archeology but poor in technical knowledge. Two shocking examples, which unfortunately did not have a great coverage on the media, are represented by the Iranian citadel of Arg-e Bam (3^{rd} BCE - 3^{rd} CE), a $200,000\ m^2$ complex made of sun-dried mud brick, wiped off the map by a magnitude 6.5 earthquake in 2003 [16] and by the complex of $1,400$ temples of the Shwedagon Pagoda (6^{th} - 10^{th} CE), in the Irrawaddy Delta region of Myanmar, which were razed to the ground by a cyclone in 2008 [17]. Notwithstanding, a natural disaster is not enough to raise public awareness of the transience of archaeological heritage. In fact, in our collective memory there is still room to remember the devastations caused by terrorist groups in Afghanistan, Iraq and Syria, who, in the last 15 years, destroyed world heritage sites and monuments of splendid civilizations spared by millennia making archaeology another casualty of their madness [18]. The dense 3D reconstruction of the destroyed Bel Temple in Palmyra [19] is an emblematic example of how much is necessary the development and employment of high quality digitization techniques in the Cultural Heritage field. Moreover, the

Digital Archaeology allows scholars from the whole world to explore new ways of digital restore old artifacts without the risk of damage them irreparably. The 3D reconstruction of a Cultural Heritage site can be performed through Imagery and techniques like Structure from Motion (SfM) [4]. We present some use-cases of this kind in sparse 3D reconstruction views browsable in The Social Picture framework. A Cultural Heritage site can be digitized also employing 3D scanners. We show a comparative benchmark of all the main hand-held (mobile) and fixed 3D scanners currently available in the market. Then, we describe a 3D web viewer that we have specifically developed for augment 3D model of artifacts (e.g., semantic notations, comments, markers on points of interest).

As regards the Medical context, we show outcomes on the female human breast surgeon research [20, 21]. The opportunity to acquire body part shapes, including soft tissues like the female human breast, with a 3D hand-held scanner has motivated our conjunct study with the medical specialists in breast reconstruction. Our main aim is to find a discriminative parametrization of female breast shape (i.e., a small set of parameters to objectively describe it). This kind of mathematical representation enables the possibility to easily define accurate metric for breast difference evaluation. This result is very attractive for breast surgeons, since it can be useful for pre/post surgeon patients monitoring and in performance and quality assessment of surgeons. It could also be an effective strategy to create clear and well-defined breast shape categories. We proposed a clinical procedure in which the female patients hold the hands behind and above the head, while an operator can digitize her breast with a 3D scanner. Then, we designed two approach for parametrization based on Principal Warps [22] and on Bag-of-Normals, respectively.

In summary, in this dissertation we present real use-cases in which a big amount of multimedia data is analyzed and summarized. The shown applications are mainly thought to manage online images, videos, and 3D data from multiple and different kind of smart devices. These applications may be the enabling technology for LTE-based multimedial services.

As part of his International Doctorate Fellowship, it should be noted that the PhD candidate spent 6 months abroad, as visiting scholar in the University of South Florida (USF), Florida, USA. He has been hosted by the Department of History of USF and worked in the Center for Virtualization and Applied Spatial Technologies

(CVAST) to several topics, i.e., 3D web viewers, color specification for archaeology, and 3D digital measurement.

The dissertation is structured as follows: Image Matching and Retrieval techniques and approaches are reported in Chapter 2. The framework named The Social Picture is deeply treated in Chapter 3. In Chapter 4 we define and describe the novel type of saliency named Social Saliency. Video Summarization and RECfusion framework are described in Chapter 5. The applications and real use-cases concerning 3D data are treated in Chapters 6 and 7, for Cultural Heritage and Medical Research, respectively. In Chapter 8 conclusions, remarks and possible future works conclude the main part of this dissertation. We inserted a work related to color specification and standardization in Appendix A and a 3D web viewers benchmark in Appendix B.

A more detailed description of the dissertation structure is reported in the next section, while authored and co-authored publications are listed in Section 1.2.

1.1 Dissertation Structure

In this dissertation, titled *Multi-Device Media Analysis and Summarization for High Bandwidth Connected Environment*, we mainly treated three kinds of media: images, videos, and 3D data. For this reason, the dissertation is properly divided into 3 corresponding parts. Dissertation structure is shown in Fig. 1.1. We started our discussion from images, as compared to the other two kinds of media they have the most simple spatial and temporal context. Firstly, we faced the image matching and retrieval issue. Findings of this chapter have been used in many algorithms employed in the framework The Social Picture (TSP), which is capable of collect, analyze, and organize huge flows of visual data. Indeed, it has been used to acquire datasets of images and investigate our new kind of saliency that we named *Social Saliency*. TSP is also a good place for store output of video and 3D data analysis and summarization. More in detail, we exploited 3D representation of Social Saliency to augment 3D data for Cultural Heritage. Two appendices are attached to this dissertation. The first one, Automatic Recognition of Color for Archaeology (ARCA), has been used in pre-processing of images and videos, while the second one, related to 3D web viewers design, presents perfect applications for spread and

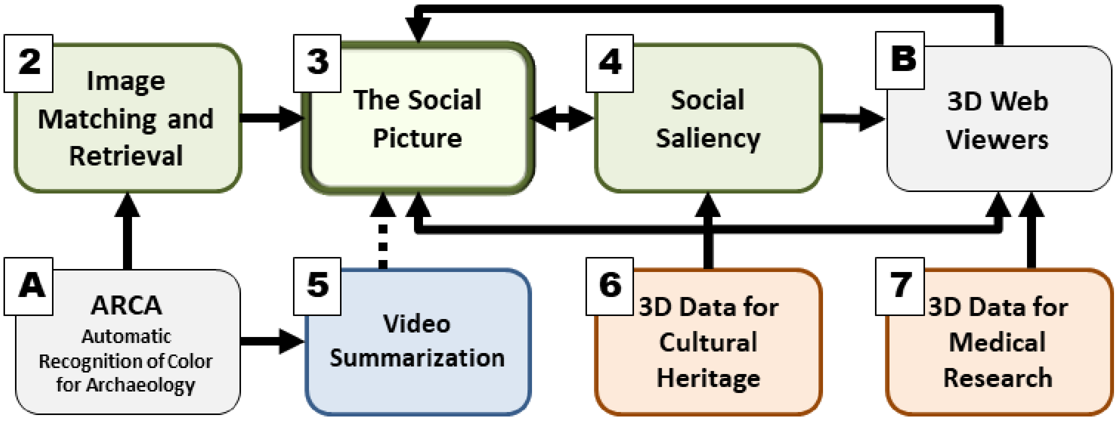

2
Image Matching and Retrieval
3
The Social Picture
4
Social Saliency
B
3D Web Viewers
A
ARCA
Automatic Recognition of Color for Archaeology
5
Video Summarization
6
3D Data for Cultural Heritage
7
3D Data for Medical Research

Figure 1.1: Dissertation Structure. Numbers in black depicts the Chapters. Chapters colored in green, blue, red, and gray are related to images, videos, 3D data, and appendices, respectively. Black arrows marks how knowledge acquired and findings from a chapter are mostly employed in the other ones.

browse results of 3D data analysis for Cultural Heritage and Medical Research. Moreover, we integrated a prototype of 3D web viewers in TSP, customized to show Social Saliency together with the output of Structure from Motion techniques based on the part of Image Matching and Retrieval treated in the beginning.

1.2 List of Publications

In this dissertation we present 4 real use-cases, co-authored and published in international journals and conferences.

1. **The Social Picture:**

   - S. Battiato, G.M. Farinella, F.L.M. Milotta, A. Ortis, L. Addesso, A. Casella, V. D'Amico, and G. Torrisi. "The Social Picture". In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM. 2016, pp. 397–400.

     – The Social Picture is a framework to collect and explore huge amount of crowdsourced social images about public events, cultural heritage sites and other customized private events. Collections can be explored through a number of advanced Computer Vision and Machine Learning algorithms, able to capture the visual content of images in order to organize them in a semantic way.

- F.L.M. Milotta, M. Bellocchi, and S. Battiato. “The Social Picture: Advanced Image Analysis Applications”. In: STAG: Smart Tools and Applications in Graphics (2017). 2017.

  – In this extension of The Social Picture, we implemented two advanced Image Analysis applications: scene summarization and features density representation through several kinds of density maps.

2. **RECfusion:**

- F.L.M. Milotta, S. Battiato, F. Stanco, V. D’Amico, G. Torrisi, and L. Addesso. “RECfusion: Automatic Scene Clustering and Tracking in Video from Multiple Sources”. In: EI – Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications. 2016.

  – In this work we proposed an extended version of the RECfusion framework. Main aims are: analysis of video streams from multi-source multi-device context, identification of scenes of interest through automatic clustering of video sequences, and time tracking of the computed scenes clusters.

- S. Battiato, G.M. Farinella, F.L.M. Milotta, A. Ortis, F. Stanco, V. D’Amico, L. Addesso, and G. Torrisi. “Organizing Videos Streams for Clustering and Estimation of Popular Scense”. In: 19th International Conference on Image Analysis and Processing (ICIAP 2017). 2017.

  – This further extension of RECfusion is able to generate a video which captures the interests of the crowd starting from a set of the videos by considering scene content popularity. Frames composing the final popular video are automatically selected from the different video streams with an improved procedure that does not employ a threshold anymore.

3. **3D Data Analysis for Cultural Heritage:**

- M.F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, F.L.M. Milotta, S. Schiavone, and F. Stanco. “Archaeometric characterization and 3D survey: new perspectives for monitoring and valorization of Morgantina silver Treasure (Sicily)”. In: (2015).

– In this work we started the design and development of a 3D web viewer for Cultural Heritage purposes.

- D. Allegra, E. Ciliberto, P. Ciliberto, F.L.M. Milotta, G. Petrillo, F. Stanco, and C. Trombatore. "Virtual unrolling using x-ray computed tomography". In: Signal Processing Conference (EUSIPCO), 2015 23rd European. IEEE. 2015, pp. 2864–2868.

  – In this paper we addressed the problem of virtual unrolling to read papyrus scroll by avoiding a dangerous physical unrolling.

- D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F.L.M. Milotta, C. Santagati, and F. Stanco. "Low cost handheld 3D scanning for architectural elements acquisition". In: Proceedings of the Conference on Smart Tools and Applications in Computer Graphics. Eurographics Association. 2016, pp. 127–131.

  – In this study we focused the attention on one emerging technology, the Structure Sensor device, in order to verify a 3D pipeline acquisition on an architectural element and its details.

- F. Stanco, D. Tanasi, D. Allegra, and F.L.M. Milotta. "3D digital imaging for knowledge dissemination of Greek archaic statuary". In: Proceedings of the Conference on Smart Tools and Applications in Computer Graphics. Eurographics Association. 2016, pp. 133–141.

  – By means of 3D scanning techniques, this contribution showcases how virtual restoration can not only improve interpretations of the scholars, but also boost the communication plans of museums, giving back to the public via a web platform a masterpiece of Greek sculpture known just by specialists.

- F. Stanco, D. Tanasi, D. Allegra, F.L.M. Milotta, G. Lamagna, and G. Monterosso. "Virtual anastylosis of Greek sculpture as museum policy for public outreach and cognitive accessibility". In: Journal of Electronic Imaging 26.1 (2017), pp. 011025–011025.

  – This work deals with a virtual anastylosis of a Greek Archaic statue from ancient Sicily and the development of a public outreach protocol for those with visual impairment or cognitive disabilities through the application of three-dimensional (3-D) printing and haptic technology.

- D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F.L.M. Milotta, C. Santagati, and F. Stanco. "Hand Held 3D Scanning for Cultural Heritage: Experimenting Low Cost Structure Sensor Scan". In: Handbook of Research on Emerging Technologies for Architectural and Archaeological Heritage. IGI Global, 2017, pp. 475–499.

  – In this chapter of book we focused on the handheld 3D scanner and provided a description of the products currently available on the market.

- M. F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, G. Milazzo, F.L.M. Milotta, L. Pellegrino, S. Schiavone, and F. Stanco. "Integrated three-dimensional models for noninvasive monitoring and valorization of the Morgantina silver treasure (Sicily)". In: Journal of Electronic Imaging 26.1 (2017), pp. 011015–011015.

  – All acquired data, i.e. 3D models, UV fluorescence and X-Ray images and chemical information, have been made available, in an integrated way, within a web-oriented platform, that represents a in progress tool to deepen the existing archaeological knowledge and production technologies and to obtain referenced information of the conservation state, before and after moving of the finds from their exposure site.

- F.L.M. Milotta, F. Stanco, and D. Tanasi. "ARCA (Automatic Recognition of Color for Archaeology): a Desktop Application for Munsell Estimation". In: 19th International Conference on Image Analysis and Processing (ICIAP 2017). 2017.

  – The following pipeline for Munsell estimation aimed towards archaeologists has been proposed in this work: image acquisition of specimens, manual sampling of the image in the ARCA desktop application, automatic Munsell estimation of the sampled points and creation of a sampling report.

4. **3D Data Analysis for Medical Research:**

- G. Gallo, D. Allegra, Y.G. Atani, F.L.M. Milotta, F. Stanco, and G. Catanuto. "Breast Shape Parametrization Through Planar Projections". In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer. 2016, pp. 135–146.

  - – The main aim of this work is to propose an innovative strategy to automatically analyze 3D breast shape in order to describe them within a quantitative well defined framework.

- D. Allegra, F.L.M. Milotta, D. Sinit‘o, F. Stanco, G. Gallo, T. Wafa, and G. Catanuto. “Description of Breast Morphology through Bag of Normals Representation”. In: 19th International Conference on Image Analysis and Processing (ICIAP 2017). Springer. 2017.

  - – In this work we focus on digital shape analysis of breast models to assist breast surgeon for medical and surgical purposes. A clinical procedure for female breast digital scan is proposed. PCA is computed and the obtained first 2 principal components are used to plot the breasts shape into a 2D space.

All authors contributed equally to these works.

Part I - Images

Chapter 2

Image Matching and Retrieval

2.1 Introduction

Image datasets used in our framework The Social Picture (TSP - Chapter 3) are made up of images gathered by social media or directly uploaded by users through a properly designed application. Datasets of images are usually browsed by means of queries based on tags, hash-tags or geodata (where present). Queries can be even done using other images and looking for the most similar ones in the dataset. This procedure is called "Content Based Image Retrieval" (CBIR) and is based on Image Matching algorithms [23]. Formally, in the Image Matching procedure, images are distinguished in one or multiple "Query" images to be matched with a single "Reference" image[1]. Since pictures are gathered by social networks (e.g., Flickr, Instagram, Panoramio), then we need to manage data at a scale bigger than the one of a traditional database. For this reason, algorithms designed for handle our dataset are thought as strictly related to Big Data approaches [24]. Moreover, queries may be launched not only on whole images but even on smaller parts of them. In other words, one can be interested in looking for a specific object or element in a scene, possibly acquired in several images by different points of view. In this case, we refer to "Visual Search", another task based on Image Matching techniques [25].

We employed a *heatmap* to represent how query images match with a reference one. Lots of different colormaps can be used in a heatmap (e.g., hsv, hot, cool, jet; Fig. 2.1). We chose the jet colormap, so colors will range from blue (low values) to red (high values). With this representation, the heatmap highlights in red parts of

[1] For the sake of curiosity, in Video Motion Estimation procedure it is possible to find a similar definition: frames are distinguished in one ore multiple "Target" frames to be compared to a single "Anchor" frame.

the reference image with a high number of positive match with the query images. In the past, heatmaps have already been used for Image Retrieval representation [26, 27]. Moreover, heatmaps are often used in geodata-based Image Retrieval tasks. For instance, it is possible to visualize a geographical map with a superimposed heatmap. In this case, heatmaps highlight geographical places in which many multimedial data have been acquired. Many libraries for implementation of this feature are currently available online; among the many, we refer to [28, 29]. In particular, we employed [29] in The Social Picture (Chapter 3) for landmarks selection purpose. However, for the rest of this chapter we will treat Image Matching-based heatmaps, and not geodata-based.

The structure of the Chapter is the following: in Section 2.2 we discuss related works about Content Based Image Retrieval (CBIR) and Image Matching. In Section 2.3 we describe the heatmap computation procedure. In Section 2.4 we report experimental settings and results. Discussion and conclusions end the Chapter in Section 2.5.

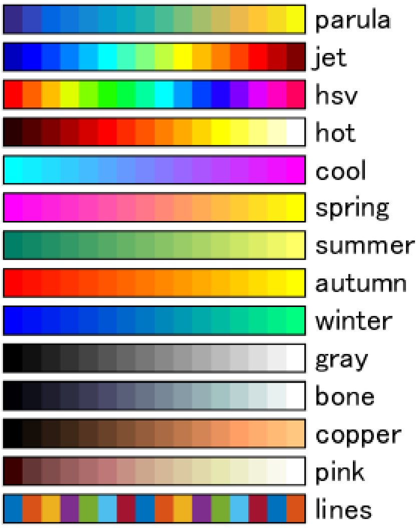


Figure 2.1: Examples of available colormaps in Matlab environment [30].

2.2 Related Works

Content Based Image Retrieval (CBIR) has been deeply investigated in the last years. Rui et al. [23] stated that CBIR is mostly related to Database Management and Computer Vision. They reported that initially CBIR was performed through a careful and complete annotation phase of all the images in a dataset. However, when image collections started to be bigger and bigger this approach was clearly no good anymore. This was the point in which Computer Vision community developed feature-based algorithms. With the term *features* they refer to color [31, 32], shape, texture [33, 34] and segmentation.

Actually, such features are pretty the same in all the CBIR algorithms. Moreover, there is a recurrent distinction between low and high level features. The gap between these two levels is known as *semantic gap*: it raises from the difficulty to relate low-level features like objects in the scene with the context of the scene itself. Works related to CBIR semantic gap include [35, 36, 37]. In Liu et al. [35] identified different semantic levels for features useful for CBIR and defined 5 categories of algorithms taking into account semantic. Categories are related to object ontologies, machine learning and metadata (for images downloaded from web). Rehman et al. [36] presented a survey of CBIR methods focusing on semantic gap and referred the method named CLUE [38] as possible solution to reduce the semantic gap through clustering of the images. Wang et al. [37] presented a method to reduce semantic gap on image collections from the Web. Their method employs textual information crawled from the Web, too. Indeed, textual indexing techniques are strictly related to CBIR. Doermann [39] presented a survey on indexing and retrieval methods for images in documents in which there is a great amount of textual information.

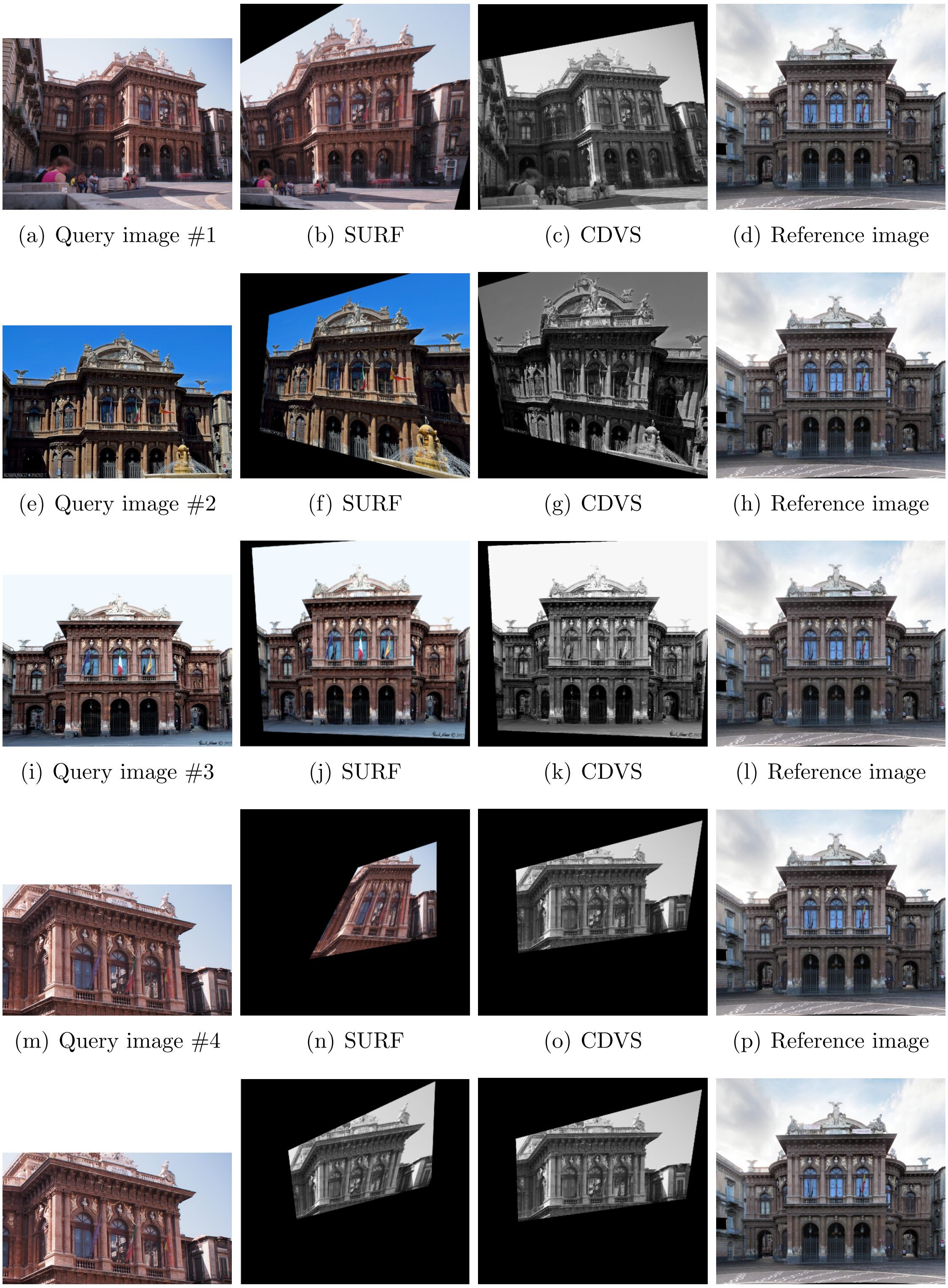
Speaking of textual indexing and retrieval, an usually referred topic is the bag-of-words method. In this information-retrieval approach a *vocabulary* is created and used for retrieval, disregarding the grammar or single words [40, 41]. In CBIR a similar method exists and is named bag-of-visual-words (BoVW). In the BoVW-based techniques images are described as a collection of visual words, as well as a book is described by the collection of its words. A *visual word* is defined as a particular descriptive region of the image. A vocabulary of visual words is built processing the whole collection of images. Visual words are further processed with several filters (e.g., gradient, color), gaining a bunch of descriptive codewords, hence the vocabulary is also named codebook. Works related to BoVW include [42, 43, 44]. Two applications of BoVW have been presented in [45, 46] for landmark search and land-use classification, respectively.

We already distinguished low level features from high level ones. Another distinction can be made for descriptors of visual features: they can be distinguished in local and global compact descriptors [11]. *Compact* means that these descriptors are optimized for matching and comparison in high bandwidth network. Optimization have an impact also on power consumption in mobile devices [11]. Examples of compact local descriptors can be found in [47, 48], where a compressed histogram of gradient and a grid-based quantization approach to code the spatial layout of local features is applied, respectively. Examples of compact global descriptors are works in which BoVW are compressed in the so called bag-of-features (BoF) [49, 45, 50] or the work in which vector of locally aggregated descriptors (VLAD) is defined for the first time [51]. The Moving Picture Experts Group (MPEG) has grouped these compact descriptors in the standardized MPEG Compact Descriptors for Visual Search (CDVS) method [11]. Telecom Italia presented an improvement to MPEG CDVS in [52], that is the one employed in this dissertation, too. This choice is supported by the fact that transformations estimated by [52] are clearly way better than the one estimated by CBIR methods, performed with a well assessed descriptor like SURF (Fig. 2.2). Moreover, CDVS is faster than SURF and other main CBIR descriptors (Section 2.4 - Table 2.7) and manages to reach comparable results only if Query Expansion (Section 2.3.2) is employed (Fig. 2.2(r)).

2.2.1 Content Based Image Retrieval Workflow

Content Based Image Retrieval (CBIR) has a consolidated workflow, well defined in [53, 54]:

1. **Keypoints detection**: keypoints are discriminative and descriptive points or part of the scene depicted in an image. Scale-invariant, rotation-invariant and geometric-transformation-invariant are important qualities that keypoints should have. Affine Covariant Detectors [53] include Harris-Affine [55, 56, 57], Hessian-Affine [55, 56], Maximally Stable Extremal Regions (MSER) [58], edge-based regions [59] and intensity extrema-based regions [60].



(a) Query image #1 (b) SURF (c) CDVS (d) Reference image

(e) Query image #2 (f) SURF (g) CDVS (h) Reference image

(i) Query image #3 (j) SURF (k) CDVS (l) Reference image

(m) Query image #4 (n) SURF (o) CDVS (p) Reference image

(q) Query image #4 (r) SURF+QueryExp. (s) CDVS (t) Reference image

Figure 2.2: Visual comparison of transformations estimated by SURF and CDVS. Transformations by CDVS are better or equal to the ones by SURF.

2. **Features extraction from detected keypoints**: once that keypoints are detected, than they are formally characterized by a proper "descriptor". Usually, the spatial neighborhood of a keypoint is analyzed and several characteristics are taken into account (i.e., gradient, lumimance, statistical behavior of chrominance). Features descriptors include Scale Invariant Feature Transform (SIFT) [61, 62], Speeded-Up Robust Features (SURF) [63] and Gradient Location-Orientation Histogram (GLOH) [54].

3. **Features matching**: once that features have been extracted and represented with specific descriptors, than a similarity metric between descriptors is defined. In other words, a distance function between descriptors is defined. This metric is used to decide when descriptors on different images are representing the same keypoint in the scene. Usually, this metric is the Euclidean distance [54], but other definitions are possible (i.e., Minkowski-type metric [35]).

4. **Geometric verification**: the output of Step 3 is a set of features pairs. Accordingly to the chosen similarity metric, matched features are very similar each other. However, at this point, matches should be considered just as putative ones. In order to decide how many matches are correct and coherent with the relation to the images context, it is needed to take into account the totality of the putative matches. A mathematical model is estimated trying to fit distance, position and rotation of features in the images. In other words, a geometric verification is estimated. Geometric verification algorithms include RANSAC [64] and its improved versions, like PROSAC [65] or SCRAMSAC [66].

5. **Inliers extraction**: all the features matches that are not coherent with the mathematical model estimated in Step 4 are considered as "Outliers", while on the countrary the others are defined "Inliers". Using inlier matches, geometric relationships between the reference image and query images is estimated. Then, it is possible to "warp" query images into the reference one. In other words, the estimated homography can be applied to a query image to deform it and make it more similar to the reference image.

2.3 Heatmap Computation

When a reference image is chosen, then a heatmap of matched query images can be computed. The heatmap is updated with relation to query images processed with the Content Based Image Retrieval workflow (Section 2.2.1). In this dissertation, we employed an improved version of the Compact Descriptors for Visual Search (CDVS), developed by Telecom Italia [52]. A detailed description of our implementation of the heatmap is given in the rest of this Section.

2.3.1 Heatmap Update Workflow

Our heatmap update workflow, as employed in The Social Picture (TSP [3, 6] - Chapter 3), is the following:

1. **Heatmap Initialization**: we define two variables. Let H_C be the heatmap, represented as a matrix of zero-values and size equals to the one of the reference image; this matrix will be used to keep track through cumulative updates of *how much* each query image contributes to heatmap. Let H_I be another data structure, used to keep track of *which* query image is giving its contribution to any heatmap part (Section 2.3.1 - Image Indexing). In other words, eventually, H_C will become the heatmap itself, highlighting the most viewed parts of reference image, while H_I will be used to know *who* are the query images that generate a particular heatmap value, in a specific coordinate of the map.

2. **Tolerance on Inliers number**: a threshold T_i is needed to decide if the match between query and reference image is positive or not. This threshold checks the number of inlier matches on query image. If the number is lower, then the image is discarded, otherwise workflow proceeds to Step 3. Notice that the "eight points algorithm" [67] requires that at least 8 inliers should be selected, in order to have a good likelihood of obtain a valid transformation, without any further assumption.

3. **Heatmap Update**: for each positively matched query image:

   3.1. **Image Transform**: Query image is "warped" (deformed) applying the homography estimated in Step 5 of Image Retrieval workflow (Section 2.2.1).

3.2. **Masking of H_C:**

3.2.1. **Cumulative update of H_C**: in heatmap update we chose to give a major weight (importance) to images with higher size. For this reason, each image contributes to H_C with a value equals to the ratio $\frac{QueryImageSize}{ReferenceImageSize}$. The contribution of query images is summed to H_C in a cumulative approach, only in the coordinates masked with warped query image.

3.2.2. **Update of H_I**: index of the query image is added in coordinates of H_I which correspond to coordinates of H_C masked with the warped query image.

4. **Heatmap Smoothing**: we apply Gaussian smoothing (2D filter dimension equals to 51, sigma equals to 150) to H_C for different purposes: noise reduction, gain a representation with blurry edge (that is more fair than the one with sharp edges) and, in particular, to go from a quantitative representation of the number of matches to a probabilistic representation. Indeed, this kind of Gaussian smoothing gets the same result of processing H_C with a Parzen Window method [68]. Hence, after smoothing, we gain a heatmap that represents area of reference image with higher/lower probability of be matched with other query images.

Matched Images Indexing

The data structure H_I is used in conjunction with H_C by users: they select a point of H_C and, through H_I, they can know who are the query images contributing in that point. H_I is a complex data structure to be handled and stored in memory. Indeed, it is a matrix of lists of indices. Although matrix can be vectorized, the set of lists still require lots of memory in reference images with great size. For instance, with a reference image whose size is 640 × 480, with an average number of 300 matched query images per coordinate, then we will get 640 × 480 lists with an average of 300 indices. A simple optimization is to quantize the set of coordinates (i.e., create a grid on the image). For instance, with a reference image whose size is 100 × 100 and quantization block whose dimension is 5 × 5, one is able to reduce the initial set of coordinates to a 20 × 20 grid, requiring a memory space that is $\frac{1}{(5 \times 5)}$ times smaller than the initial one.

On the other hand, we reduce the precision of H_I. However, considering that users select points on the heatmap using the mouse cursor, blocks with sizes like 5×5, 9×9 or even 21×21 represent a good solution and compromise to keep a fair experience while exploring H_C (browse indexed queries) and lower the required complexity and memory resources.

Back-Projection Verification

Sometimes, the threshold set in Step 2 of Heatmap Update Workflow (Section 2.3.1) may be not enough to guarantee a good image transormation in Step 3. If this wrong transformed images contribute in a cumulative update of H_C (Step 3.2.1), then noise is introduced and even worst, when Query Expansion is employed (Section 2.3.2).

A Back-Projection Verification step can be added [27], in order to discard wrong transformations. This step modifies Step 3.1 of Heatmap Update Workflow as follows:

3. **Heatmap Update**: for each positively matched query image:

    3.1. **Image Transform with Back-Projection Verification**: Query image is "warped" (deformed) applying the homography estimated in Step 5 of Image Retrieval workflow (Section 2.2.1). Then, we compute Image matching between warped query image and the original query image. Warped query image is back-projected in the original query image. Using the same tolerance on inliers number chosen in Step 2 we check if back-projection is verified. In case of successful outcome, workflow proceeds to Step 3.2, otherwise image is discarded.

A visual comparison between image transformations without and with Back-Projection Verification is shown in Fig. 2.3 and Fig. 2.4, respectively.

2.3.2 Query Expansion (Re-Query)

In Step 1 of Image Retrieval workflow (Section 2.2.1), we said that keypoints should be as most as possible scale-invariant, rotation-invariant and geometric-transformation-invariant. However, sometimes these statements apply partially or nothing at all.

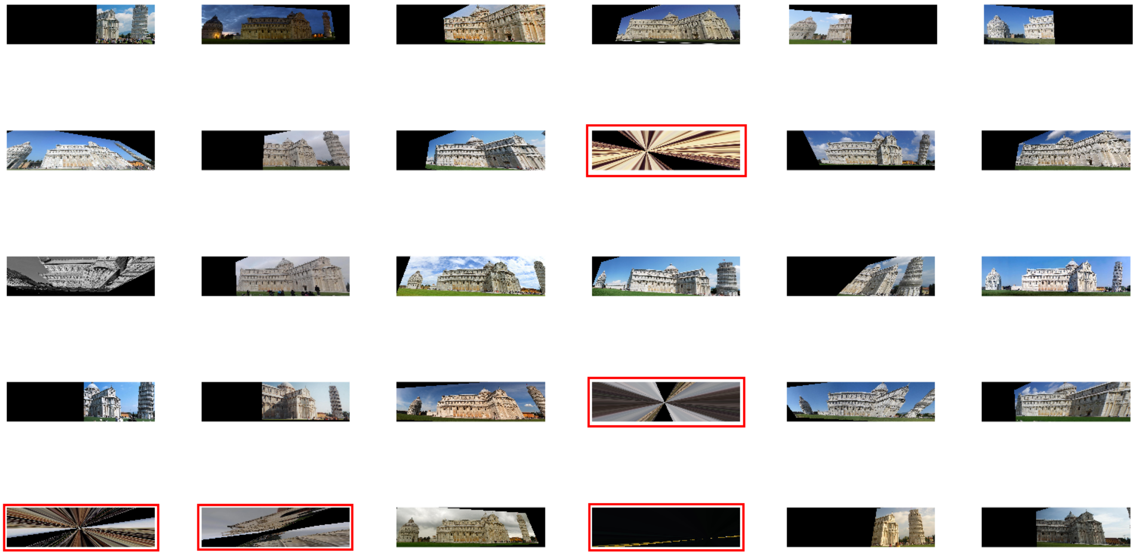


Figure 2.3: Image transformations without Back-Projection Verification. Totally wrong transformations are highlighted with red boxes. (Images from test *CDVS-8* - Section 2.4)

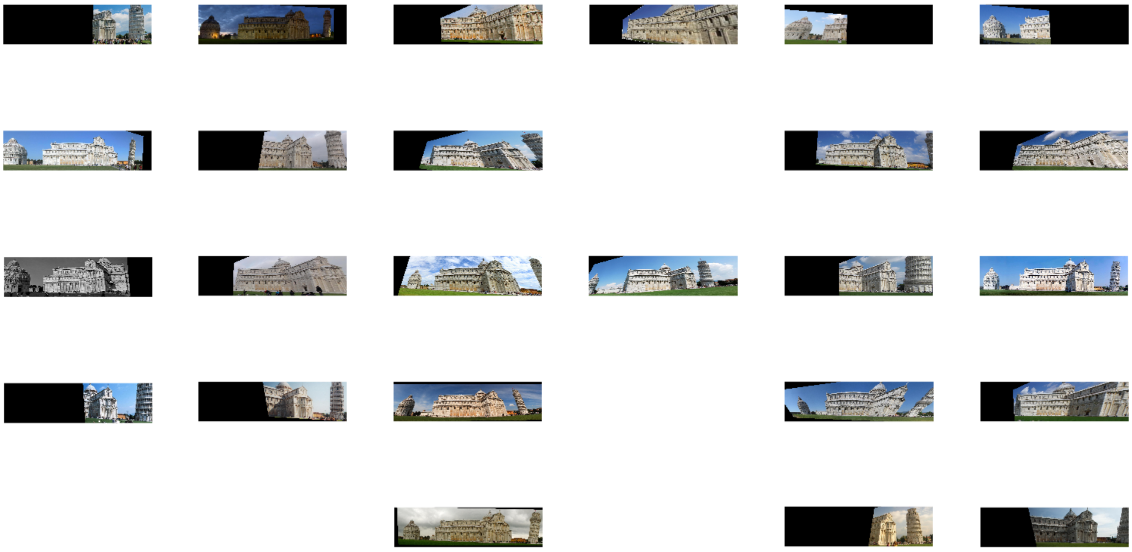


Figure 2.4: Image transformations with Back-Projection Verification. Totally wrong transformations are correctly discarded. (Images from test *CDVS_BP-3* - Section 2.4)

For instance, this happens when a query image depicts the same scene of the reference image, but from a point of view very different (e.g., angle of view, distance of the observer, ligth conditions). Image Retrieval algorithms will hardly match the two images. Moreover, query images may contain elements of the surrounding environment that are just partially depicted in the reference image, hence they will not be matched. For solve these kind of issues, a single step of Image Matching is not enough. It is possible to match two query images between themeselves and then find the nested relationship for match the query image discarded in the first step of Image Matching. This procedure is named Query Expansion (or Re-Query) [27, 69].

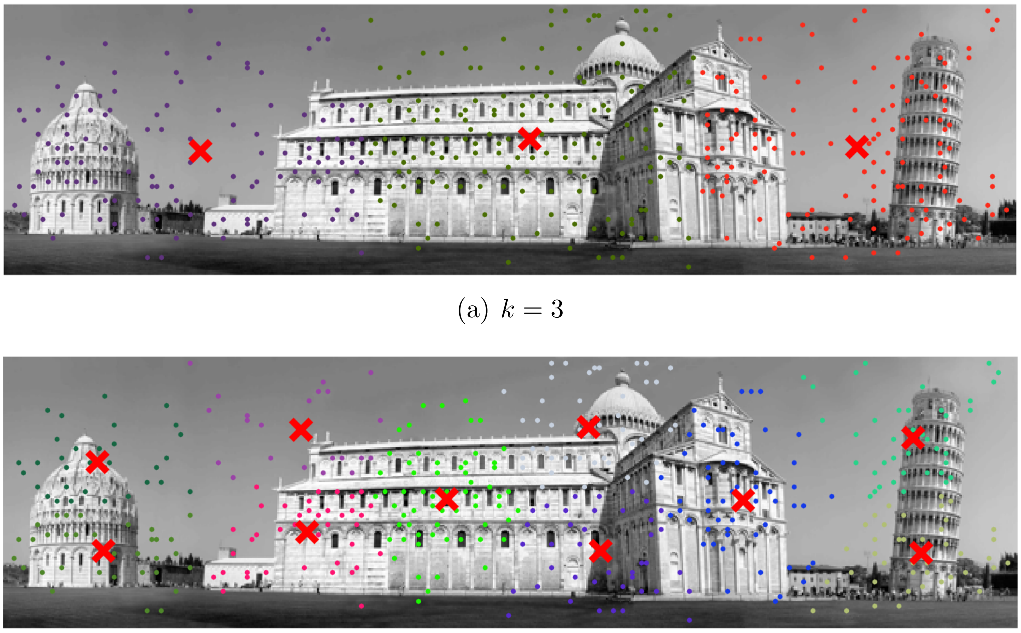
Starter Points

How to start Query Expansion? The exhaustive solution is to consider each query image as a reference one, at least once. However, this approach is high time consuming due to number of required comparisons. Another solution could be to sort all the query images with respect to their number of inliers, and consider just a subset of them. However, in this way, Query Expansion may be done only on already well matched query images, ignoring images potentially matchable, but discarded. Hence, we chose a compromise: we perform k-means on spatial 2D coordinates of features extracted by CDVS [11, 52] in the reference image (Fig. 2.5). We named these k pairs of coordinates "Starter Points". Then, we employed H_I (Section 2.3.1) and created a list of all query images indexed in Starter Points. The list is sorted by number of inlier keypoints (descending order). Finally, the sorted listed can be used within the Query Expansion workflow.

We extracted Starter Points only from the initial reference image. This means that, although it was possible to further reiterate starter points extraction also on query images in the sorted list, we employed for them another approach. Properly, we chose the second strategy suggested before as possible solution for query expansion initialization: sort all the query images with respect to their number of inliers, and consider just a subset of them.

Query Expansion Workflow

The Query Expansion workflow is the following:



(a) $k = 3$

(b) $k = 10$

Figure 2.5: k-means spatial clustering of CDVS extracted features [11, 52]. Starter Points (marked with a red X) are centroids of clusters computed by k-means.

Let Q_{Lev} be the current level (iteration) of Query Expansion. Let Q_{Nes} the number of query images selected from the sorted list, for each k-th starter point, as described in the previous paragraph; "Nes" stands for *nesting*, as this parameter changes the degree of nesting of the Query Expansion workflow. Let Q be the queue of query images to be considered as new reference image in iteration $Q_{Lev} + 1$.

1. For $Q_{Lev} = 1$ do:

    1.1. Compute Image Matching between initial Reference Image and Query Images.

    1.2. Extraction of k *Starter Points*. For each starter point:

        1.2.1. Sort with descending order the query images, with respect to their number of inlier keypoints.

        1.2.2. Enqueue in Q the first Q_{Nes} query images in the sorted list.

    1.3. Increase Q_{Lev}.

2. For $Q_{Lev} > 1$ repeat:

    2.1. Compute Image Matching between Reference Image(s) in Q and not yet matched Query Images.

    2.2. Sort with descending order the query images, with respect to their number of inlier keypoints.

    2.3. Enqueue in Q the first Q_{Nes} query images in the sorted list.

    2.4. Increase Q_{Lev}.

Some clarification: each quary image can be enqueued in Q only once and query images already matched with some reference image are not matched again. This is needed to avoid loop in the workflow. Moreover, homographies estimated in Step 2.1 are always related to the initial reference image (through a chain of homographies). Eventually, the procedure ends if one of the following stop criteria applies:

- Maximum Q_{Lev} has been reached.
- There are no more query images to be matched.
- Q is empty (there are no more reference images to be matched).

Other stop criteria may be defined. For instance, maximum numbers of matched images or query images enqueued in Q may be employed to further limit computational time.

To summarize, Query Expansion workflow can be represented as a tree shaped data-structure: the initial reference image is the root, while Q contains the other nodes in the tree. In $Q_{Lev} = 1$ there are k subtrees rooted in the initial reference image. Each one of these subtrees has Q_{Nes} children. For $Q_{Lev} > 1$, each node has Q_{Nes} children too. Homographies are always related to the root, through a chain of homographies that take into account all the homographies between the nodes in the path from each node to the root. This means that homographies in levels farer from the root have a more localized "impact" on the heatmap. Hence, thanks to Query Expansion, we are able to capture smaller details on initial reference image that are difficult to highlight with a single step of Image Matching.

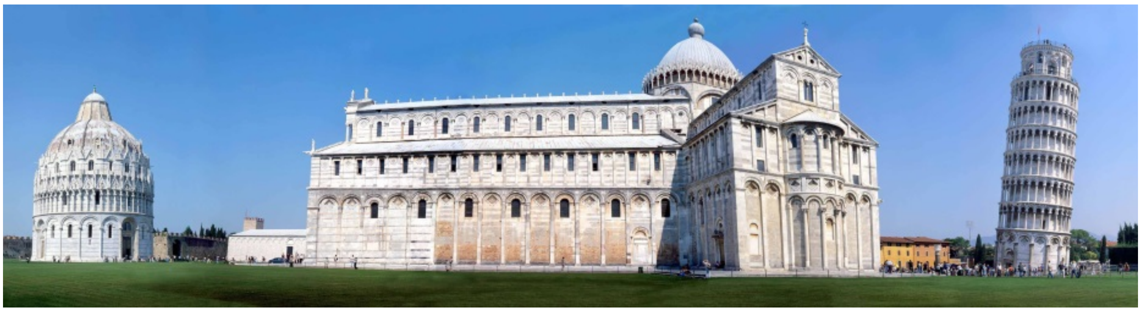


Figure 2.6: Reference Image of Pisa Collection used in heatmap computation tests.

2.4 Heatmap Computation Results

Heatmap is a valuable tool for data analysis and summarization that has been implemented within our framework The Social Picture (TSP) (Chapter 3). We tested different descriptors for heatmap computation: Speeded-Up Robust Features (SURF) [63], Maximally Stable Extremal Regions (MSER) [58], Scale Invariant Feature Transform (SIFT) [61] from VLFeat library [70] and the Compact Descriptors for Visual Search (CDVS [11]) implementation by Telecom Italia R&D team [52]. We employed a collection named *Pisa* of 2923 images, downloaded by social media (Flickr and Panoramio) and stored in TSP. Pisa collection mainly contains images of the landmark called *Piazza dei Miracoli*, since it was created through geodata-based queries targeting that place. There are three main subjects in the scenario: the famous Tower of Pisa, the Cathedral and the Saint John Baptistery. Images have been manually labeled with labels referred to these three main subjects. We distinguished 1651 "positive" images with at least a portion of the three main subjects partially visible, and 1272 "negative" images (noise). We chose a reference image (Fig. 2.6) in which all the main subjects are clearly visible. Then, we changed tuning parameters of selected descriptors, looking for the best computed heatmap.

In the results, we reported a field *Quality*: this is a subjective metric from 1 to 5 (represented with black stars: "★", as done in the common recommendation systems), computed as an average of the judgment given by 3 computer vision experts. They evaluated how the heatmap looks like, given the bias that buildings are more important than the background, and that Tower of Pisa should be the most important building in the scenario. They also evaluated how much the transformations (homographies) computed in Step 5 of Content Based Image Retrieval workflow (Section 2.2.1) look correct. For this purpose, they browsed query images indexed in the H_I matrix, as defined in Section 2.3.1. Clearly, creation of 1651 maps of ground truth for the verification of transformations is not feasible, so this evaluation is based on a subjective judgment, too.

Results are reported in Tables 2.2, 2.3, 2.4, 2.5 and 2.6, for SURF, MSER, VLFeat, CDVS and CDVS with Back-Projection Verification tests, respectively. Notice that Back-Projection Verification has been employed in SURF, MSER and VLFeat tests, even if not explicitly written in their TestIDs. Parameters listed in tables are defined in Table 2.1. In tables we reported Query Expansion parameters, however when $Q_{Nes} = 1$ and $Q_{Lev} = 1$, then this means that Query Expansion is not computed for that test. In the other cases, Query Expansion is computed employing k-means with $k = 3$, except for test *CDVS_BP-4*, where $k = 10$ (as stated in Table 2.6).

Performances of conducted tests are summarized in Table 2.7, where we sorted tests results by True Positive Rate (TPR), computational time and quality. Computed heatmaps are shown in Fig. 2.7. TPR values are very low for heatmaps ranked with high quality: this statements happen because we preferred an algorithm with high True Negative Rate (TNR) (that means, low False Positive Rate). Indeed, TNRs value are very high in almost the totality of tests. In this way, we retain in heatmap computation only good images (where landmarks or details of them are clearly visible), reducing noise. In other words, in uncertain cases it is better to discard than to update the heatmap. SURF, MSER and VLFeat have long computational time. We obtained the best result with them using SURF, after 11 days. The obtained heatmap is really good, but the needed time makes this method unfeasible. Then, we employed CDVS. It is way faster: just $30'$ are needed to compare reference image with all query images in *Pisa* collection. Improved with Back-Projection and Query Expansion, we were able to increase the number of total comparisons computed in the same time by other tested methods. With Query Expansion we reached a deeper level of detail. In this way, with test *CDVS_BP-2* we gain our best result in less than one day, hence almost 10 times better than the previous one obtained with SURF in 11 days.

Table 2.1: Legend of parameters names used in Tables 2.2, 2.3, 2.4, 2.5, 2.6 and 2.7.

T_i	Threshold on Inliers number (Section 2.2.1)
Q_{Lev}	Query Expansion maximum Level (Iterations) (Section 2.3.2)
Q_{Nes}	Query Expansion number of Nested query images (Section 2.3.2)
IndBS	Block Size for Matched Images Indexing (Section 2.3.1)
TPR	True Positive Rate
TNR	True Negative Rate
Time	Computational Time
Quality	Subjective quality of transformation

TD	MSER - Matlab [71]: Threshold Delta (Default: 2)
MAV	MSER - Matlab [71]: Max Area Variation (Default: 0.25)
T_S	MPEG-CDVS [52]: Threshold on score (Default: 4)
MMode	MPEG-CDVS [52]: Match Mode **LG**: local descriptor + global d.; **L**: local d.; **G**: global d.
MD	RANSAC - Matlab [72]: Max Distance (Default: 1.5)
RIT	RANSAC - Matlab [72]: maximum number of iterations
BP	Back-Projection Verification (Section 2.3.1)

Table 2.2: Heatmap computation performance employing SURF [63].

TestID	T_i	Q_{Nes}	Q_{Lev}	IndBS	TPR	TNR	Time	Quality
SURF-1	15	1	1	1	19.26%	99.06%	~1 d	★★★☆☆
SURF-2	12	1	1	1	29.50%	96.70%	~1 d	★★☆☆☆
SURF-3	15	1	1	1	15.87%	99.06%	~1 d	★★★☆☆
SURF-4	15	1	1	1	11.51%	99.37%	~1 d	★☆☆☆☆
SURF-5	23	1	1	1	10.66%	99.53%	~1 d	★★★☆☆
SURF-6	15	2	1	1	26.35%	98.82%	~2 d	★★★★☆
SURF-7	15	3	1	1	31.19%	98.90%	~3 d	★★★★☆
SURF-8	20	3	3	1	38.28%	98.35%	~11 d	★★★★★

Table 2.3: Heatmap computation performance employing MSER [58].

TestID	T_i	Q_{Nes}	Q_{Lev}	IndBS	TD	MAV	TPR	TNR	Time	Quality
MSER-1	12	1	1	1	0.20	0.25	11.57%	99.53%	~2 d	★★☆☆☆
MSER-2	13	1	1	1	0.20	0.25	10.84%	99.53%	~2 d	★★★☆☆
MSER-3	14	1	1	1	0.20	0.25	9.63%	99.76%	~2 d	★★☆☆☆
MSER-4	13	1	1	1	0.20	0.50	10.84%	99.61%	~2 d	★★★☆☆
MSER-5	13	1	1	1	0.10	0.25	10.30%	99.45%	~2 d	★★★☆☆
MSER-6	13	1	1	1	0.10	0.50	10.72%	99.53%	~2 d	★★★☆☆

Table 2.4: Heatmap computation performance employing VLFeat [70].

TestID	T_i	Q_{Nes}	Q_{Lev}	IndBS	TPR	TNR	Time	Quality
VLFeat-1	12	1	1	1	11.21%	97.41%	~1 d	★★☆☆☆

2.5 Conclusion

In this Chapter, we presented several Content Based Image Retrieval (CBIR) methods. We defined our heatmap, as employed in the framework The Social Picture

Table 2.5: Heatmap computation performance employing CDVS implementation by Telecom Italia R&D team [52].

TestID	**T_i**	**Q_{Nes}**	**Q_{Lev}**	**IndBS**	**T_S**	**MD**	**MMode**	**RIT**	**TPR**	**TNR**	**Time**	**Quality**
CDVS-1	13	3	3	21	4	25	LG	1000	63.11%	90.41%	~3.5 h	★☆☆☆☆
CDVS-2	13	4	4	21	4	25	LG	1000	82.19%	86.40%	~11 h	★☆☆☆☆
CDVS-3	20	3	3	21	4	35	LG	1000	23.62%	98.51%	~3 h	★☆☆☆☆
CDVS-4	20	3	4	21	4	35	LG	1000	25.44%	97.96%	~4.5 h	★☆☆☆☆
CDVS-5	13	1	1	21	0.5	25	G	2000	3.94%	99.53%	~20'	★★★☆☆
CDVS-6	13	1	1	21	0.5	25	L	5000	3.94%	99.53%	~15'	★★★☆☆
CDVS-7	13	1	1	21	0.5	35	L	2000	3.51%	99.76%	~20'	★★★☆☆
CDVS-8	13	1	1	21	0.5	35	L	5000	3.51%	99.76%	~25'	★★☆☆☆
CDVS-9	13	1	1	21	0.5	35	L	10000	3.51%	99.76%	~30'	★★☆☆☆
CDVS-10	13	1	1	21	0.7	35	L	5000	2.79%	100.00%	~25'	★★★☆☆
CDVS-11	13	3	4	21	0.7	35	L	5000	45.12%	95.36%	~5 h	★☆☆☆☆
CDVS-12	13	1	1	21	0.7	35	G	10000	1.70%	100.00%	~20'	★☆☆☆☆
CDVS-13	13	3	3	21	0.75	35	G	10000	19.26%	98.66%	~3 h	★★★★☆

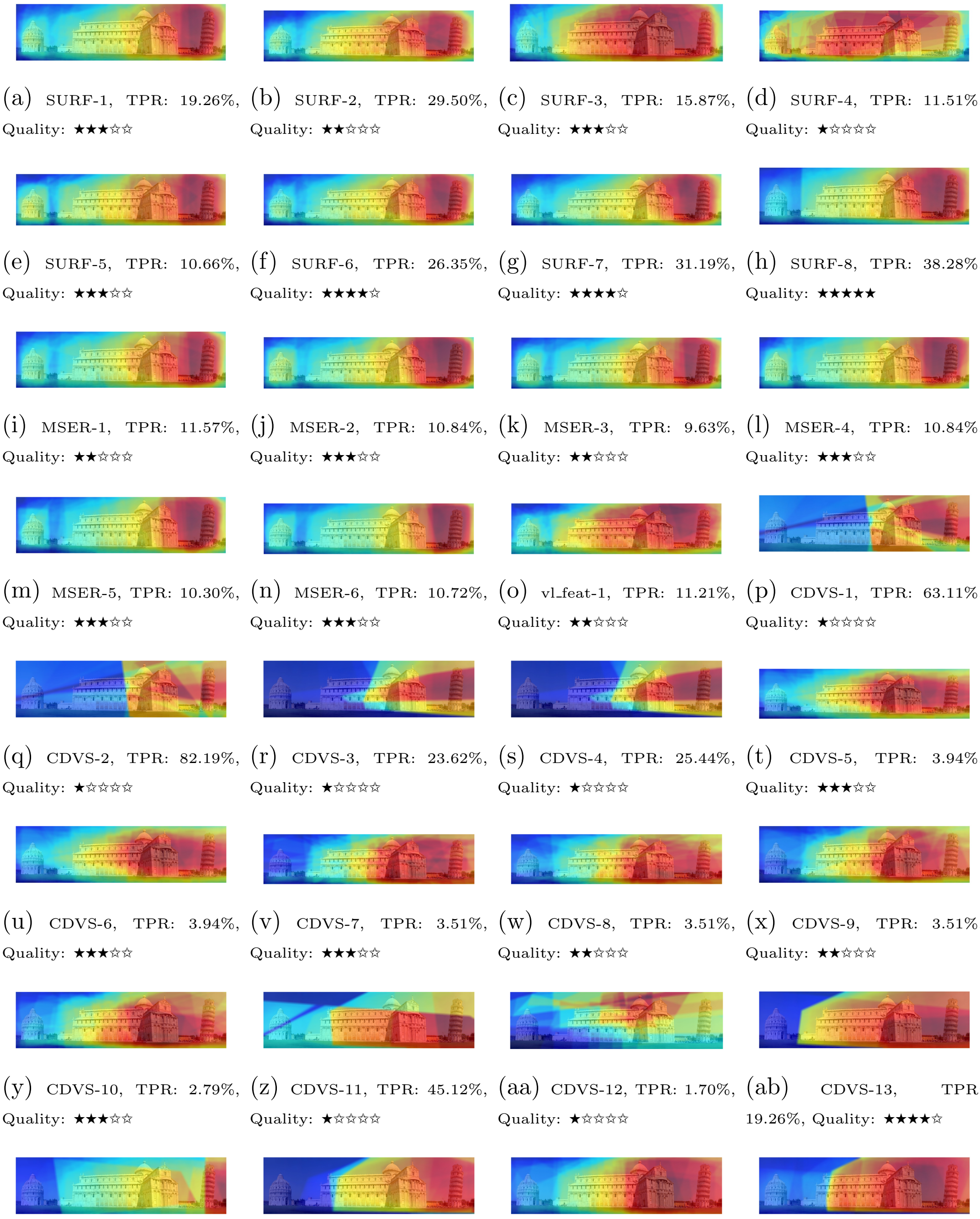
Table 2.6: Heatmap computation performance employing CDVS implementation by Telecom Italia R&D team [52] with Back Projection Verification (Section 2.3).

TestID	**T_i**	**Q_{Nes}**	**Q_{Lev}**	**IndBS**	**T_S**	**MD**	**MMode**	**RIT**	**TPR**	**TNR**	**Time**	**Quality**
CVDS_BP-1	13	3	3	21	0.5	35	G	5000	12.05%	98.90%	~3.5 h	★★★☆☆
CVDS_BP-2	13	4	4	21	0.5	35	G	5000	47.73%	96.31%	~18 h	★★★★★
CVDS_BP-3	13	1	1	21	0.5	35	L	5000	2.54%	99.92%	~20'	★★★★☆
CVDS_BP-4	13	$(k = 10)$ 3	1	21	4	35	L	10000	28.47%	97.72%	~4 h	★★★☆☆

(TSP - Chapter 3). Heatmap represents a valuable tool for analysis and summarization of large images collections. We compared 4 CBIR descriptors: Speeded-Up Robust Features (SURF) [63], Maximally Stable Extremal Regions (MSER) [58], Scale Invariant Feature Transform (SIFT) [61] from VLFeat library [70] and the Compact Descriptors for Visual Search (CDVS [11]) implementation by Telecom Italia R&D team [52]. We shown how CDVS, improved with Back-Projection Verification and Query Expansion, outperforms other tested method. In TSP heatmap is used with an incremental approach: once that the image collection is created, than heatmap is computed for the first time. From that moment on, new images added to the collection should be compared and matched only with images in the Query Expansion structure, hence is fast to update the heatmap in an incremental way. Major details will be given in Chapter 3, where TSP is described.

Table 2.7: Heatmap computation performances comparison of tested methods, sorted by True Positive Rate (TPR), computational time and quality (as defined in Section 2.4).

Sorted by TPR (Decreasing order)				**Sorted by Time (Increasing order)**				**Sorted by Quality (Decreasing order)**			
TestID	**TPR**	**Time**	**Quality**	**TestID**	**TPR**	**Time**	**Quality**	**TestID**	**TPR**	**Time**	**Quality**
CDVS-2	82,19%	~11 h	★☆☆☆☆	*CDVS-6*	3,94%	~15’	★★★☆☆	*CDVS_BP-2*	47,73%	~18 h	★★★★★
CDVS-1	63,11%	~3.5 h	★☆☆☆☆	*CDVS-5*	3,94%	~20’	★★★☆☆	*SURF-8*	38,28%	~11 d	★★★★★
CDVS_BP-2	47,73%	~18 h	★★★★★	*CDVS-7*	3,51%	~20’	★★★☆☆	*SURF-7*	31,19%	~3 d	★★★★☆
CDVS-11	45,12%	~5 h	★☆☆☆☆	*CDVS-12*	1,70%	~20’	★☆☆☆☆	*SURF-6*	26,35%	~2 d	★★★★☆
SURF-8	38,28%	~11 d	★★★★★	*CDVS_BP-3*	2,54%	~20’	★★★★☆	*CDVS-13*	19,26%	~3 h	★★★★☆
SURF-7	31,19%	~3 d	★★★★☆	*CDVS-8*	3,51%	~25 m	★★☆☆☆	*CDVS_BP-3*	2,54%	~20’	★★★★☆
SURF-2	29,50%	~1 d	★★☆☆☆	*CDVS-10*	2,79%	~25’	★★★☆☆	*CDVS_BP-4*	28,47%	~4 h	★★★☆☆
CDVS_BP-4	28,47%	~4 h	★★★☆☆	*CDVS-9*	3,51%	~30’	★★☆☆☆	*SURF-1*	19,26%	~1 d	★★★☆☆
SURF-6	26,35%	~2 d	★★★★☆	*CDVS-3*	23,62%	~3 h	★☆☆☆☆	*SURF-3*	15,87%	~1 d	★★★☆☆
CDVS-4	25,44%	~4.5 h	★☆☆☆☆	*CDVS-13*	19,26%	~3 h	★★★★☆	*CDVS_BP-1*	12,05%	~3.5 h	★★★☆☆
CDVS-3	23,62%	~3 h	★☆☆☆☆	*CDVS_BP-1*	12,05%	~3.5 h	★★★☆☆	*MSER-2*	10,84%	~2 d	★★★☆☆
SURF-1	19,26%	~1 d	★★★☆☆	*CDVS-1*	63,11%	~3.5 h	★☆☆☆☆	*MSER-4*	10,84%	~2 d	★★★☆☆
CDVS-13	19,26%	~3 h	★★★★☆	*CDVS_BP-4*	28,47%	~4 h	★★★☆☆	*MSER-6*	10,72%	~2 d	★★★☆☆
SURF-3	15,87%	~1 d	★★★☆☆	*CDVS-4*	25,44%	~4.5 h	★☆☆☆☆	*SURF-5*	10,66%	~1 d	★★★☆☆
CDVS_BP-1	12,05%	~3.5 h	★★★☆☆	*CDVS-11*	45,12%	~5 h	★☆☆☆☆	*MSER-5*	10,30%	~2 d	★★★☆☆
MSER-1	11,57%	~2 d	★★☆☆☆	*CDVS-2*	82,19%	~11 h	★☆☆☆☆	*CDVS-5*	3,94%	~20’	★★★☆☆
SURF-4	11,51%	~1 d	★☆☆☆☆	*CDVS_BP-2*	47,73%	~18 h	★★★★★	*CDVS-6*	3,94%	~15’	★★★☆☆
VLFeat-1	11,21%	~1 d	★★☆☆☆	*SURF-2*	29,50%	~1 d	★★☆☆☆	*CDVS-7*	3,51%	~20’	★★★☆☆
MSER-2	10,84%	~2 d	★★★☆☆	*SURF-1*	19,26%	~1 d	★★★☆☆	*CDVS-10*	2,79%	~25’	★★★☆☆
MSER-4	10,84%	~2 d	★★★☆☆	*SURF-3*	15,87%	~1 d	★★★☆☆	*SURF-2*	29,50%	~1 d	★★☆☆☆
MSER-6	10,72%	~2 d	★★★☆☆	*SURF-4*	11,51%	~1 d	★☆☆☆☆	*MSER-1*	11,57%	~2 d	★★☆☆☆
SURF-5	10,66%	~1 d	★★★☆☆	*VLFeat-1*	11,21%	~1 d	★★☆☆☆	*VLFeat-1*	11,21%	~1 d	★★☆☆☆
MSER-5	10,30%	~2 d	★★★☆☆	*SURF-5*	10,66%	~1 d	★★★☆☆	*MSER-3*	9,63%	~2 d	★★☆☆☆
MSER-3	9,63%	~2 d	★★☆☆☆	*SURF-6*	26,35%	~2 d	★★★★☆	*CDVS-8*	3,51%	~25’	★★☆☆☆
CDVS-5	3,94%	~20’	★★★☆☆	*MSER-1*	11,57%	~2 d	★★☆☆☆	*CDVS-9*	3,51%	~30’	★★☆☆☆
CDVS-6	3,94%	~15’	★★★☆☆	*MSER-2*	10,84%	~2 d	★★★☆☆	*SURF-4*	11,51%	~1 d	★☆☆☆☆
CDVS-7	3,51%	~20’	★★★☆☆	*MSER-4*	10,84%	~2 d	★★★☆☆	*CDVS-2*	82,19%	~11 h	★☆☆☆☆
CDVS-8	3,51%	~25’	★★☆☆☆	*MSER-6*	10,72%	~2 d	★★★☆☆	*CDVS-1*	63,11%	~3.5 h	★☆☆☆☆
CDVS-9	3,51%	~30’	★★☆☆☆	*MSER-5*	10,30%	~2 d	★★★☆☆	*CDVS-11*	45,12%	~5 h	★☆☆☆☆
CDVS-10	2,79%	~25’	★★★☆☆	*MSER-3*	9,63%	~2 d	★★☆☆☆	*CDVS-4*	25,44%	~4.5 h	★☆☆☆☆
CDVS_BP-3	2,54%	~20’	★★★★☆	*SURF-7*	31,19%	~3 d	★★★★☆	*CDVS-3*	23,62%	~3 h	★☆☆☆☆
CDVS-12	1,70%	~20’	★☆☆☆☆	*SURF-8*	38,28%	~11 d	★★★★★	*CDVS-12*	1,70%	~20’	★☆☆☆☆



(a) SURF-1, TPR: 19.26%, Quality: ★★★☆☆

(b) SURF-2, TPR: 29.50%, Quality: ★★☆☆☆

(c) SURF-3, TPR: 15.87%, Quality: ★★★☆☆

(d) SURF-4, TPR: 11.51%, Quality: ★☆☆☆☆

(e) SURF-5, TPR: 10.66%, Quality: ★★★☆☆

(f) SURF-6, TPR: 26.35%, Quality: ★★★★☆

(g) SURF-7, TPR: 31.19%, Quality: ★★★★☆

(h) SURF-8, TPR: 38.28%, Quality: ★★★★★

(i) MSER-1, TPR: 11.57%, Quality: ★★☆☆☆

(j) MSER-2, TPR: 10.84%, Quality: ★★★☆☆

(k) MSER-3, TPR: 9.63%, Quality: ★★☆☆☆

(l) MSER-4, TPR: 10.84%, Quality: ★★★☆☆

(m) MSER-5, TPR: 10.30%, Quality: ★★★☆☆

(n) MSER-6, TPR: 10.72%, Quality: ★★★☆☆

(o) vl_feat-1, TPR: 11.21%, Quality: ★★☆☆☆

(p) CDVS-1, TPR: 63.11%, Quality: ★☆☆☆☆

(q) CDVS-2, TPR: 82.19%, Quality: ★☆☆☆☆

(r) CDVS-3, TPR: 23.62%, Quality: ★☆☆☆☆

(s) CDVS-4, TPR: 25.44%, Quality: ★☆☆☆☆

(t) CDVS-5, TPR: 3.94%, Quality: ★★★☆☆

(u) CDVS-6, TPR: 3.94%, Quality: ★★★☆☆

(v) CDVS-7, TPR: 3.51%, Quality: ★★★☆☆

(w) CDVS-8, TPR: 3.51%, Quality: ★★☆☆☆

(x) CDVS-9, TPR: 3.51%, Quality: ★★☆☆☆

(y) CDVS-10, TPR: 2.79%, Quality: ★★★☆☆

(z) CDVS-11, TPR: 45.12%, Quality: ★☆☆☆☆

(aa) CDVS-12, TPR: 1.70%, Quality: ★☆☆☆☆

(ab) CDVS-13, TPR: 19.26%, Quality: ★★★★☆

(ac) CDVS_BP-1, TPR: 12.05%, Quality: ★★★☆☆

(ad) CDVS_BP-2, TPR: 47.73%, Quality: ★★★★★

(ae) CDVS_BP-3, TPR: 2.54%, Quality: ★★★★☆

(af) CDVS_BP-4, TPR: 28.47%, Quality: ★★★☆☆

Figure 2.7: Heatmaps computed by tested methods. True Positive Rates (TPRs) and Qualities are reported.

Chapter 3

The Social Picture (TSP)

3.1 Introduction

Social networks have become increasingly useful to understand people opinion and trends. Particularly, social media have changed the communication paradigm of people sharing multimedia data: users express emotions and share experiences in social networks. In social events (e.g., parties, concerts, sport matches) users are gradually changing in the so called *prosumers*, as they do not just *use as consumers* but also *produces* and share multimedia data related to what has captured their interest with mobile devices. The redundancy in these data, together with annexed metadata (e.g., geolocation, tags, mood-tag), can be exploited to infer social information about the attitude of the audience. For instance, systems such as MoViMash [73], ViComp [74] and RECfusion [7, 8, 9] are able to generate a video which describes the crowd interest starting from a set of videos by considering scene content popularity (Chapter 5). Indeed, the popularity of a visual content is an important cue for understand the mood of crowd attending to an event or estimate *how much* parts of a cultural heritage are perceived as interesting. Large scale visual data from social media and other multimedia information gathered by multiple sources (e.g., mobile devices) can be processed with Machine Learning and Computer Vision algorithms in order to infer knowledge about social contexts [75] or organize images by visual content.

In [7], we presented our framework called The Social Picture (TSP), in which images are gathered from social networks or uploaded directly in the repository by users through a mobile app and a website. The framework is capable of collect, analyze and organize huge flows of visual data, and to allow users the navigation of image collections generated by the community. In TSP three categories of image collections are distinguished: social events, private events and cultural heritage landmarks. The collections are processed with several tools which include automatic clustering of images, intensity heatmaps and automatic image captioning. These tools allow TSP to provide users a number of representative image prototypes related to each stored collection, exploitable for different purposes (e.g., selection of the most meaningful pictures of a painting during a showcase in a museum). Automatic clustering is implemented using a Convolutional Neural Network (CNN) representation [76] and employing an AlexNet architecture [77]. For each image, the *fc7* features are extracted and the t-SNE algorithm [78] is employed to compute a 2D embedding representation characterizing the pairwise distances between visual features. The intensity heatmap is another tool implemented in the framework. It consists of a map of values related to the number of collected pictures containing visual areas similar to the ones of a specific landmark building or area of interest. Users can interact with the heatmap selecting points on the map and retrieving images that contribuited to generate intensity values in that specific point (Chapter 2). Finally, the automatic image captioning [79] is a tool used to create and suggest descriptions of images, that comes useful for text-based queries perfomed by users.

In this Chapter, we employed VisualSFM (VSFM) [4] to compute visual matching between images within a cultural heritage collection of TSP and to obtain a 3D sparse reconstruction of landmarks. Using VSFM, we define new features added to TSP and present two advanced Image Analysis applications. In the first one, we consider the cameras as nodes in a fully connected graph in which the edges weights are equal to the number of matches between cameras. The spanning tree of this graph is used to explore images in a meaningful way, obtaining a scene summarization. In the second application, we define three kinds of density maps with relation to image features: density map, weighted-density map and social-weighted-density map. Results of a test conducted on a collection from TSP is shown.

The contents of this Chapter are based on our published papers [3, 6]. The structure of the Chapter is the following: overview of VSFM and its integration in our framework is described in Section 3.2, together with the definition of the Model (the data structure) employed for the process of advanced Image Analysis applications implementation. Then, we describe a scene summarization method and density maps in Sections 3.3 and 3.4, respectively. Discussion and conclusions end the Chapter in Section 3.5.

3.2 VSFM integration and Model definition

VisualSFM (VSFM) [4] is a powerful tool of 3D reconstruction from a set of photos, exploiting Structure from Motion (SfM), publicly available online [5]. VSFM maintains high accuracy by regularly re-triangulating feature matches that initially fail to triangulate. VSFM performs a linear-time incremental SfM method [80], hence it fits our scenario in which we want to estimate the 3D reconstruction of images within a collection, where new images could be added in any moment by users.

In this Chapter, we used 2924 photos from the cultural heritage collection named *Pisa* that was already been used as study-case in [3] and Chapter 2 for Heatmap computation based on image retrieval and image matching.

Once that images are loaded into VSFM, it extracts the visual features (SIFT [61] and GIST [81]) from them. Then, VSFM matches the extracted visual features and computes 3D sparse reconstruction. VSFM builds more than one single 3D-reconstructed model, as is possible that different scenes exist within a single dataset of photos or VSFM is not able to associate a set of coherent photos to the same model (due to too much different points of view). Among the built models, we chose the model with the highest number of cameras, since it is the most complete 3D reconstruction. VSFM saves this result in a N-View Match (NVM) file. Then, we designed a proper parser, in order to read the NVM file. The obtained data structure has been enriched with other meta-data in our framework (e.g., GPS tag, focal length). It represents our *Model* in the implementation process of advanced Image Analysis applications (Figure 3.1).

The NVM file has an own template [5]: for each model, all cameras (images) and 3D points in the reconstruction are listed. For each point, there is a full list of all the cameras that "view" that point. In other words, for each point, NVM file stores information about which camera uses that point as visual feature for matching in the SfM procedure. Instead, given a camera, the relationship about what are the features used by that camera is implicit. In order to make this latter relationship

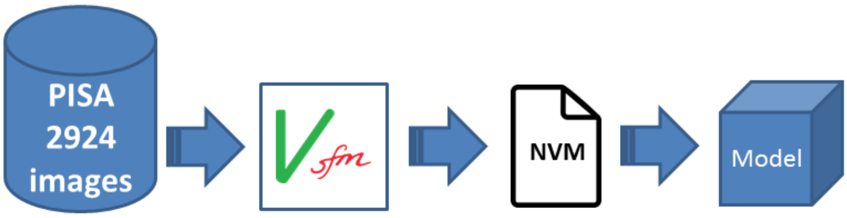


Figure 3.1: VSFM integration initial workflow: images of a collection are processed with VSFM and a *Model* is obtained through parsing of the NVM file and aggregation from our framework of meta-data.

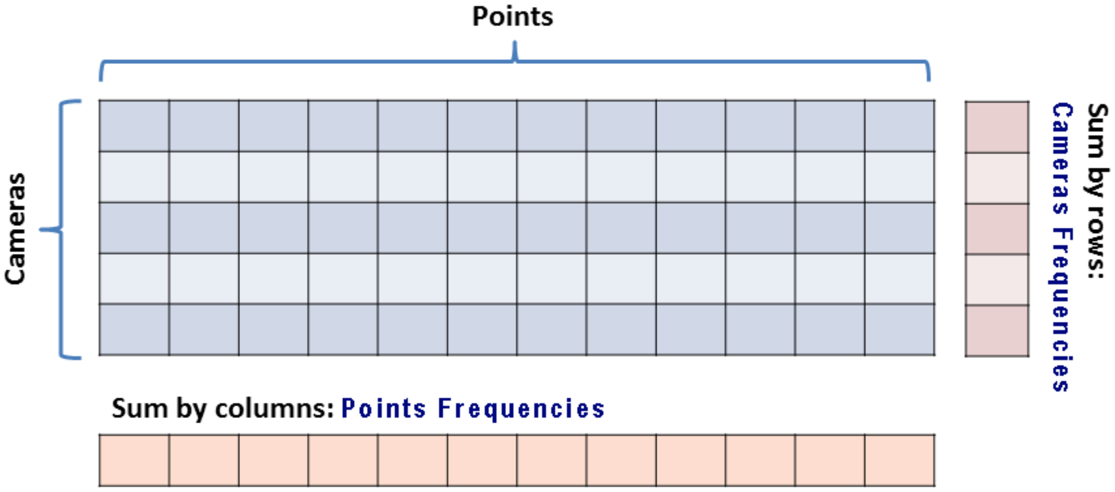

Points
Cameras
Sum by rows: Cameras Frequencies
Sum by columns: Points Frequencies

Figure 3.2: $Camera-Point-Correspondences$ (CPC) matrix and row-wise and column-wise sums. CPC is of boolean type.

explicit, we define a $C \times P$ matrix, where C is the number of cameras and P is the number of points. This matrix is named *Camera-Point-Correspondences* (CPC). The CPC matrix is of boolean type: it has a value 1 in position (c, p) if and only if the camera c used the point p as visual feature for matching. The row-wise sum of CPC gives as result the total number of points that each camera view, while the column-wise sum gives as result how many times each point has been viewed by cameras (Figure 3.2). Both of these sums can by normalized w.r.t. the maximum row-wise and column-wise. The normalization is used to obtain the "frequencies of view some point" for cameras and the "frequencies of been viewed by a camera" for points. These values can be used in the 3D reconstruction computed by VSFM, in place of the *color vertex* of 3D points. Hence, they can be viewed in a 3D-points

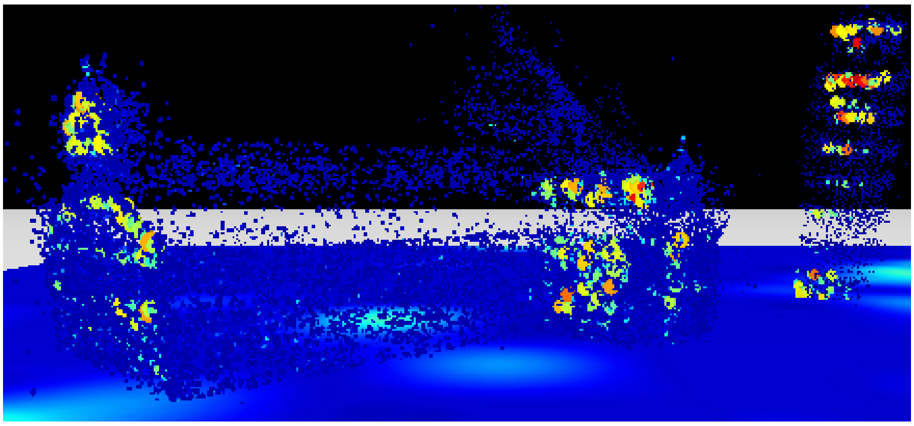


Figure 3.3: A detail of 3D-points frequency representation on The Social Picture web platform: the points with the highest feature-frequency are represented with bigger shape and red colors, while on the countrary points with the lowest feature-frequency are represented with smaller shape and blue colors.

frequency representation (Figure 3.3).

We also defined a matrix that represents relationships between cameras used in the 3D sparse reconstruction. This second matrix is named *Camera-Camera-Matches* (CCM). For each row c in CPC we compute the sum of the logical AND between c and the other cameras $\bar{c}$. The so obtained values are stored in CCM, in the proper cell in position (c,$\bar{c}$). In this way, the matrix CCM is symmetric w.r.t. the principal diagonal, which is null since is meaningless to consider self-matches (Figure 3.4). The other values in CCM are non-negative. For values equal to 0 we can assume that the corresponding cameras do not match at all.

The computation of the CPC and CCM matrices is known as *scene summarization* of the dataset [82].

3.3 Scene Summarization

Scene summarization is stated as "the issue of select a set of canonical images that represents the visual content of a scene". In [83, 82] a graph-based approach for scene summarization is presented, where the graph is built matching visual features between images. We follow a similar method, employing the Model defined in Section 3.2, obtained from the VSFM output enriched with meta-data from our framework.

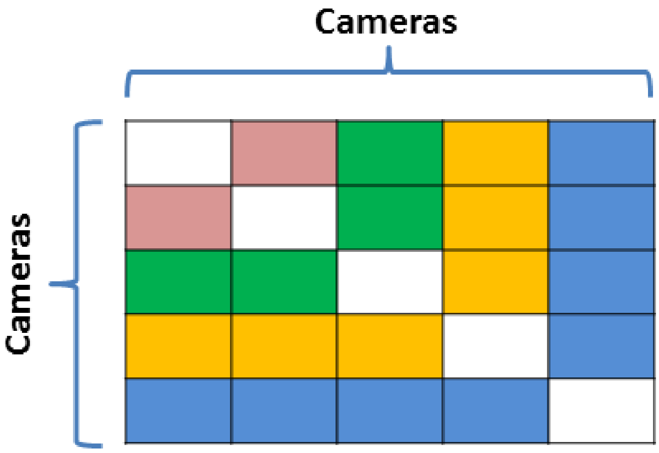


Figure 3.4: *Camera-Camera-Matches* (CCM) matrix. This symmetric matrix is obtained from the CPC boolean matrix as sum of the logical AND between each camera (represented as a row in CPC) and the other ones.

Values in CCM matrix can be seen as the edge-weights of an undirected graph. Indeed, cameras can be considered as nodes in a fully connected undirected graph in which the edge-weights are equal to the number of matches between the cameras. The spanning tree of this graph can be used to explore in a meaningful way the images, in a sort of tour of the selected collection. We employed the known minimum-spanning-tree (MST) construction algorithm from Kruskal [84], as it is particularly performing approachin terms of complexity: $O(V \cdot E)$ in the worst case, where V are the nodes and E are the edges of the graph.

The values stored in the CCM matrix represent how much cameras (images) are similar to each other. For this reason, we considered the values in CCM matrix as cost. However, since the Kruskal algorithm sorts edges in increasing order of cost, we modified the original values of CCM as follows. Firstly, the reciprocal of each value in CCM is computed, in order to convert maximum values in minimum, and viceversa. Then, we set to 1 (the highest possible value in CCM) the values equal to 0, as it is not possible to define the reciprocal for values equal to 0. In this way, all the new values are normalized in the range $(0; 1)$, accordingly to the non-negative costs requirement of the algorithm of Kruskal.

A *score* is computed for each node (camera) in the MST. We defined this score as the sum of the edge-weights adjacent to each node. It can be computed for each camera as the row-wise sum of CCM (or equally, column-wise, as CCM is

Figure 3.5: Scene Summarization: two consecutive frames of Depth-First-Search graph traversal of the Minimum-Spanning-Tree of Pisa Collection.

symmetric). Then, we chose the node with the highest score as the source node of the graph traversal, implemented with a Depth First Search (DFS). This choice is taken because very similar images are in the same path of the MST. When a node has more than one adjacent node (e.g., the node is root of a subtree in the MST), then it can be considered as the centroid of images contained in its subtree [83].

An example of DFS graph traversal on MST of the selected collection has been publiced online on [85]. Two consecutive images of the traversal are shown in Figure 3.5. This kind of scene summarization through a video is hardly evaluable with an objective metric. Instead, it is usually evaluated with a subjective consensus [83, 82]. We are still in the process of investigate a better evaluation method. More in details, we state that scores of the nodes may be exploited to discard the most unrelated images during the traversal. However, to assess a valid threshold for score values a wider experimentation should be conducted, with an higher number of collections. We demand this investigation for future works.

3.4 Feature Density Maps

Exploiting CPC matrix we can select all the points used as features by a given camera. Through the $Model$, which is parsed from NVM file, we also know the position of the features in the image. Given the positions of the visual features, we can define several kinds of maps (Figure 3.6):

- **Density Map (*D-map*)**: characterizes the spatial density of the inlier visual features of an image; the inlier visual features are the one used for matching

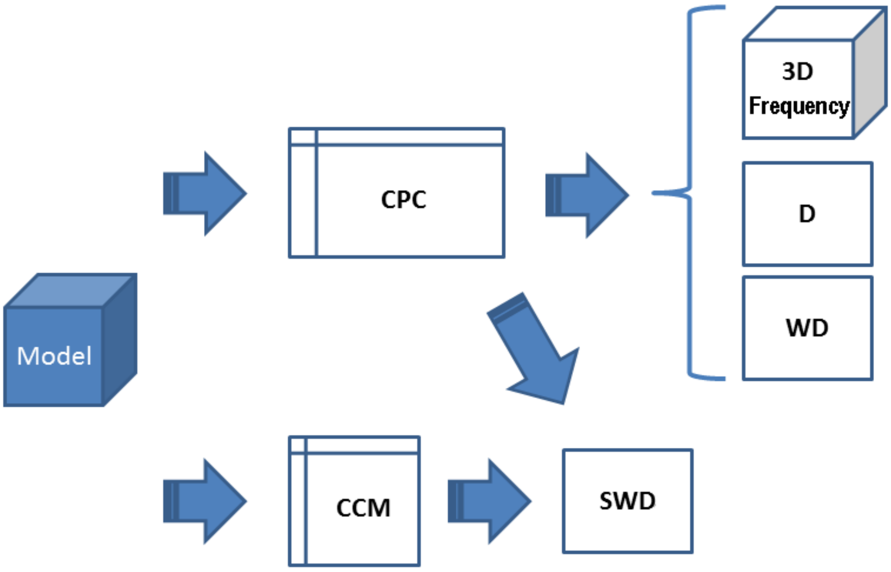
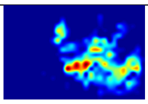
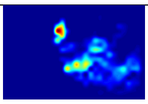
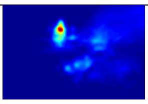
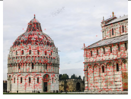
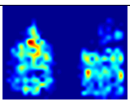
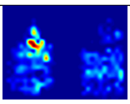
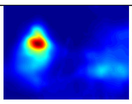
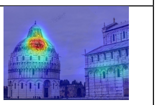
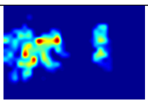
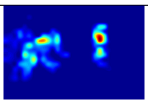
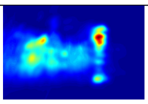
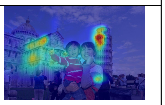
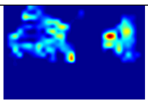
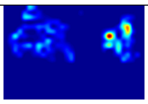
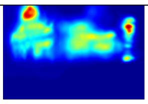
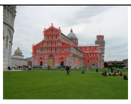
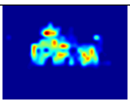
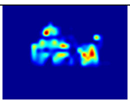
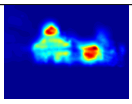
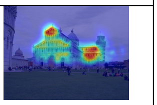
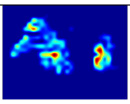
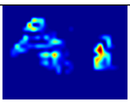
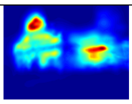
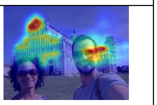


Figure 3.6: Creation of Feature Density Maps from the *Model*.

(Section 2.2.1 - Inliers extraction) and 3D sparse reconstruction. In other words, *D-map* highlights locations of the image in which visual features are more dense.

- **Weighted-Density Map (*WD-map*)**: characterizes the spatial density of the inlier visual features of an image, weighted w.r.t. their feature-frequency (as computed from *CPC* matrix). In other words, *WD-map* characterizes the density and importance of visual features within an image.
- **Social-Weighted-Density Map (*SWD-map*)**: similar to the *WD-map*, but it also take into account the inlier visual features from images matching procedure ((Section 2.2.1). In other words, *SWD-map* characterizes the density and "social" importance of visual features within an image, since the feature-frequency values are obtained taking into account also the images acquired by different points of view, which might have *WD-maps* potentially different from the one of the reference image.
- **Cumulative Map (*C-map*)**: similar to the *SWD-map*, but the *C-map* does not characterize density of robust visual features. The *C-map* highlights parts of the reference images frequently viewed in similar images in the dataset,

Table 3.1: Examples of Density Maps from Pisa Collection. From left to right, ID in the *Pisa* collection (ID), original image (Image), visual features marked on the original image (Features), Density-Map (D-Map), Weighted-Density Map (WD-Map), Social-Weighted-Density Map (SWD-Map) and a blended image between the original one and its SWD-Map are shown.

ID	Image	Features	D-Map	WD-Map	SWD-Map	Blended
1						
16						
45						
69						
74						
87						

computing the perspective transformations between them. So, the *C-maps* gives a smoother idea of what is socially salient from the point of view of the crowd. The intensity heatmap described in Chapter 2 can be considered a C-map.

Some example of density maps is shown in Table 3.1. A deeper description of each map is given in the following subsections.

3.4.1 Density Map (*D-map*)

The density of visual features is obtained quantizing the 2D space of the image and counting how many features are contained in each quantized interval. Values obtained are normalized w.r.t. the maximum obtained. This map is named *Density Map* (*D-map*). Note that visual features contained in the NVM file (and so in the *Model*) are all "inlier", this means that they are the one used for matching and

3D sparse reconstruction. They are a subset of the whole possible visual features, already filtered by VSFM. The majority of the inlier visual features can be usually found near the *keypoints* particularly robust to scale and rotation variations (Section 2.2.1 - Keypoints detection).

3.4.2 Weighted-Density Map (*WD-map*)

The *D-maps* can be further refined taking into account a weight for each visual feature. The weight is equal to the feature-frequency, as computed through the column-wise sum from CPC matrix. In this way, each feature contributes differently to the computation of the density map during the quantization of the 2D space of the image. Values obtained are normalized w.r.t. the maximum obtained in the whole map. The so obtained map is named *Weighted-Density Map* (*WD-map*). The *D-maps* characterize where *the majority* of the visual features can be found, while the *WD-maps* characterize where *the most important and robust* visual features can be found. Indeed, it is possible that a low number of very salient features generates a value in a WD-map higher than a bunch of low salient features. Given a reference image, the position of visual features in the *WD-map* is the same as in the *W-map*, since we just changed the weight of the features. What we want to highlight in the *WD-maps* is that in this kind of maps we are stressing the importance of the features from the point of view of the whole dataset, not just of a single image.

3.4.3 Social-Weighted-Density Map (*SWD-map*)

It is easy to find what images match with a reference one exploiting CCM matrix. We select the row in CCM related to the reference image. Then, we look for images with a number of matching features higher than 0. We can either set a threshold T_i on the number of matching features (Section 2.3.1 - Tolerance on Inliers number), in order to filter the image matching, obtaining the most reliable ones that reduce the risk of erroneous transform estimations. In the experiment described in this Chapter, we set $T_i = 30$. For all the images that have more than T_i features in common, we estimate the perspective transformation between them. A T transform matrix is obtained for each transform estimation. In *SWD-map* computation, we do not directly transform the images, but their *WD-maps*, without the normalization w.r.t. the maximum described in Section 3.4.2. All the non-normalized *WD-maps*, transformed with their own T, are summed to the non-normalized *WD-map* of the reference image. Eventually, this final result is normalized w.r.t. the maximum obtained by the sum of all the weighted feature densities. This map is named *Social-Weighted-Density Map* (*SWD-map*), since it is generated by a "social knowledge" that is able to decide what features of an image are meaningful.

The rationale behind the computation of *SWD-maps* is that images similar to a reference one might contain information that is not present in the reference image. For instance, similar images might contain objects that in the reference image is hidden due to occlusion or a bad point of view. Through *SWD-maps* some parts of the reference image might be judged as meaningful by a pool of similar images and each one of them independently define what is meaningful for itself. A pro of *SWD-maps* is the capability to highlight salient-features regions even if there is some occlusion in the image (Images 45, 69 and 87 in Table 3.1), while a con is the issue related with erroneous transformations, that might generate wrong density estimation in the final *SWD-map*.

3.4.4 Cumulative Map (*C-map*)

The *Cumulative Maps* (*C-maps*) are similar to the *SWD-map*: given a reference image, we use the visual features computed by VSFM to estimate a perspective transform T between the reference image and all its matching images. However, differently from the *SWD-maps*, in the *C-maps* we transform the matching images using a mask, properly a matrix with only values 1. All the masks, transformed with their own T, are summed to the mask of the reference image. This final result is normalized w.r.t. the maximum summed value obtained. Differently from the *SWD-maps*, the *C-maps* do not characterize density of robust visual features, but they highlight parts of the reference images frequently viewed in similar images in the dataset. The intensity heatmap described in Chapter 2 can be considered a C-map, where updates are performed with weights equal to 0. Visual features are often found near *keypoints*, but salient objects are not necessarily made of keypoints. Moreover, users are not interested in visual features when they acquire a photo. So, the *C-maps* gives a smoother idea of what is socially salient from the point of view of the crowd.

3.5 Conclusion

In this Chapter we described our framework The Social Picture (TSP). VisualSFM (VSFM) [4] has been employed to compute visual matching between images within a cultural heritage collection of TSP and to obtain a 3D sparse reconstruction of landmarks. Using VSFM, we defined new features added to TSP, such as the 3D-points frequency, and presented two advanced Image Analysis applications: scene summarization and density-maps.

Through the scene summarization we were able to create a video with a set of canonical images representing the visual content of a selected collection. This kind of summarizing-video is hardly evaluable with an objective metric. Instead, it is usually evaluated with a subjective consensus [83, 82]. We defined a score for each image in the scene and stated how it may be exploited to discard the most unrelated images during the traversal. We demanded the assessment of a valid threshold for score values to a wider future experimentations.

We shown how density-maps can be used together with the Structure-from-Motion (SfM) technique to highlight parts of the image with robust visual features. Several types of density-maps have been defined with different aims. Eventually, we shown that Social-Weigthed-Density (SWD) maps represents a good tool to stress the presence of visual features even when a strong occlusion is present in the image. Currently, SWD-maps shown in this chapter have not been publicly released yet, but they could be shared under request.

Chapter 4

Social Saliency

4.1 Introduction

In Chapter 3 we presented our framework The Social Picture (TSP), in which images are gathered from social networks or uploaded directly in the repository by users through a mobile app and a website. In Chapter 2, we defined one of the tool implemented in TSP: the heatmap. It is employed to represent, during the Content Based Image Retrieval (CBIR) procedure, areas of reference image with higher/lower probability of be matched with other query images. Hence, TSP is the place in which images are gathered and stored, while the heatmap is one of the tools included in the framework, designed for visualize the results of CBIR analysis and to summarize the contents of a collection, at a glance. Then, in Chapter 3, we also shown how the use of advanced techniques like Structure from Motion (SfM) can open the way to new kind of data-analysis, starting from images collected in TSP. In Section 3.4, we defined several types of feature density maps, which represent other possible summarizations of a collection. In particular, we shown how Cumulative Maps (*C-maps* - Section 3.4.4) are nothing more than another definition of heatmap. Finally, we defined the most complex feature density map: the Social-Weighted-Density Map (*SWD-maps* - Section 3.4.3).

SWD-maps are a very interesting summarization tool, as they allow us to see the data from a new view-point: the *social* one. In other words, taking into account the social aspect, we become aware of the social context that links them together. SWD-maps of images acquired in the same place highlight parts of the environment with visual features perceived as important, properly *salient*, by the crowd. Indeed, this is exactly the definition of saliency: "spatial regions in the visual field that attract attention" [14].

For this reason, we studied the literature related to saliency estimation and we found that, at the best of our knowledge, *social saliency*, intended as a saliency estimated from a social point-of-view, represents a novel definition of saliency. Although there are similar concept of saliency (i.e., co-saliency [86, 87], multi-camera saliency [88], likelihood of joint attention [89, 90]), we focused on an estimation method based on the analysis of images gathered from social networks or uploaded directly in the repository by users, in a social way, from the point of view of the crowd. Hence, in this Chapter, we define Social Saliency and investigate if it is possible to learn a saliency model from collections in TSP, in order to estimate the social saliency of a new image subsequently added in the framework.

We tested two methods to learn saliency models: one is based on Support Vector Regression (SVR [91]) and the other one on a Counting Convolutional Neural Network (C-CNN) named *Hydra C-CNN* [92]. More in details, we compared results of linear and non-linear SVR with the ones obtained by Hydra.

The structure of the Chapter is the following: in Section 4.2 we discuss related studies about saliency estimation. Social Saliency estimation method is described in Section 4.3. Discussion and conclusions end the Chapter in Section 4.4.

4.2 Related Works

4.2.1 Saliency Definition

In Section 4.1, we reported the definition of *saliency* as given by Duncan [14]. Other similar definitions are given in literature, i.e., "*saliency* characterizes some parts of a scene (which could be objects or regions) that appear to an observer to stand out relative to their neighboring parts" [13] or "*saliency* is the ability of a vision system (human or machine) to select a certain subset of visual information for further processing. This mechanism serves as a filter to select only the interesting information related to current behaviors or tasks to be processed while ignoring irrelevant information" [93]. Saliency raises from visual attention models related to biological reactions [12]. The first model of visual saliency has been defined by Itti, in the 1998 [94]. The human attentional phenomenon is composed by two main stages: pre-attentive and attentive. In the former, eyes focus for 25-50ms on each item within the field-of-view with the aim of select locations sufficiently distinctive [95]. In the latter, features are combined into high level entities like objects [96]. Pre-attentive and attentive stages are usually referred as bottom-up and top-down processes [95, 14, 13]. They are really different: in the bottom-up approach, features automatically raise from the relationships with surrounding (e.g., objects on the foreground appear highly different from the background due to color, contrast, sharpness, etc.), whilst in the top-down approach features are influenced by elements like context, expectations of the user, cultural and social biases, prior knowledge of the scene or a task given to the watching user (e.g., object-search) [95, 14, 13]. Given this distinction, saliency methods are usually subdivided in three categories: bottom-up, top-down and mixed (integrated) methods. Each one of these categories is modeled separately as an independent task [14].

Saliency methods can be subdivided also in 3 other complementary levels, w.r.t. what saliency detection highlights [14, 13, 97]: Visual Attention Model (VAM), Saliency Object Detection (SOD), and Saliency Object Segmentation (SOS). This further classification is strictly related to the task. More in details, VAM refers to eye fixations prediction, SOD refers to the ability of detect salient objects in the scene highlighting them with a bounding box, and SOS refers to the ability of detect and also segment salient objects in the scene [97].

4.2.2 Saliency Methods

Among the many different saliency methods proposed through the last decades [13, 14], we refer the followings [97]:

- Visual Attention Model (VAM):

  – Itti et al. [94], based on image input. Itti is one of the oldest and major tested saliency method. Frintrop et al. [98] proposed several improvements to Itti.

  – Hou et al. [99], based on information redundancy and spectral contents in the frequency domain.

  – Harel et al. [100], based on segmentation and a graph structure.

  – Garcia et al. [101], based on adaptive whitening of color and scale features.

- Saliency Object Detection (SOD):

  – Hou et al. [102], based on information redundancy and spectral contents in the frequency domain.

  – Goferman et al. [103], based on content awareness.

- Saliency Object Segmentation (SOS):

  – Achanta et al. [104], based on frequency domain tuning.

  – Jiang et al. [105], based on context and segmentation.

Advanced methods of saliency estimation combine features from background and foreground together [106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117]. It is possible to find methods based on color data-driven approaches, edges extraction and frequency domain [106, 113, 118]. Other works are based on segmentation techniques, which are employed to segment salient regions of the image or to separate background and foreground layers [108, 113, 117, 119]. We can also distinguish methods based on Convolutional Neural Networks (CNNs) and deep learning (often named Deep Neural Networks - DNNs) [92, 107, 110, 112, 114, 116, 120]. In these methods based on neural networks, there is a step named *training*, in which the saliency model is *learned* from datasets of images where ground-truth saliency maps are available. For instance, in [114] the saliency maps are computed as a combination of the ones obtained through basic methods (e.g., Itti [94], Harel [100], Garcia [101]), while the DNN employs a Support Vector Machine (SVM) for regression (properly, a SVR [91]). The main saliency estimation methods currently available have been benchmarked by the Massachussets Institute of Technology (MIT). An updated list of tested methods is available in [121]. Indeed, authors are used to submit their methods to [121], in order to be listed in the public benchmark obtaining an objective quality assessment of their proposed methods.

Saliency may be employed in different scenarios. For instance, it is used in context aware image resizing [94, 122, 123], visual object tracking [124], image segmentation [125], or features and keypoints extraction [126].

4.2.3 Saliency Datasets

"Different tasks" means also "different datasets", which are needed to properly asses the methods. Among many, we refer the followings [97]:

- MIT Eye Fixations Dataset [127], aimed for VAM methods, it consists of $1,003$ images hand labeled with eye fixations data gathered by 15 subjects;
- Saliency in Context (SALICON) [128], aimed for VAM methods, it consists of $10,000$ images hand labeled with eye fixations data gathered by 16 subjects;
- Microsoft Research ASIA Dataset (MSRA) [129], aimed for SOD methods, it consists of $25,000$ images hand labeled with bounding boxes annoations;
- THUS10000 Dataset [130], aimed for SOS methods, it consists of $10,000$ images (derived from MSRA) hand labeled with segmentation masks around objects;
- DUT-ORMON Dataset [131], generic purpose dataset, it consists of $5,168$ images hand labeled with ground-truth masks useful for any one of the three methods.

A very well updated, organized, and detailed list of datasets for saliency benchmarking is available in [121].

4.2.4 Saliency Benchmarks and Comparative Works

We described how saliency methods and datasets are strictly related to specific tasks (i.e., eye fixation estimation in VAM methods, object detection in SOD methods, search of specific objects). This statement brought to an objective problem of comparison between this plethora of methods. In literature, there are many comparative works [97, 114, 120, 132, 133]. As said before in this Section, a very good benchmark is frequently kept update and publicly available in the website of the Massachusetts Institute of Technology (MIT) [121].

Comparative results are also based on a number of different metrics, which includes Receiver Operating Characteristic (ROC) curves [13], Area Under Curve (AUC) [13] or the Pearson Correlation Coefficient (COR or CC) [134]. One of the most promising objective metrics is the Normalized Scanpath Saliency (NSS) [133]. Particularly, NSS is claimed to be a very promising metric that could become the main one in the MIT benchmark [121].

4.2.5 Social Saliency Related Works

As in Section 4.1, we gave a novel definition of saliency, namely *Social Saliency*, even if similar (but not equal) definitions exist. All of these kinds of saliency models are based on a combination of several inputs from different sources: they could be computed analyzing frames in a video stream, common context in a bunch of images or the likelihood of joint-attention.

In the video context, Zhang et al. [108] implemented real-time object detection in video streams with 80 frames per second, while Mauthner et al. [135] proposed an encoding method to approximate the joint distribution of feature channels in video frames (color or motion).

Luo et al. [88] gave the definition of *multi-camera saliency*, employing a Label Consistent dictionary-learning algorithm based on K-Singular Value Decomposition [136] (LC-KSVD - It is similar to a k-means clustering method). In [88] a *learning phase* is needed, in order to discriminate the saliency between different point-of-views.

Zhang et al. [86, 87] gave the definition of *co-saliency*. Also in this case, a *learning phase* is needed: SVM is employed to learn what are the regions that tend to have similar saliency values. The method described in [87] is based on Self-Paced Multi-Instance Learning (SP-MIL [137]): self-paces stands for a continuous process of *training-&-update*, while "multi-instance learning" means that the model is learned from a set of different images from the same or similar context.

Saliency defined as *likelihood of join-attention* is probably the most similar definition to the social saliency one. Indeed, in [90] they named the joint-attention exactly with the term social saliency. However, in [90] saliency values are generated and thought differently from our definition of social saliency. In [90] users (who take a picture) are represented as little atoms or particles capable of generate a sort of magnetic field around them. In their method, is possible to define a "center of mass" for the population of users and also a "point of joint attention", in which the combination of magnetic fields from the users reach the global maximum. This combination could be seen as a 2D function, hence several local maximum points may even be defined, resulting in secondary points of joint attention. This 2D saliency map can be reprojected in the 3D world highlighting, with Augmented Reality techniques, parts of the environment with high likelihood of joint-attention. Instead, our idea of *social saliency* is something related to visual features assessed by elements of a scene depicted in lots of images gathered from a social media scenario. For us, social saliency is not a likelihood of joint-attention computed w.r.t. how many people are grouped in a specific place, as it is in [90].

Indeed, we tackled the saliency estimation issue as a visual-features counting problem. Onoro et al. [92] developed a Counting-CNN: its main task is to learn a model to automatically estimate how many entities (e.g., car, people) appear in a patch of an image. They started from another C-CNN designed by Zhang et al. [138] using an implementation based on the Caffe framework [139]. Then, they improved the initial C-CNN introducing a scale-aware architecture named Hydra, based on the previous work from Li et al. [110]. With this improvement, they are able to gain a better counting estimation even when the same number of entities are represented in patch with different scale factors. In our method, we modified the architecture in [92], in order to use images in our framework The Social Picture (TSP - Chapter 3) and Social-Weighted-Density Maps (*SWD-maps* - Section 3.4.3) during the training phase.

4.3 Social Saliency Estimation Method

We estimated the social saliency attention model employing collections of images in the framework The Social Picture (TSP - Chapter 3). Particularly, we integrated the tool named Visual Structure from Motion (VSFM [4]) and we started our saliency estimation from the model described in Section 3.2. We worked in Matlab environment (Fig. 4.1).

In our method, we want to learn a saliency model from the Social-Weighted-Density (SWD - Section 3.4.3). Once the model is learned, we can be able to use it for estimate social saliency in new images uploaded in TSP.

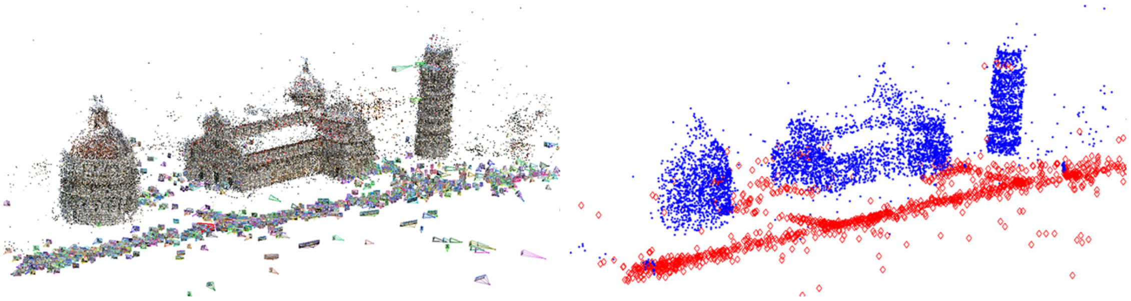

(a) VSFM (b) Matlab

Figure 4.1: Cloud points representation of the NVM Model computed through VSFM (Section 3.2). Views in the (a) VSFM GUI and (b) Matlab GUI.

We tested two methods to learn saliency models: one is based on Support Vector Regression (SVR [91]) and the other one on a Counting Convolutional Neural Network (C-CNN) named *Hydra C-CNN* [92]. We compared results of linear and non-linear SVR with the ones obtained by Hydra. Validation of the methods have been performed through k-fold cross-validation, with $k = 10$. We shown the results, as average of outcomes from cross-validation, in the form of Area Under Curve (AUC [13]) and Pearson Correlation Coefficient (COR or CC [134]).

4.3.1 Experimental Settings

Our dataset is a collection of 2924 photos from a cultural heritage collection in TSP, named *Pisa*, that was already been used as study-case in [3] and Chapter 2 for Heatmap computation based on image retrieval and image matching. The average size of the images is 1048×1260, with 3 color channels (RGB). SWD-maps have been computed for each image in this collection and are used as ground-truth saliency estimation. Notice that these SWD-maps are not-normalized: this means that values in the maps are counters from 0 to a maximum value, which may be different between maps. SVR and CNN methods have different experimental settings, detailed in the following.

Setting of Support Vector Regression

We extracted features from images employing a sliding window approach. Following the multiresolution pyramid approach, images have been rescaled with a 0.5 rescale factor. In this way, we reduced images size, decreased computational time, and retained only the visual features most invariant to scale transform. We set the window size (or block size) to 51 pixels and step to 50 pixels. Taking into account the rescale factor and the images average size (1048×1260), we extracted an average of 100 blocks from each image. Moreover, each block counts 3 color channels, hence each feature is made by $51^3 = 132,651$ values. We appended the coordinates of the window center (pivot), in order to capture information related to the position, reaching a total of $132,653$ values. The model that we are going to learn should be able to correlate these values with the single one from SWD-map placed in the pivot coordinates. In Machine Learning definitions, the block values represent our X features term, the SWD-map value in the pivot is Y and the learned model M is the function allowing to compute Y given X, properly $M(X) = Y$.

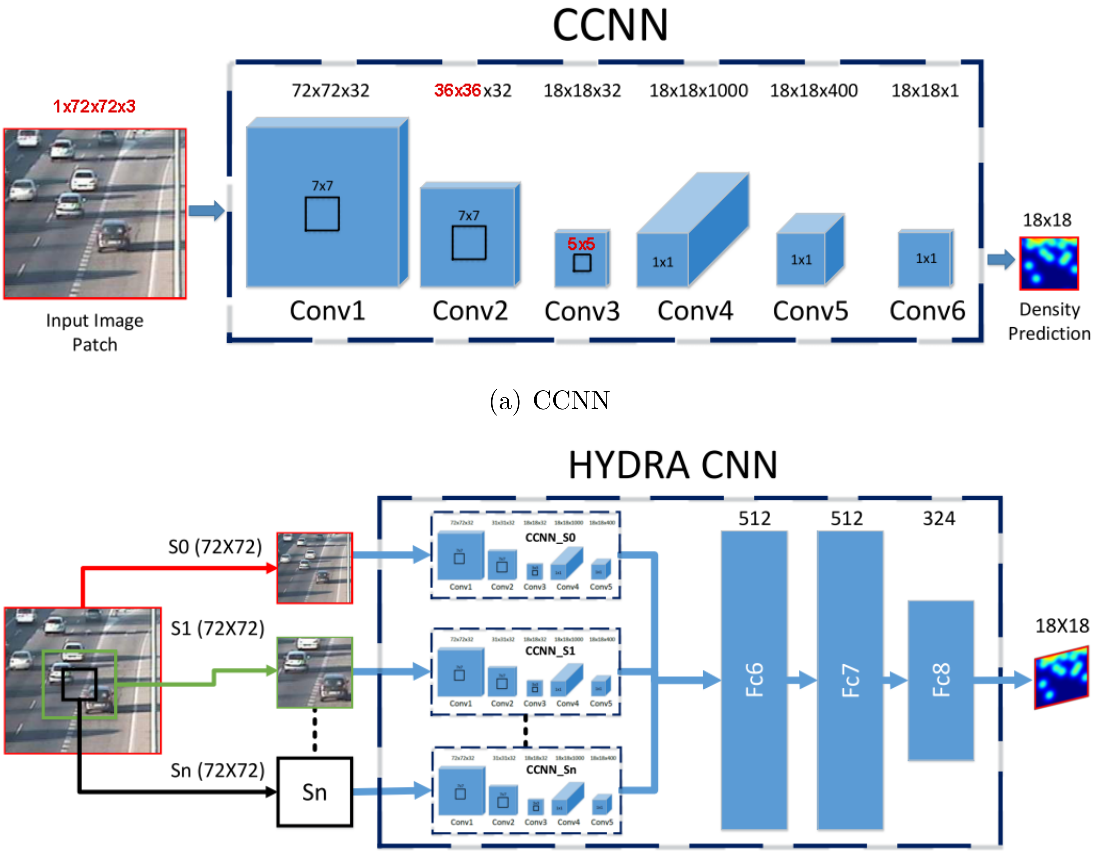
We trained our model in Matlab through Support Vector Regression with linear and non-linear kernels. For the linear approach, we employed the function *fitrlinear*, with the following parameters [140]:

$$fitrlinear(X, Y, Regularization, ridge, Solver, lbfgs)$$

where regularization equal to ridge is used for the computation of the objective function and the solver is the minimization technique (in this case it is equal to Limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) [141]). For the non-linear approach, we employed the function *fitrsvm* instead, with the following parameters [142]:

$$fitrsvm(X, Y, KernelFunction, gaussian, KernelScale, auto, Standardize, true)$$

where the kernel function is defined as a Gaussian (or Radial Basis Function - RBF) non-linear kernel, the kernel scale is set automatically by the procedure itself, and standardize equal to true means data are weighted-standardized. This step of weighted-standardization is manually done in the linear approach through subtraction of the mean and division per standard deviation.



(a) CCNN

(b) Hydra CCNN

Figure 4.2: Hydra architectures (images taken from [92]). Numbers written in red are correction of typos in [92].

Prediction is done using the function *predict*, with the following parameters [143]:

$$predict(M, X) \rightarrow Y$$

where M is the model learned (gained as output) with *fitrlinear* or *fitrsvm*.

Setting of Hydra C-CNN

Hydra can have several levels of complexity in the architecture [92]. The simplest is the C-CNN with the following order of convolutional (C), relu(R), and max-pool (P) layers (Fig. 4.2(a)): C1, R1, P1, C2, R2, P2, C3, R3, C4, R4, C5, R5, and C6.

CCNN takes patches of size 72 × 72 (with 3 color channels - RGB) as input, and in the final layer (C6) it computes a value representing the counting of salient elements in the patch. We employed a more complex architecture: a Hydra CCNN with 4 heads (Fig. 4.2(b)).

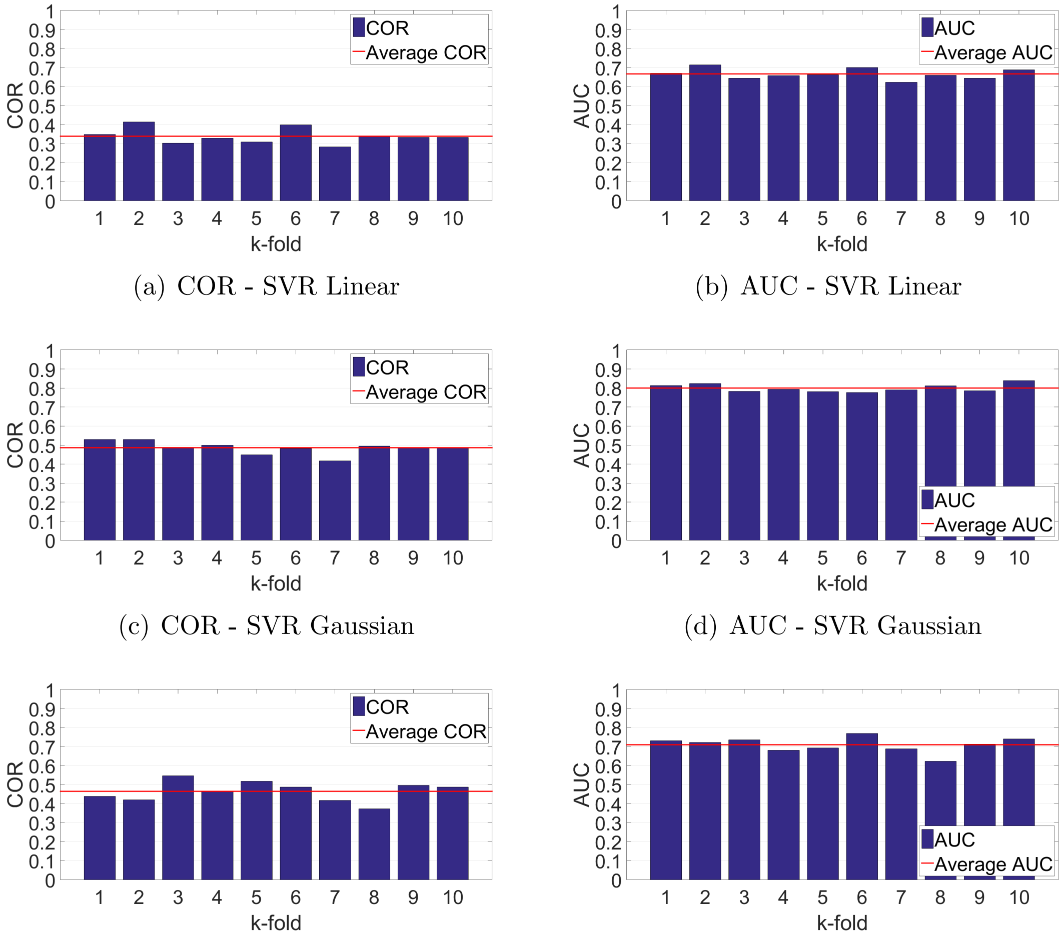
We empirically changed the configuration parameters used in [92], since our ground truth is different from the one used by Onoro et alii. They employed a ground truth computed as sum of Gaussian functions centered on each dot of a set of 2D manually annotated points (their aim is to solve the counting problem, not to estimate saliency). Although SWD-maps may be seen as the sum of Gaussian functions, their value ranges are totally different if compared to the ones employed in [92]. This is the reason why we adopted different configuration parameters. The number of randomly extracted patch has been decreased from 800 to 400, and the loss weight during *training phase* has been decreased from 1 to 0.01, as done for CCNN. These changes allowed a minor computational time and amount of required resources (i.e., memory and CPU).

We trained our model in Matlab through the code available at [144], properly modified in order to take in input SWD-maps.

4.3.2 Experimental Results

Experimental results are shown in Fig. 4.3, as average of outcomes through k-fold cross-validation with $k = 10$, in the form of Pearson Correlation Coefficient (COR or CC [134]) and Area Under Curve (AUC [13]). We compared results of linear (Fig.s 4.3(a) and 4.3(b)) and non-linear SVR (Fig.s 4.3(c) and 4.3(d)) with the ones reached by Hydra (Fig.s 4.3(e) and 4.3(f)). We obtained an average COR equal to 0.34 for linear SVR, equal to 0.49 for non-linear SVR, and equal to 0.45 for Hydra. Similarly, we obtained an average AUC equal to 0.67 for linear SVR, equal to 0.79 for non-linear SVR, and equal to 0.71 for Hydra.

COR values range from 0 to 1, that is from totally uncorrelated to totally correlated data. Hence, we gain models that are fair correlated. AUC values also range from 0 to 1, where the worst case is obtained when 0.5 values appear (this is the case of a random classifier). We gain fair values in this case, too. Code for metric evaluation is available at MIT saliency benchmark website [121]. However, our ground truth (the SWD-maps) are a different and new kind of saliency attention



(a) COR - SVR Linear

(b) AUC - SVR Linear

(c) COR - SVR Gaussian

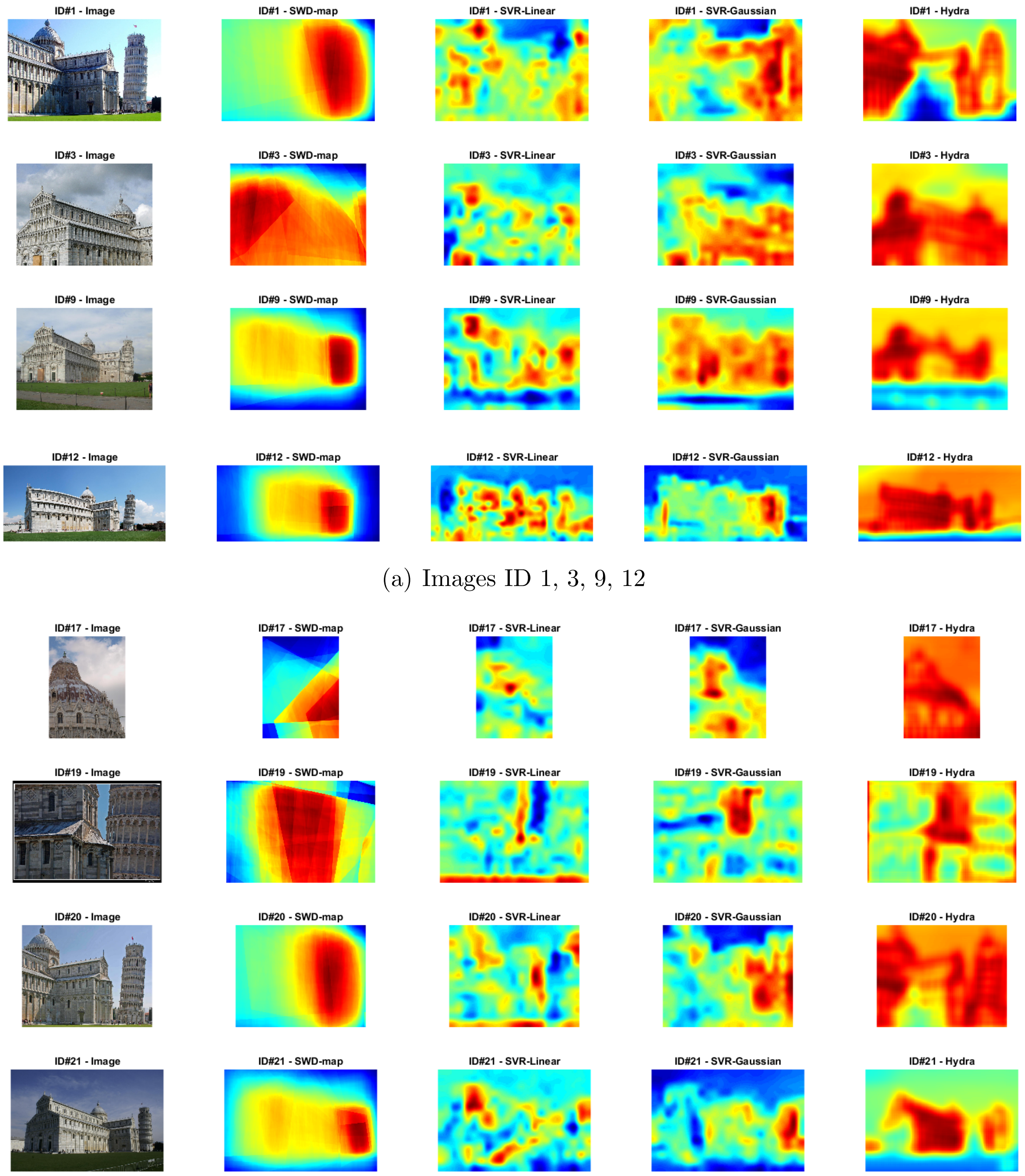
(d) AUC - SVR Gaussian

(e) COR - Hydra

(f) AUC - Hydra

Figure 4.3: Experimental results. COR is the Pearson Correlation Coefficient [134], while AUC is the Area Under Curve [13]. SVR Linear denotes the linear approach, while SVM Gaussian the non-linear one.

model, hence we needed to change MIT's publicly available Matlab code for metric evaluation. Although codes are different, we can at least take a look to COR and AUC values usually obtained on datasets in the MIT benchmark, in other saliency estimation contexts. From this comparison, we may say that obtained results sound reasonable. Clearly, more effort in standardization of our dataset and employment of experimentally meaningful tools, like the metric evaluation code of MIT, should be put in the future updates of the system. From this first results, we may say that this



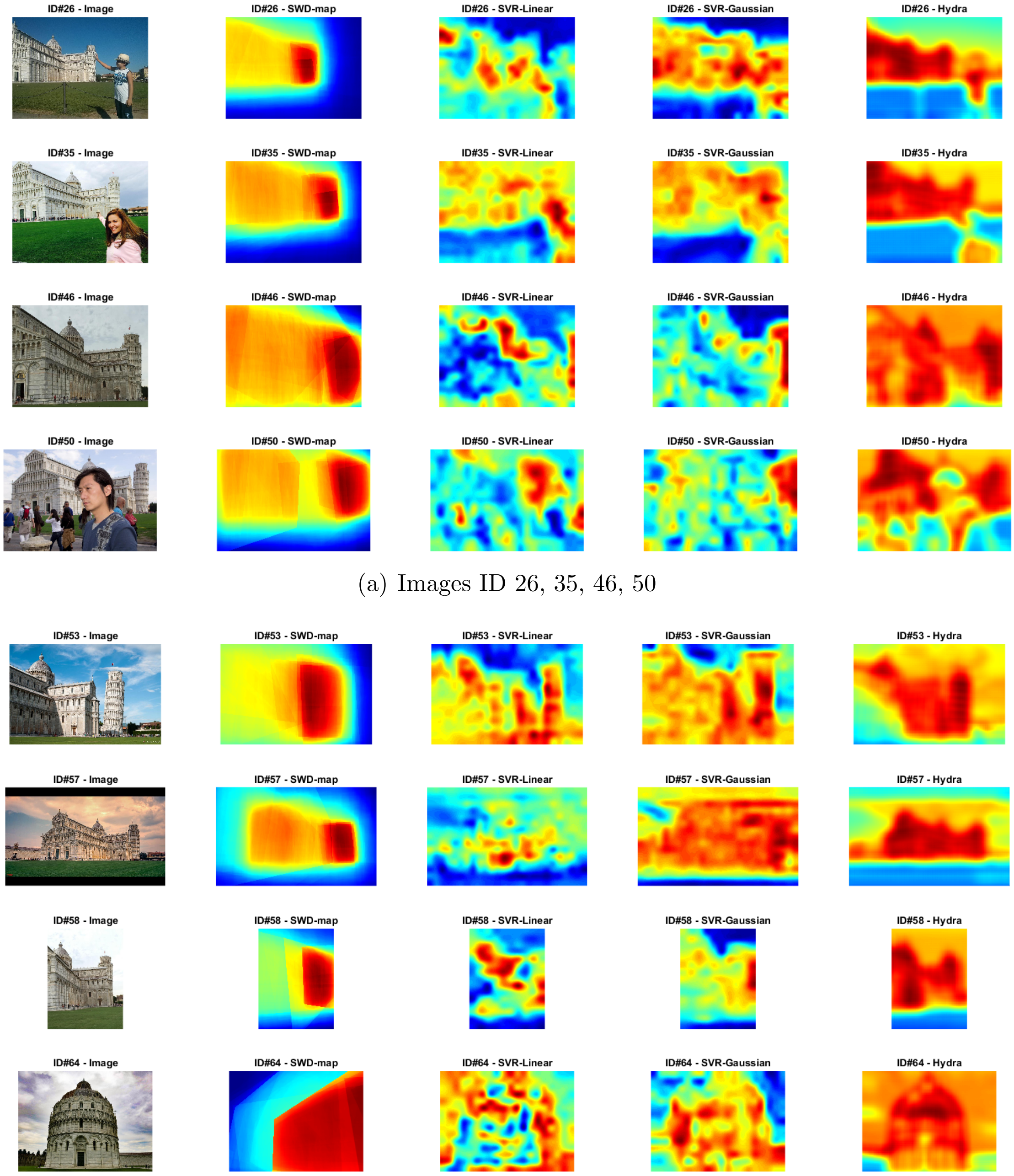
(a) Images ID 1, 3, 9, 12

(b) Images ID 17, 19, 20, 21

Figure 4.4: Visual comparative examples.

seed work on Social Saliency is promising and has room for further improvements.

In Fig.s 4.4 and 4.4 some visual comparison is shown. From this comparison it



(a) Images ID 26, 35, 46, 50

(b) Images ID 53, 57, 58, 64

Figure 4.5: Visual comparative examples.

is possible to see how sometimes SWD-maps are affected by *shift effects* in Query Expansion process (Section 2.3.2). For instance, in images 17, 19, and 64 during the image matching process the majority of the images have been matched always with the same region of the reference image. This behavior results in a shifted SWD-map.

Linear SVR model highlights salient regions in a very sparse and segmented way. This is clearly the worst obtained result. The Gaussian SVR also generates segmented maps, but they are more similar to the corresponding SWD-maps. Indeed, SVM non-linear model reaches the best COR and AUC values. Finally, the Hydra model generates salient maps very similar to Salient Object Segmentation methods 4.2. Although Hydra model reached lower COR and AUC values compared to non-linear SVR, saliency maps obtained with this method seems to be able to highlight salient parts better than the other two SVR models. However, at the current point we are not able to assess if Hydra model has really learned to distinguish architectural parts of the buildings. Indeed, it could have learned how to recognize the grass, considering salient everything but the lawn. A deeper experimentation is needed, taking into account also different environments and contexts.

4.4 Conclusion

In this Chapter, we defined Social Saliency and investigated if it is possible to learn a saliency model from collections in TSP, in order to estimate the social saliency of a new image subsequently added in the framework. We reported the literature related to saliency estimation and we found that, at the best of our knowledge, *social saliency*, intended as a saliency estimated from a social point-of-view, represented a novel definition of saliency. Although there are similar concept of saliency (i.e., co-saliency [86, 87], multi-camera saliency [88], likelihood of joint attention [89, 90]), we focused on an estimation method based on the analysis of images gathered from social networks or uploaded directly in the repository by users, in a social way, from the point of view of the crowd.

We tested two methods to learn saliency models: one is based on Support Vector Regression (SVR [91]) and the other one on a Counting Convolutional Neural Network (C-CNN) named *Hydra C-CNN* [92]. We compared results of linear and non-linear SVR with the ones obtained by Hydra, and we found that non-linear SVR obtained higher Pearson Correlation Coefficient (COR) and Area Under Curve (AUC) values than Hydra model. Linear SVR gained the lowest COR and AUC values, and its saliency maps are too sparse and segmented. The Gaussian SVR also generated segmented maps, but they are more similar to the corresponding ground truth SWD-maps. Although Hydra model reached lower COR and AUC values compared to non-linear SVR, saliency maps obtained with this method seems to be able to highlight salient parts better than the other two SVR models, and are very similar to Salient Object Segmentation (SOS) maps.

A deeper experimentation is needed, in order to increase the soundness of this novel method. More effort in standardization of our dataset and employment of experimentally meaningful tools, like the metric evaluation code of MIT, will be put in the future updates of the system. Hydra-based architecture is the most promising one. We are planning to acquire images collections with different environments and contexts, in order to allow the learning of a more generic visual attention model and reduce the overfitting coming from a single collection.

Part II - Videos

Chapter 5

Video Summarization

5.1 Introduction

During a social event, the audience typically uses its personal devices to record video clips related to the most interesting moments of the event. As a result, several videos will be related to the same visual contents, and this redundancy can be exploited to infer the most interesting moments of the event over time, according to the people interests on the observed scenes. Acquired scenes directly reflect the audience's response to the event. In this context the estimation of the most popular scene is of interest. The issue of crowd-popularity estimation through automatic video processing is not trivial due to the variability of the visual contents observed by multiple devices: different points of view, pose and scale of the objects, lighting conditions and occlusions. The differences between device models should be also taken into account, since they imply different characteristics of the lens, color filter arrays, resolution and so on. For instance, even using two devices with similar (or equal) sensors the colors recorded will not necessarily be the same because devices responses are processed with different non-linear transformations due to the differences on the Imaging Generation Pipelines (IGPs). They can vary from device to device and even on an per-image basis [145, 146, 147].

In this Chapter we describe a system called RECfusion, thought to estimate the popularity of scenes related to multiple video streams. Streams are analyzed with the aim to create a continuous video flow, obtained by mixing several input channels, taking into account the most popular scenes over time to reflect the interests of the crowd. Then, clusters of different scenes are tracked over time. This allows to have not only the most popular scene at each time, but also other scenes of interest and give the possibility to introduce a scenes story log that allows the user to select the scene of interest among all the detected ones. Firstly, the system automatically processes multiple video flows from different devices to understand the most popular scenes for a group of end-users. Then, the extracted information is used to cluster devices according to observed scenes over time. Initially in [8] a thresholding method was used for the cluster tracking procedure. We employed an algorithm based on a voting procedure, in order to be able to generalise over different sets of video, and compared its performance with respect to a threshold-based approach.

The contents of this Chapter are based on our published papers [7, 8, 9]. The structure of the Chapter is the following: in Section 5.2 we discuss related studies about crowd-saliency inference from multi-device videos. In Section 5.3 an overview of the RECfusion framework is given, together with the description of its three main modules: intraflow analysis, interflow analysis and cluster tracking. In Section 5.4 the employed dataset is introduced, whereas in Section 5.5 we report experimental settings and results. Discussion and conclusions end the Chapter in Section 5.6.

5.2 Related Works

Different papers about crowd-saliency inference from multi-device videos have been proposed in literature in the past. Works in [148, 149] exploit Structure from Motion (SfM) to estimate a 3D reconstruction of the scene and the pose of employed devices. Hoshen et al. [150] uses egocentric video streams considering a single camera model acquired by different participants to create a single popular video of an event. However, in [148, 149, 150] the number of different popular scenes and the number of devices are known a priori. This information about the amount of devices and popularity of scenes is used to estimate a number of regions of interest used as prototypes in the grouping phase. Saini et al. [73] developed the framework MoViMash with the purpose of replicate the behavior of a movie director: the system learns from a labeled sets of video frames “how to” and “when” perform transitions between different views. However, this technique is hardly adaptable for a real-time context, since for each different recorded scene a proper learning phase should be tuned up. ViComp is another framework similar to MoViMash [74]. In ViComp the final output video consists in a combination of several video streams from multiple sources. The combination is obtained by selecting high quality video segments according to their audio-visual ranking scores. It selects the best video stream among a pool of available ones exploiting BRISQUE to evaluate spatial and Luminance Projection Correlation (LPC) to evaluate spatio-temporal assesment. BRISQUE, which stands for Blind/Referenceless Image Spatial Quality Evaluator, quantifies degradation and noise caused by video compression [151] and estimates camera pan and tilt [152].

Video streams gathered in a crowdsourcing paradigm represents a key factor in the social media context. Cues about what is really salient from the point of view of the audience can be retrieved by analysing captured videos and representing the estimated mood of the crowd. The subject identified as the most salient could be further investigated to understand the causes that brought to such a focus. The achieved findings represent a powerful cue to make specific improvements to change the degree of appeal of the scene. Furthermore, multi-video semantic retrieval could be useful in many application scenarios different from crowd-popularity estimation.

The aforementioned approaches achieve significant results but, compared to them, our approach (RECfusion) does not need any prior knowledge or training stage and is able to combine videos from an unknown number and types of recording devices. RECfusion is a framework with a popularity-based video selection approach: it clusters the video streams and selects the best video stream from each cluster exploiting clustering metrics.

5.3 RECfusion system overview

RECfusion is a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Given a set of video streams as input, the framework can group these video streams according to the viewed similarity and popularity of the scenes over time, then it automatically suggests a video stream to be used as output, acting like a “virtual director”. With the aim to mitigate the aforementioned differences in the color representation of the devices, due their different IGPs, the video frames are pre-processed by an equalization algorithm. This step helps the further computations that compares frames captured by different devices [147, 153, 154, 155]. After this normalization, the system extracts an image representation from each frame. The algorithm takes a frame as input and returns a descriptor. The aim is to have a descriptor that maximize the differences between semantically different frames and minimize the differences between semantically similar ones. The color data of the frame are valuable cues for several algorithms, with applications in segmentation, image retrieval, object tracking and recognition. We reported in Appendix A a framework specifically developed to tackle Color Standardization issue for Archaeology. However, to be useful for practical issues, colors should refer directly to intrinsic properties of the objects and should be independent by acquiring device [147]. Generally, two main approaches are used to overcome this issue: computational color constancy and color invariance, in which the colors are changed accordingly to color invariant extracted features [156].

- Computational color costancy: firstly, a scene light estimation is performed. Then, a color correction based on the previous light estimation is also performed. RGB values are consequently modified.

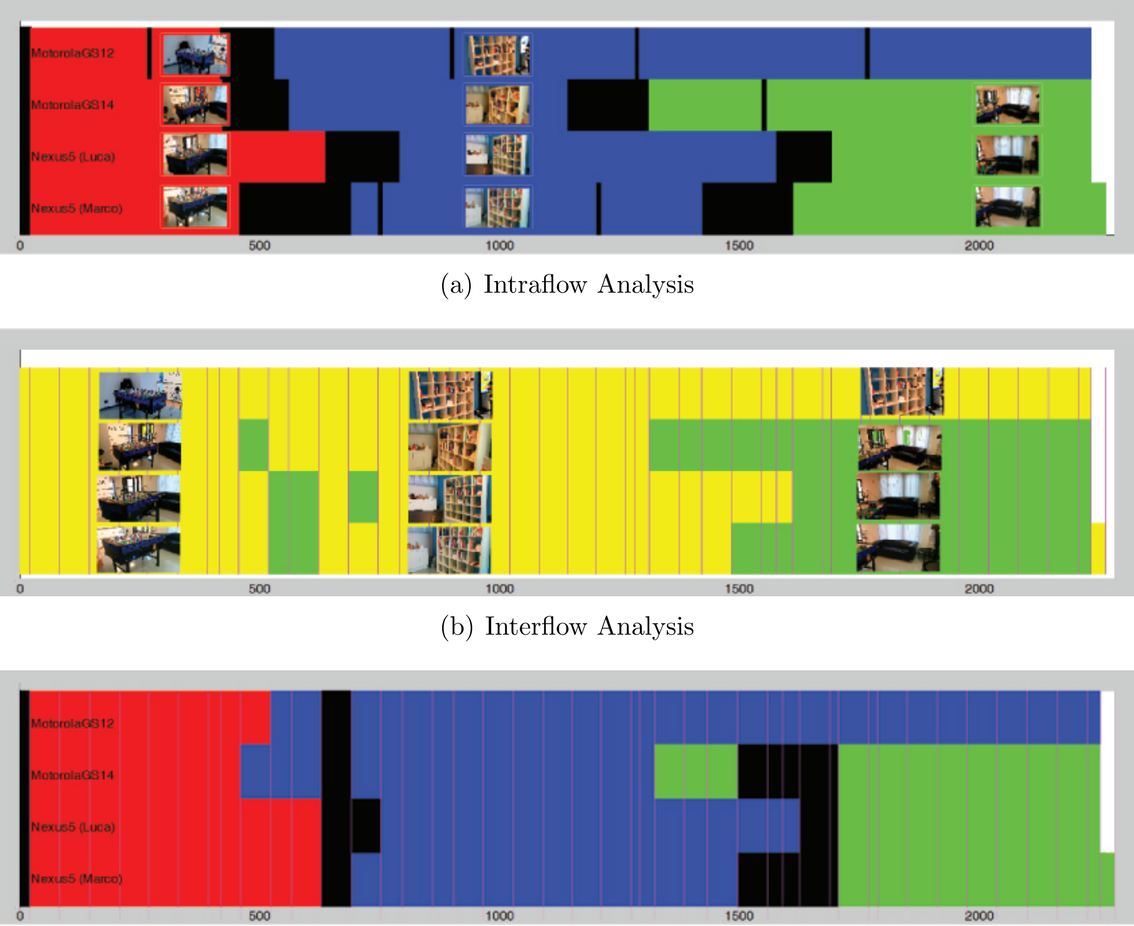
- Color invariance: RGB values are changed accordingly with several color invariant features extracted.

Although these two approaches are generally valid, Finlayson et al. [146] demonstrate that they (and all of their derivatives) are not good solutions when applied to image retrieval issues, since their performance is not strong against changing light conditions and no ones of them is designed for the multi devices scenario. Hence, we employed a definition of light conditions (and almost devices) independent representation given in [147]. The method is based upon the observation that changes of light conditions or device directly change the RGB values of the frame, while order of sensors response remains the same. Indeed, RGB values are the result of the processing of the electrical signals generated by sensors. These signals, although specific for each kind of device, maintain the same order independently from the acquiring device. Finally, equalization of RGB channels, as described in [147], is performed. After the normalization of the color domain, video streams are analysed in our approach in three phases (Fig. 5.1), detailed in the following. Firstly, RECfusion segments different scenes, transitions and unstable intervals within a video stream through the intraflow analysis. Then, by exploiting the interflow analysis, it finds devices that observe the same scene. Finally, clusters are identified over time among video streams. The final video can be rendered automatically by the system, which chooses the most popular cluster within each time-slot, or manually by the user, which can select the output video among video streams or scene clusters suggested by the system.

5.3.1 Intraflow Analysis

The intraflow analysis segments the sequence of frames of a single video stream (Fig. 5.1(a)). During intraflow analysis frames of each video are processed comparing their visual contents. For each frame of the video flow, we extract keypoints using the SIFT detection algorithm [62]. The set of the extracted SIFT features represents a template for the acquired scene. In this way, the comparison between frames could be done as the comparison between SIFT templates. At each instant, the last stable set of SIFT features extracted from the last detected scene is kept as a reference template and compared with the current frames. A reference template might change due to several factors, such as: a change of scene or light condition, temporarily occlusion of the main subject, transformations of the video subjects, sudden movements of the operator, and so on. When the comparison between the current frame end the reference template generates a sensible variation of features (i.e., low matching score), then the algorithm refreshes the reference template and splits the video producing a new segment. All the frames within a segment have coherent visual contents.Two frames are considered “similar” in relation to the number of matchings between their SIFT keypoints. To make the matching more reliable, we reject matchings where keypoints are too far in terms of spatial coordinates by assuming smooth transition between consecutive frames [7]. We used a threshold distance of 100 pixels for images with resolution 1280 × 720 or 1920 × 1080. Also, we do not consider SIFT extracted from the border of frames to make more robust the feature matching procedure. In order to detect sudden changes of the number of matchings we defined a slope function which is computed on a frame at time T as follows:

$$slope(T) = \frac{h}{w} = \frac{l \sin\theta}{l \cos\theta} = \tan\theta \tag{5.1}$$



(a) Intraflow Analysis

(b) Interflow Analysis

(c) Cluster Tracking

Figure 5.1: RECfusion results applied on Foosball dataset. Chronograms show results of the three main steps of RECfusion (intraflow analysis, interflow analysis and cluster tracking). Foosball dataset is composed by 4 video streams having a duration of ~ 2300 frames (~ 90 seconds). Each video stream is represented as a row in the chronograms. Vertical red lines mark the end of time-slots. (a) Intraframe analysis: red, blue and green frames are respectively the first, second and third scene of each video stream. Noisy frames are depicted in black. (b) Interframe analysis: yellow and green clusters are respectively the first and second cluster of each time-slot. (c) Cluster tracking: red, blue and green clusters are respectively the first, second and third cluster of the whole video set. Noisy clusters are depicted in black.

This function represents the variation of the number of matchings h in a range interval w (duration of time) centered in a frame at time T. This value is related to the tangent of an angle θ which is proportional to the gradient of the matching curve.

For instance, if the slope function has a peak greater than 10 (i.e., $\theta = 85°$), then algorithm attempts to create a new template. However, For major stability, a new template can be defined only if it has a duration greater than 2 seconds, otherwise it is considered as noise. In other words, a template is considered a *stable template* if the number of matching SIFTs do not change too much in time. A *backward checking* is required in order to understand if a new defined template regards a new scene or it is related to a previously observed one. The algorithm compares the new defined template with the past ones, starting from the last found template. Two different templates of the same scene could be rather different due to the elapsed time between them. During this step, we use a geometric verification to exclude the spatial matchings, in order to check if two templates describe the same scene. The threshold distance for matching changes from 100 pixels to one-third of the height of the image. In this way, we focus on the most likely salient point of the scene, that is supposed to be found in the center of the frame. This approach aims to subdivide the scene in three horizontal parts and exclude matchings between the upper side of one image with the lower one of the other image. If the percentage of matchings after the geometric verification is greater than 50% of the original matchings, then the two matching templates are assigned to the same scene. Each reference template is labeled with a $SceneCounter$ and all video frames achieving a robust match are labeled with the same $SceneCounter$. Note that all the frames are required to decide if a template should be considered as a new or an updated one are labeled as a transition interval. On the other hand, all the frames between the instant when a new scene template have to be upgraded by the system and the instant when that template is finally upgraded are labeled as a transition interval.

5.3.2 Interflow Analysis

The interflow analysis is computed for each time-slot. It segments video frames labeled by intraflow analysis and assigns a $ClusterCounter$ with respect to all the video streams in that specific time-slot (Fig. 5.1(b)). In our case, frames have already been segmented by intraflow analysis; through the output of that analysis (the $SceneCounters$ and the set of SIFT features used as scene template) We want to group together the devices that are looking at the same scene over time. Then,

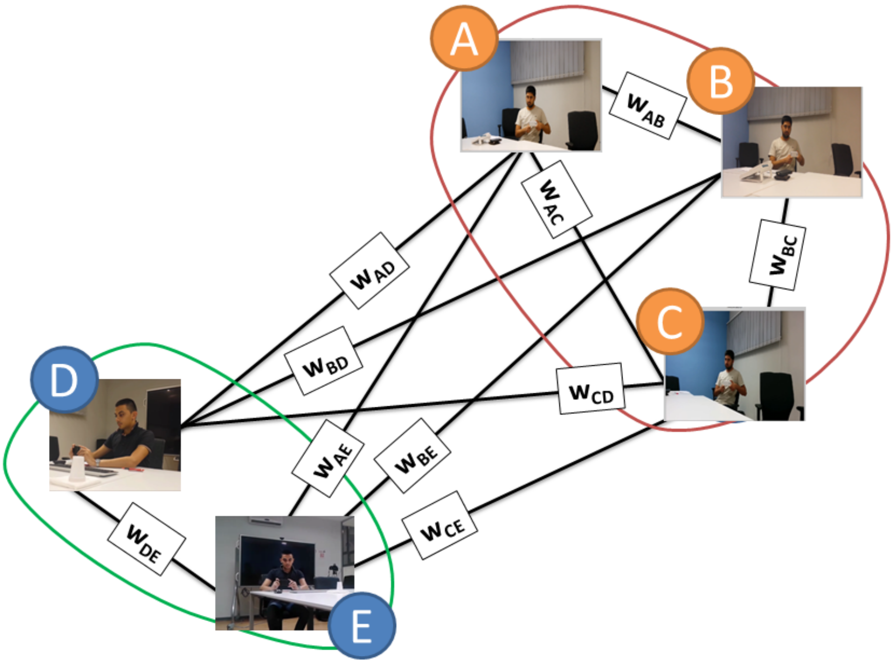


Figure 5.2: Graph of interflow analysis: arcs are weighted with interflow distances (Eq. 5.3.2). Two clusters can be defined, accordingly to interflow analysis definition.

the most popular scene, which is related to the group with the highest number of devices, could be used to produce the final video.

The descriptor used in the interflow analysis is based on weighted color histograms [157]. In this context the device invariance should be granted as well as possible. For this reason we firstly apply an histogram equalization, as suggested in [147]: each RGB channel is independently equalized and this allow us to reduce the variability introduced by the IGPs related to the different devices. The equalization is followed by a quantization of the color space (8 colors for each channel). Weights are obtained by using a gradient map as suggested in [157]. The gradient map is useful to highlight the structures of the objects that appear in the scene, making more robust the descriptor. As in [157], histograms are compared by exploiting the following distance:

$$d(h_{D_a}, h_{D_b}) = \frac{\sum (h_{D_a} - h_{D_b})^2}{\sum (h_{D_a})^2} \tag{5.2}$$

Table 5.1: Overview of the main variables employed in Cluster Tracking procedure.

Variable Name	Assigned by	Assigned to (scope)	Meaning
$SceneCounter_{ID}$	Intraflow Analysis	Scenes of a single Video Stream	Distinguish scenes within the same Video Stream
$ClusterCounter_{ID}$	Interflow Analysis	Frames of a single Time Slot	Distinguish scenes within the same Time Slot
$LoggedCluster_{ID}$	Cluster Tracking	Scenes in the whole Dataset	Distinguish scenes within the whole Dataset
$TrackedCluster_{ID}$	ClusterTracking	Scenes in the whole Dataset	Identify the most similiar LoggedClusterID in the previous Time Slots

where h_{D_a} and h_{D_b} are the weighted histograms related to the two frames of two different devices D_a and D_b.

Hence, to cluster devices accordingly to visual content at every time-slot, we firstly segment the sequence of frames of each video stream with the intraflow analysis (Section 5.3.1), and then we use the weighted color histogram representation to compare frames of each time-slot.

The different scenes obtained with the intraflow analysis could be considered as nodes of a complete graph in which arcs are weighted with interflow distances (Eq. 5.3.2) between the scenes acquired by devices (Fig. 5.2). The clustering procedure selects a frame among the unclustered frames and assigns it to the most similar cluster. We used an average linkage approach to compare a frame with a cluster: the distance between a frame and a cluster is given by the average distance between the frame and all the elements within the cluster [7]. A frame is assigned to the cluster with the minimum average distance if the distance in Eq. 5.3.2 is lower than a threshold. In [7] we empirically observed that Eq. 5.3.2 returns positive values lower than 2, so this threshold was set to 1. If all the average distances between the frame and clusters are higher than the threshold, then the clustering procedure creates a new cluster containing only the considered frame.

5.3.3 Cluster Tracking

For each time-slot, the output video is composed by selecting the frame belonging to the most popular scene [7]. Scenes from other clusters might be interesting for users, so it is useful to track all the clusters upon time. Exploiting this tracking, clusters can be recognized through time. We are going to define several variables with complex names and relationships. For the sake of readability and comprehensibility, the variables defined in the following are also summarized in Table 5.1.

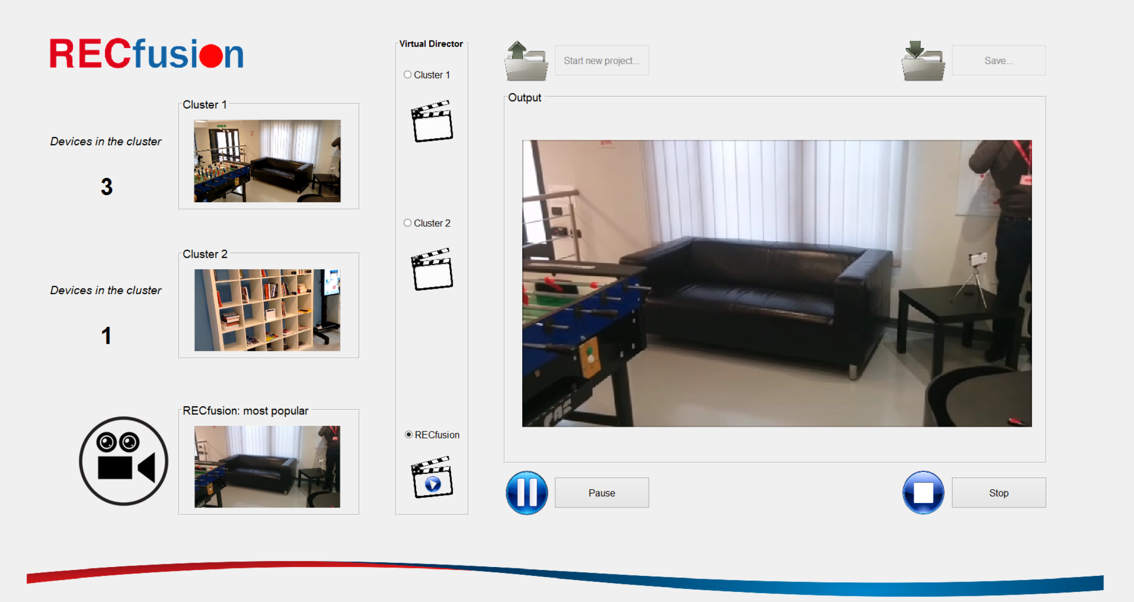


Figure 5.3: RECfusion Graphical User Interface showing the Cluster Tracking framework. On the left, active clusters with respective amount of recording devices and automatically suggested video stream (called *RECfusion: most popular*) are shown. User can browse the “Virtual Director” panel to dinamically change the active video stream. On the right side, active video stream with classic video player commands is shown.

To understand the meaning of the Cluster Tracking module we have to step back to intraflow analysis. The intraflow analysis segments the sequence of frames of a single video stream, and assigns a $SceneCounter$ to each segmented scene. However, frames taken by two different video streams but labeled with the same $SceneCounter$ can represent different scenes, since $SceneCounters$ are discriminative only within a single video stream. The interflow analysis segments video frames in a time-slot and assigns a $ClusterCounter$ to the scenes of the video streams. Interflow analysis exploits the $SceneCounters$ and the set of SIFT features templates from intraflow analysis. Similarly to $SceneCounters$, the $ClusterCounters$ are to be considered only within a single time-slot. Therefore, we developed a cluster tracking procedure in order to track clusters representing the same scene in every video stream and time-slot (Fig. 5.1(c)). In [8] a Graphical User Interface implementing cluster tracking typical video player commands (like Start, Pause, Stop, . . .) is described (Fig. 5.3).

We propose a cluster tracking procedure based on a voting routine that combines results of the intraflow and interflow analyses. Once interflow procedure has assigned a $ClusterCounter$ to several $SceneCounters$, this set of scenes will characterize the same cluster also in further time-slots, so cluster tracking procedure an unique $LoggedCluster_{ID}$ to this set of scenes. Differently from the $ClusterCounters$, the $LoggedCluster_{IDs}$ are intended to be always discriminative. Cluster tracking procedure tracks the clusters in each time-slot assigning them $TrackedCluster_{IDs}$ equals to the most similar $LoggedCluster_{ID}$. In order to define the most similar $LoggedCluster_{ID}$, cluster tracking procedure requires an initialization phase (at first time-slot). In this phase, the assigned $LoggedCluster_{IDs}$ are equals to the $ClusterCounters$. Then, from the second time-slot on, clusters will be associated to an existent $LoggedCluster_{ID}$ or to a new one, depending on a voting routine. The same routine is also used to track the $LoggedCluster_{IDs}$ with proper $TrackedCluster_{IDs}$.

The voting routine can be divided into 2 phases: casting of vote and voting decision. In the former phase, for each time-slot, each scene votes with three different possible values: $TrackedCluster_{ID}$ at the previous time-slot, $LoggedCluster_{ID}$ or unlogged scene (V_N), if the scene is *Noise*, already logged or unlogged, respectively. Once all the votes are casted in a time-slot, then we look for a non ambiguous voting decision (i.e., a majority is found). Majority of unlogged scenes is not admitted, so in this case we simply remove these votes from the voting decision. Depending on the reached decision, new $LoggedCluster_{IDs}$ might be instantiated, while $TrackedCluster_{IDs}$ at current time-slot is eventually updated. We will compare the new proposed method with respect to a cluster tracking method based on a threshold T_{CT} [8]. This threshold was used as an hyperparameter to decide whenever to create a new $LoggedCluster_{ID}$ or not. The issue with this threshold employed in [8] is that its value should be fine-tuned for each video set in order to achieve the best results in cluster tracking procedure.

Centroids weighted-update

The cluster tracking procedure also stores and keeps up to date the centroids of logged clusters in a buffer called *ClusterLog*. For each time-slot, the distance between all the videos in a cluster and its centroid logged in the *ClusterLog* is computed, in order to automatically select the best representing video stream of that cluster. The distance is computed with Equation 5.3.2, while centroids are updated with the following weighted-update Equation:

$$L'_a = \frac{u}{u+1} L_a + \frac{1}{u+1} C_b \tag{5.3}$$

where L'_a is the updated logged-centroid, L_a is the former logged-centroid, C_b is the centroid of tracked cluster and u is the number of updates performed (the denominator is increased by 1 to take into account also the first insertion). So, in the first update the weigth of C_b is 1/2, in the second update is 1/3 and so on. In this way, logged-centroids become increasingly stable after each update.

Cluster Tracking GUI

In [8] a Graphical User Interface implementing typical video player commands (like Start, Pause, Stop, ...) is described. The user can select one of the active clusters, which are shown on the left side of the GUI within the "Virtual Director" panel, and dynamically view it on the right side of the GUI, switching from a cluster to another (Fig. 5.3). The output stream is continuously recorded as it has been mounted directly by the user. The GUI reports the number of devices in each cluster and the user can exploit this information to estimate the popularity of each cluster. Then, like a director, he can choose the most popular cluster from the point of view of the community, automatically suggested by the framework, or from his own one.

5.4 Datasets

To perform experiments we have used the RECfusion dataset [7] which is publicly available at [158]. This dataset is made up of three video sets:

1. *Foosball*: indoor context, some people appear in the scene. The number of contributing devices for this video set is 4, with an average number of frames per video stream of 2250 (44 time-slots). There are three main subjects in this video set: a foosball, a couch and a bookcase.

2. *Meeting*: indoor context, two people appear in the scene. The number of contributing devices for this video set is 5, with an average number of frames

per video stream of 2895 (60 time-slots). There are two main subjects in this video set (two people).

3. *S. Agata*: outdoor context, lots of people appear in the scene. The number of contributing devices for this video set is 7, with an average number of frames per video stream 1258 (34 time-slots). There are two main subjects in this video set: the reliquary of S. Agata and the facade of a church.

In the experiments, we exploit also a video set from the dataset used in Ballan et al. [159]. This dataset is named *Magician*. It is related to an indoor context, where one person appear in the foreground. The number of contributing devices for this video set is 6, with a fixed number of 3800 frames per video stream (77 time-slots). There are two main points of view in this video set: one above and one in front of the magician. We have chosen *Magician* video set because it is slighty different from the videos currently in RECfusion dataset. In *Magician* all the video streams are focused on a single target and are acquired as a "casual multi-view video collection" [159]. This means that backgrounds in video streams are very different from each other and that severe camera motion could often appear. The casually filmed events represent a challenging scenario for detector like SIFT (exploited in our intraflow analysis, see Section 5.3.1), so we add *Magician* video set to our tests in order to stress and evaluate scene analysis and cluster tracking performances. We have also compared obtained results with the benchmark dataset proposed in Hoshen et al. [150]. This dataset has been acquired with wearable devices and, like *Magician* video set, it is challenging since every video is strongly affected by motion.

5.5 Experimental settings and results

We define time-slots as sets of 60 consecutive frames. We select the last instant of time for every time-slot as the representative of that interval. Validation are made exploiting the Ground Truth with respect to these representative frames. To evaluate the performances of the proposed method, we compute two quality measures described in [7]. Specifically, for each clustering step we consider:

Table 5.2: Validation Results of Popularity Estimation.

Scenario	**Devices**	**Models**	P_a/P_r	P_g/P_r
Foosball	4	2	1.02	1
Meeting	2	2	1.01	0.99
Meeting	4	4	0.99	0.95
Meeting	5	5	0.89	0.76
SAgata	7	6	1.05	1
Magician	6	6	0.73	0.73
Concert [150]	3	1	1.06	1
Lecture [150]	3	1	1.05	0.86
Seminar [150]	3	1	0.62	0.62

- P_r: ground truth popularity value (number of cameras looking at the most popular scene) obtained from manual labelling;
- P_a: popularity score computed by the system (number of the elements in the popular cluster);
- P_g: number of correct videos in the popular cluster (i.e., the number of inliers in the popular cluster).

From the above defined scores, the weighted mean of the ratios P_a/P_r and P_g/P_r over all the clustering steps are computed. The ratio P_a/P_r provides a score for the popularity estimation, whereas the ratio P_g/P_r verifies the visual content of the videos in the popular cluster and provides a measure of the quality of the popular cluster. Note that P_a/P_r is a score: when is lower than 1, then it means that system is under-estimating the popularity of the cluster, while, conversely, it results in an over-estimation.

The results of the comparison between tested video sets are shown in Table 5.2. The first five rows are related to RECfusion dataset, whereas the last three rows are related to the dataset proposed in [150]. Although the constantly head motion of the wearable recording devices in videos from [150], the framework reaches good results and seems to be promising room for improvement in the field of wearable devices. Conversely, we found a drop in the performances when there is a severe difference of scale between videos in a video set. Indeed, we exploited *Meeting* video set to evaluate the drawback in performances when there are high differences

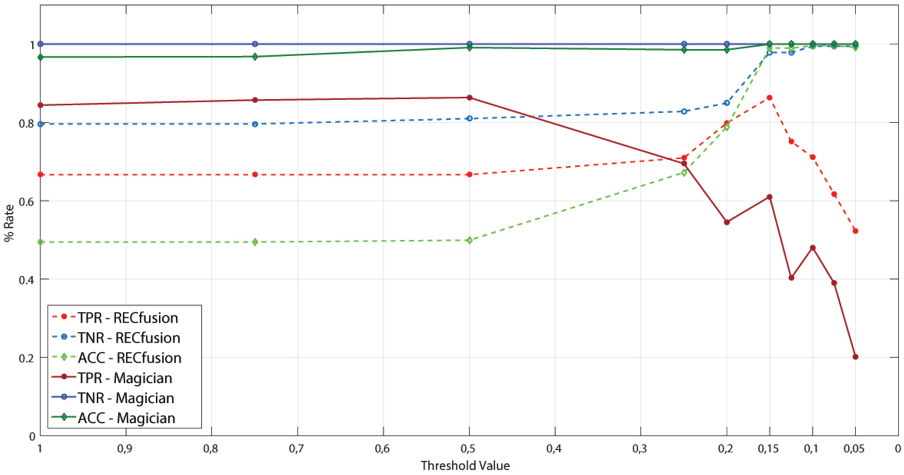


Figure 5.4: A comparison of *TPR (True Positive Rate, or Recall)*, *TNR (True Negative Rate, or Specificity)* and *ACC (Accuracy)* between RECfusion_dataset_2015 and *Magician* video set cluster tracking validations using the threshold-based procedure from [8]. As can be seen, *Magician* requires a fine tuned threshold to increase *TPR*, *TNR* and *ACC* values.

between resolution of devices. We compared three cases, with 2, 4 and all the 5 devices in *Meeting* video set, respectively. Other analysis outputs could be found at [160].

In the vote-based procedure we removed the threshold T_{CT}, used in [8] as an hyperparameter to decide whenever to create a new logged-cluster or not. In [8] the value of T_{CT} was empirically set equals to 0.15 founding the best overall value between True Positive Rate, True Negative Rate and Accuracy of clustering tracking procedure on RECfusion dataset. In Fig. 5.4 a comparison between the average values of *TPR (True Positive Rate, or Recall)*, *TNR (True Negative Rate, or Specificity)* and *ACC (Accuracy)* of RECfusion dataset and *Magician* video set whit several values of T_{CT} is shown. As can be seen, the value of T_{CT} equals to 0.15 is not the best value to be used by cluster tracking procedure, while $T_{CT} = 0.5$ should be used instead. For this reason, in [9] we proposed the threshold independent cluster tracking procedure described in Section 5.3.3. We computed *TPR*, *TNR* and *ACC* values for each video set described in Section 5.4 and compared them with the results obtained in [8]. The comparative validation results are shown in Table 5.3.

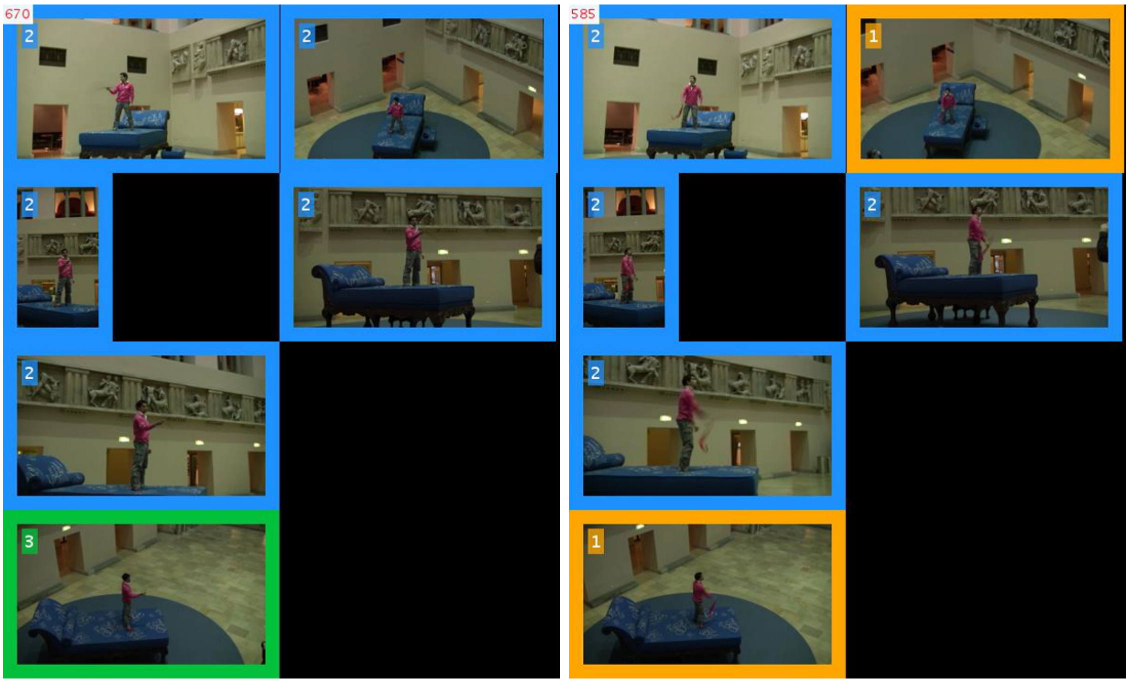
These results show that the proposed vote-based cluster tracking procedure

Table 5.3: Validation results between cluster tracking procedure threshold-based and vote-based.

		TPR (RECALL)		TNR (SPECIFICITY)		ACC (ACCURACY)	
DS	Scene	[8]	PROPOSED	[8]	PROPOSED	[8]	PROPOSED
Foosball	1	0,91	**0,92**	0,70	**1,00**	0,69	**1,00**
	2	0,69	**0,97**	**0,98**	0,91	**0,99**	0,97
	3	0,41	**0,74**	**1,00**	**1,00**	0,50	**1,00**
	MEAN	0,67	**0,87**	0,89	**0,97**	0,73	**0,99**
Meeting	1	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,80	**1,00**	**0,95**	0,93	**0,83**	0,67
	3	0,43	**0,50**	**1,00**	**1,00**	0,70	**1,00**
	MEAN	0,74	**0,83**	**0,98**	**0,98**	0,84	**0,89**
S.Agata	1	0,71	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,87	**0,97**	**0,49**	0,14	**0,80**	0,68
	3	**0,48**	0,00	**1,00**	**1,00**	**0,60**	0,00
	MEAN	**0,69**	0,66	**0,83**	0,71	**0,80**	0,56
Magician	1	0,73	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,45	**0,56**	**1,00**	**1,00**	**0,98**	0,91
	MEAN	0,59	**0,78**	**1,00**	**1,00**	**0,99**	0,96

reaches *TPR* values much higher than the threshold-based procedure, while results on TNR and ACC are comparable between the two procedures. Just in the *Meeting* video set the proposed vote-based procedure is slighty outperformed: this is a limitation of the procedure. Indeed, cluster tracking procedure relies on intraflow analysis, so if the latter defines N scenes, then the former is able to distinguish at most N scenes. Hence, differently by threshold-based procedure used in [8], that can generate a bunch of small sparse clusters if T_{CT} is not fine tuned, in this case only a limited number of clusters is tracked. In *Meeting* video set two people are recorded and there are only two distinguished clusters focusing on each one of them. Sometimes interflow analysis generates a cluster containing both of the two people. This is treated by the cluster tracking vote-based procedure as *Noise*, since intraflow analysis has never labeled a scene in which the people are recorded together.

A final remark is about *Magician* video set. We added it to our dataset in order to evaluate scene analysis and cluster tracking performances in a video collection with a single scene, where all the users are focused on the same target and videos are affected by severe camera motion. Cluster tracking results with threshold-based procedure from [8] are really bad, indeed we got the worst average performance on this video set (Table 5.3). On the other hand, the proposed vote-based procedure reached good values of *TPR*, further assessing the soundness of this new cluster tracking approach. A comparison between threshold-based and vote-based methods



(a) (b)

Figure 5.5: Example of cluster tracking on *Magician* videos set. (a) Two clusters (marked in blue and green) with 5 and 1 video streams respectively, (b) Two clusters (marked in blue and yellow) with 4 and 2 video streams respectively.

is shown in Fig. 5.5, while the output videos showing the whole result of cluster tracking vote-based procedure could be found at [160].

5.6 Conclusion

In this Chapter we described RECfusion, a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Given a set of video streams as input, the framework can group these video streams by means of similarity and popularity, then it automatically suggests a video stream to be used as output, acting like a “virtual director”. We compared RECfusion intraflow and interflow analysis validations with Hoshen [150]. We have added a video set from Ballan et al. [159] to our RECfusion dataset showing that RECfusion is capable of recognize and track the scenes of a video collection even if there is a single scene, where all the user are focused on the same target and videos are affected by severe camera motion. In [9], we proposed a vote-based cluster tracking procedure and compared it with the one threshold-based described in [8]. From this comparison we found that vote-based procedure reaches very good results totally automatic and independently by a hyperparameter fine tuning phase, but with the tradeoff of be unable to create and track an unlimited number of clusters. As future works and possible applications, we are planning to augment the framework with features specifically focused on Assistive Technology or Security issues (i.e., highlight/track bad behaviour in the life style, log the visited places, search something or someone that appears in the scene).

Part III - 3D Data

Chapter 6

3D Data Analysis for Cultural Heritage

6.1 Introduction

3D scanning has gone a long way since its first appearance in cultural heritage digitization and modeling. In the recent years some new low cost, fast, accurate emerging technologies are flooding the market. 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 and to test their effectiveness in terms of easiness of 3D data collection, processing, analysis, mesh resolution and metric accuracy against the size and features of the objects.

The contents of this Chapter are based on our published papers [161, 162, 163, 164, 165, 166]. In this chapter we describe several real case studies of 3D data analysis for Cultural Heritage. The structure of the Chapter is the following: in Section 6.2 we describe in detail the handheld 3D scanning technology and the employed Structure scanner [167]. Then, we report the real case studies: a doorway of the Monastery of Benedettini in Catania [164] in Section 6.3, the Morgantina Silver Treasure [162] in Section 6.4 and Kouros from Leontinoi [165] in Section 6.5. Discussion and conclusions end the Chapter in Section 6.6.

6.2 Handheld 3D scanning

The 3D scanners are devices able to collect geometry information about a real-world object or environment. Then, this information is 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. [168, 169, 170, 171, 172, 173, 174, 175, 176]. Since these devices work employing many different technologies and their cost changes in a wide price range, then 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).

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 estimate their geometrical properties. On the other hand, sensors that do not introduce waves in the environment are called *passive sensors*. In this latter case, methods for 3D acquisition include by stereo vision or Structure-from-Motion (SfM) [177].

From another perspective, the 3D scanning devices can be categorized w.r.t. their portability. In recent years, thanks to the miniaturization and integration of the electronic and optical sensors, small and compact high performance 3D scanners have been released [178, 179]. 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 [163]. For the relative low-cost and usability, most of these devices became consumer electronics product, while others are still used in professional context. However, they represent a great resource in the field of Cultural Heritage. Below, we report a list of the main handheld 3D scanners and their knows specifications in Table 6.1:

Microsoft Kinect is mainly used in home videogames entertainment. However, some example of applications of these devices to Cultural Heritage could be found in the works of Cappelletto [180] and Remondino [174].

Scanify Fuel 3D is a handheld device that employs combination of photometric and stereography techniques to acquire depth information. In this way, it can reach a high accuracy.

Google Project Tango is a Google device that employs motion tracking to understand position and orientation of the device. 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 thought for architectural elements such as doors, statue etc. An example of Cultural heritage application can be find in [161].

Structure Sensor is a small active scanner produced by Occipital. It employs structured light technology to guarantee a good quality scan with a low required time and with an affordable price. This device has been employed in the study conducted on this Chapter, hence more details are provided in the following Sections.

Table 6.1: Specifications of the described handheld scanners [163].

Scanner	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)

6.2.1 Employed device: Structure sensor

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 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 our study we explore the possibility of using Structure Sensor also for bigger architectural element like a doorway or a statue.

We employed the Structure Sensor Scanner [167] in our case studies as done for Medical Research described in Chapter 7. Similarly to Microsoft Kinect [163], this device has an operative range capability from 0.4m to 12m, although it is recommended a distance in the range 0.4 and 3.5 meters [167, 163]. Indeed, in the closer range from the sensor the device reaches a declared 3D point accuracy of 0.5mm. The accuracy could become greater (worst) 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. 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. 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.

6.3 Doorway of the Monastery of Benedettini in Catania

In this case study, our goal is to provide a full low cost and open source 3D pipeline highlighting potentialities and weaknesses of handheld 3D scanners, when compared to time-of-flight scanners. We chose an eighteen century doorway in Benedettini monumental complex in Catania (UNESCO heritage) located in the gallery at first floor of the monastery. Nowadays, this doorway is used as offices for the Department of Humanities of Catania University, but it was used to provides access to the cell of a friar. 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 we 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 [181].

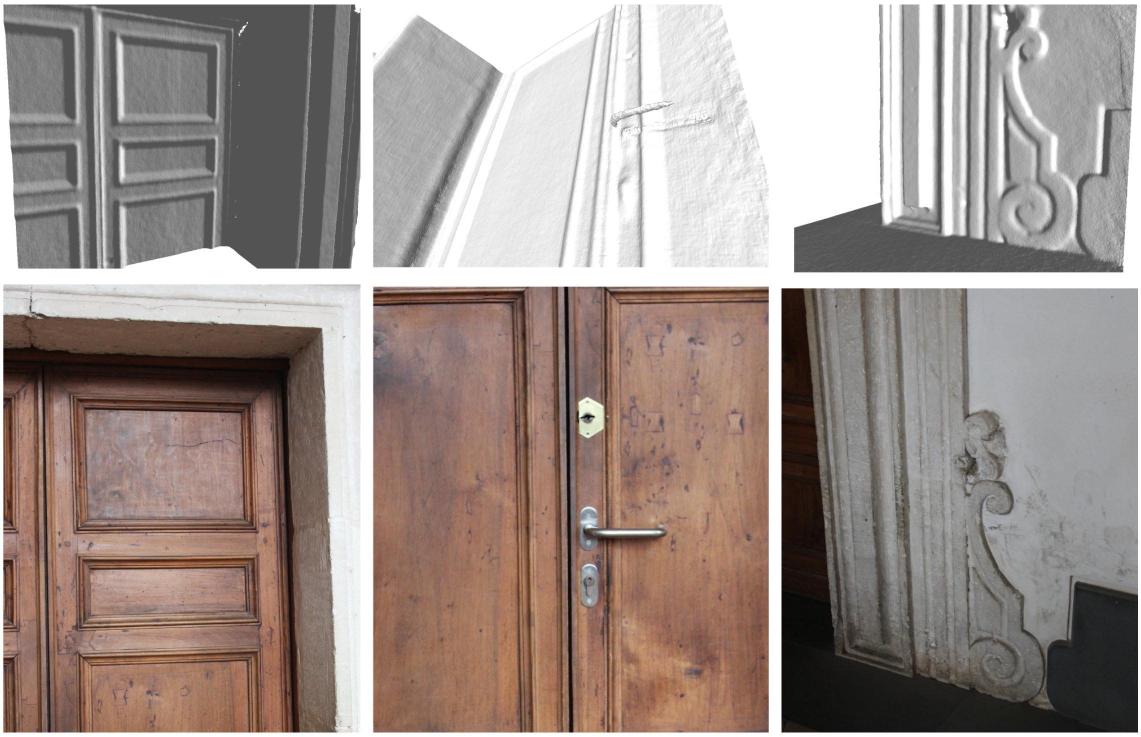


Figure 6.1: Doorway of the Monastery of Benedettini in Catania. View of the Structure Sensor behavior on three different materials (from left to right): wood, metal, limestone.

Relatively to issues discussed in Section 6.2.1, our case study is located in an indoor environment not affected by direct sunlight interferences. The doorway 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 6.1). Our case study presents a geometry complex enough to enable a good acquisition with the Structure sensor.

6.3.1 Method description

As said in Section 6.2.1, 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 for obtain a quality sufficiently good. For this reason, we decide to set a scanning volume of $1m^3$ and subdivide the acquisition of the door into several single acquisitions. We acquired a total of 23 parts, starting from the bottom left position until the top right. All 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. We process and align the meshes exploiting the software Meshlab [182]. 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 were present, using several filters from MeshLab (i.e., Quadric Edge Collapse Decimation, Remove Isolated Vertices) [182]. 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.

6.3.2 Comparison with Time of Flight 3D scanning

In this subsection we deal with the experimental results obtained during the visual and metric accuracy tests. As ground truth we use a ToF mesh model. The pipeline followed is by the time used in literature [174, 173, 183, 184, 185] and foresees alignment of the different models in the same reference system and distance calculation between the meshes, by means of Hausdorff distance algorithm application [186]. Considering performances of the handheld scanner and purposes of this Chapter, we consider both details (i.e., a capital and frames and moldings of the jams and entablature) and the overall doorway.

During ToF laser scanner acquisition (using a HDS 3000 by Leica Geosystem [187]) we decided to carry out three scans: one frontal and two lateral and we chose a scan step very dense (about $2mm$) to have a very detailed point cloud. In these cases, as reported in previous literature works [188], the size of the noise exceeds the sampling rate in such a way 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, we carried out the merging of the scans into a unique model, then we applied the pipeline employed in Ref. [188] 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

Table 6.2: Doorway of the Monastery of Benedettini in Catania. Experimental results. All values are expressed in mm.

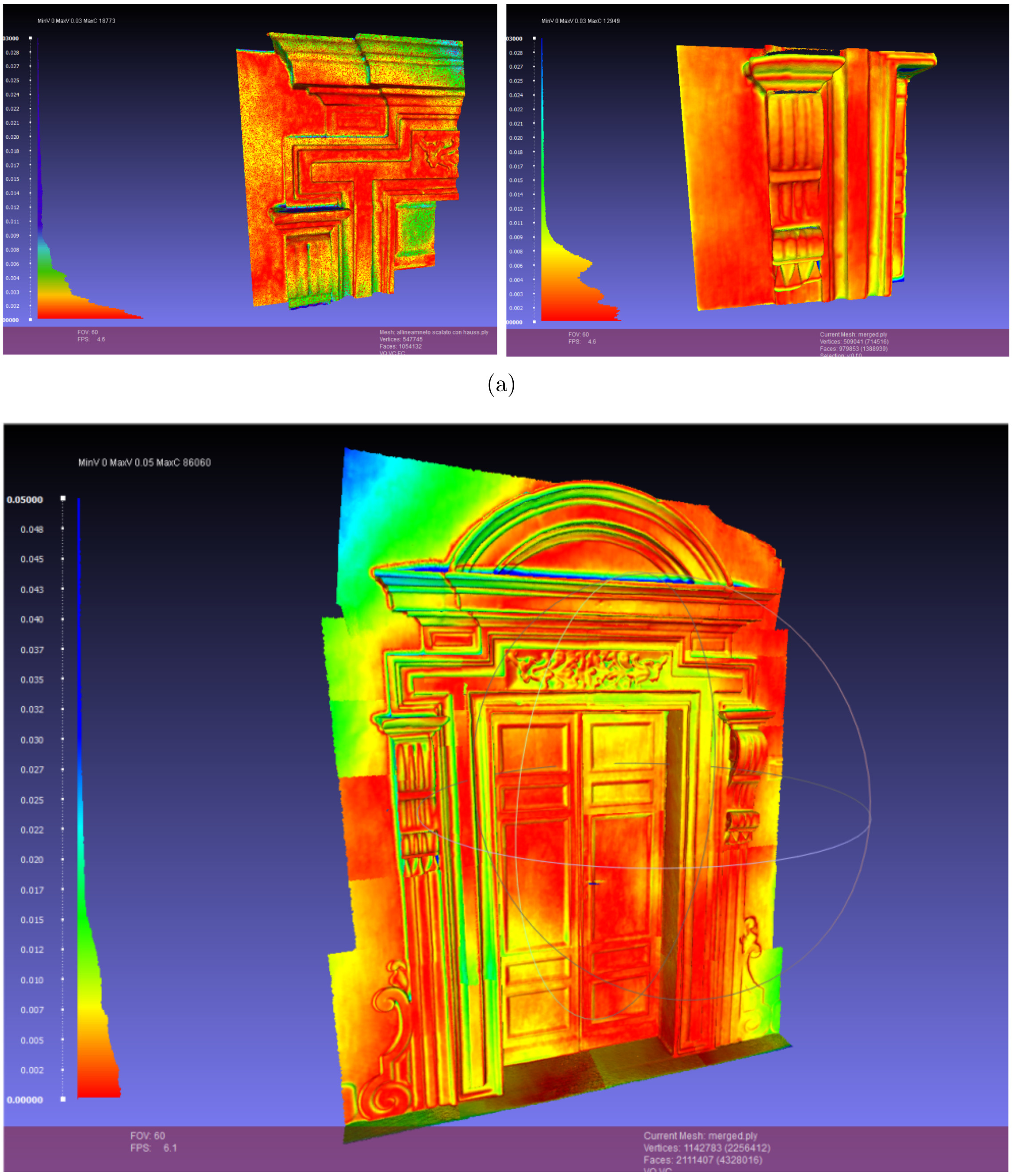
Model	Hausdorff Range	Mean	RMS
Capital	$0 - 30$	4.344	6.879
Entablature	$0 - 30$	4.775	6.797
Overall Model	$0 - 50$	9.619	1.4104

and less noisy with respect to ToF reconstructions. This is in line with the typology of used sensor. The comparison between the three models (details and the overall doorway) and their corresponding ToF scans was carried out in Meshlab [186]. 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. We calculate Hausdorff distance with the following range values $0 - 50mm$, in order to take into account these mismatches.

The experimental results are shown in Table 6.2. 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 (Fig. 6.2), where the red color highlights the minimum distance between the two meshes, while the blue the maximum one.

6.4 Morgantina Silver Treasure

The silver treasure of Morgantina, one of the most valuable collections of the *Museum of Aidone (Sicily)*, consists of 16 pieces of worked silver that were returned to Italy in 2010 following an agreement among the *Ministero dei Beni e le Attività Culturali (Regione Siciliana)* and the *Metropolitan Museum of Art of New York* (Fig. 6.3). Through police inquiry and data derived from direct archaeological excavation in a specific area of the ancient city of Morgantina, the origin of the collection was determined to be the so-called House of Eupòlemos, where the valuables were hidden probably during the Second Punic War [189, 190, 191].



(a)

(b)

Figure 6.2: Doorway of the Monastery of Benedettini in Catania. Hausdorff distance and subsequent quality histogram between TOF model and Structure Sensor model of (a) two chosen details and (b) the whole doorway.

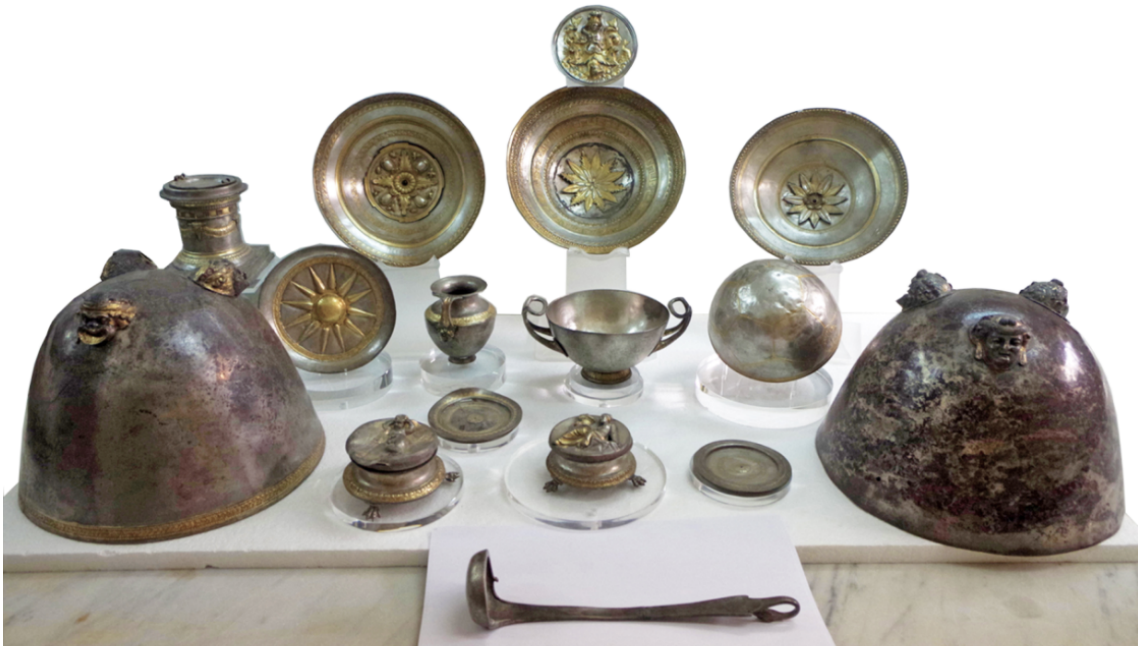


Figure 6.3: Morgantina Silver Treasure. The silver hoard from the Eupolemos's House—Archaeological Museum of Aidone (Sicily).

Since the agreement signed in 2006 provides for the alternating temporary exhibition of the silver treasure for 4 years at the Museum of Aidone and then 4 years at the Metropolitan Museum, a careful campaign of noninvasive analysis was prepared to document the conservation status and previous treatments of the collection. The stability of optimal conservation conditions is crucial for the life of a work of art. The archaeological silver artifacts are rare. At first sight, some are in very good condition, and it is easy to forget that they are very fragile and brittle, and their state of conservation depends on their history and the burial time. Often they can have some bodies damaged by microfissures, tiny-cracks, burial accretions, corrosion, and missing parts. The conservation of these artifacts should depend on a detailed understanding of the state of conservation, especially the kind of decay and embrittlement. The noninvasive investigations, carried out on the silver artifacts from Morgantina, provide a first overall picture of the conservation of the 16 silver items.

With the aim to realize a new tool to increase the existing archaeological knowledge and to obtain referenced information of the conservation state, before and after moving of the collection from their exposure site during the next scheduled temporary exhibition, three-dimensional (3D) models and diagnostic data have been acquired and organized for the first time, in an integrated way, within a web-oriented platform. The noninvasive methods have provided complementary results for a more comprehensive evaluation of the state of conservation and of the execution technique. In particular, the diagnostic study was directed to

- distinguish the original material from degradation and/or restoration materials;
- obtain a deeper knowledge of the production technique;
- assess the current state of conservation and acquire useful data for scheduled monitoring.

Finally, one of the main purposes was to produce significant scientific material for an innovative and interactive display to offer to visitors, even during the period of absence of the silvers through a virtual model available in the museum or in a dedicated website, which can provide customized visits of the Morgantina treasure.

6.4.1 Web-Oriented Interface Framework

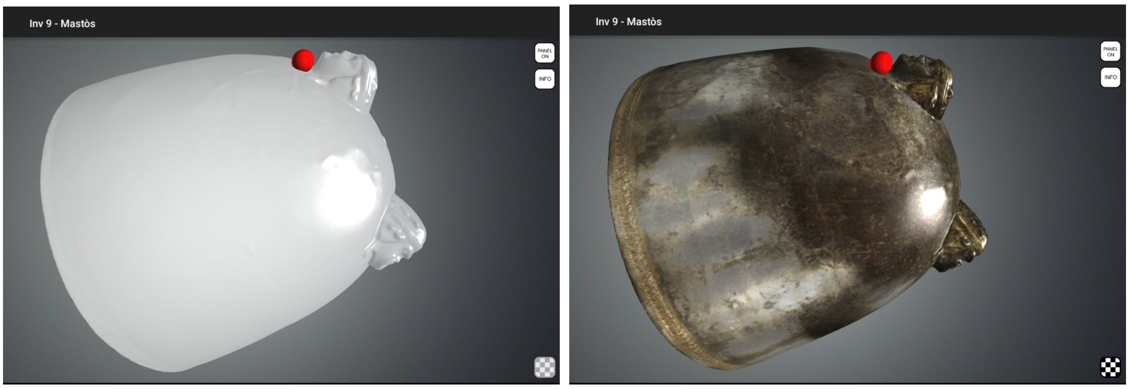
To make the 3D models and the archeometric data effectively available in a user friendly and integrated way, a web-oriented interface framework has been developed. Its main functionalities are the cataloguing of new 3-D scans and the management of additional metadata that can be implemented during the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

Through 3D scanning technologies employed in Cultural Heritage field [192, 171, 172, 193, 194], the Morgantina silver gilt treasure collection has been acquired to obtain 3D digital models. In this work, our intent also provides a user-friendly digital system to allow fruition and scientific analysis of the treasure pieces. To this aim, we have developed a software for two platforms: a web application developed by Unity 5.0 [195] and a mobile Java application for Android. In AppendixappA:app, we report a benchmark og 3D web viewers available online, which include [196, 197]. We aim to realize a customized software that includes functionalities specifically designed for Cultural Heritage scholars. Currently, the framework described in this work implements elementary functionalities, as it is a prototype thought to test the feasibility of a more longterm project.

Its main specifications are the cataloging of already existing or totally new 3D scans and the management of additional metadata. The digital version of the artifacts is augmented with *semantic annotations* about the history, measurements data, expert comments, and so on. The meaning of "semantic annotation" is the action and results of describing (part of) an electronic resource through metadata [198, 199]. Firstly, a comprehensive description of software specification is given. This description focuses on the system functionalities, which are independent from the platforms employed (web or Android). Secondly, a brief discussion about the technical details and the exclusive properties of the two platforms is reported.

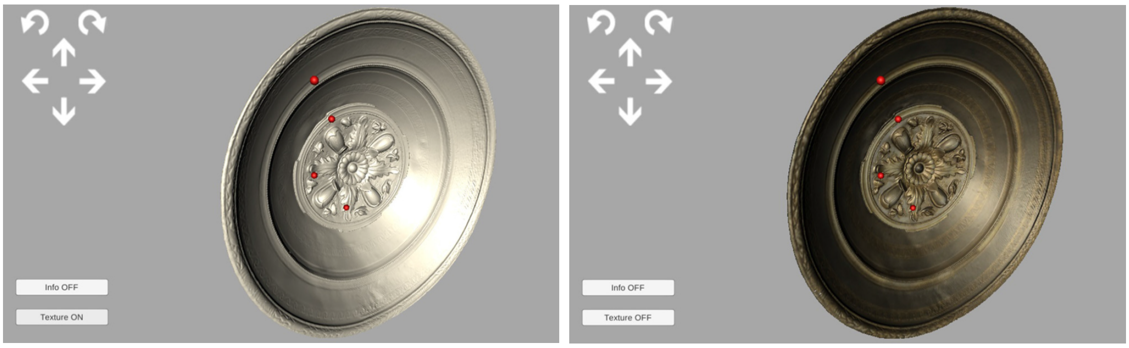
Through the graphical user interface (GUI), user browses a list of the available 3D models that are selectable for the investigation. Each one of the 3D models is placed in a different 3D environment, which offers conventional navigation functionalities such as rotating and zooming. Hence, the surface 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 proper buttons on the GUI (it depends on the available input peripheral device). Two visualization modalities are offered by the system to properly analyze the mesh surface: *shaded mode* and *textured mode*. Shaded mode is designed to permit an accurate geometric examination, as general shape and surface particulars look well marked with no texture data. This kind of surface analysis sounds well for a visual detection of alterations in the original form of the artifacts (e.g., deformation and missing parts). Texture mode shows material and color information of the 3D models, which are very valuable to check the conservation state of external surface. Chemical reactions (e.g., oxidation) or pigment scratches easily emerge when texture analysis is performed. Fig. 6.4 shows shade and texture modes for the Android platform, while Fig. 6.5 shows shade and texture modes for the Unity platform, respectively.

The main feature of the proposed system is the semantic annotation that enriches the original 3-D 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 for reporting analytic results and for the comparison of the same artifacts at different times. Interactive parts of the meshes



(a) (b)

Figure 6.4: Morgantina Silver Treasure. Android Platform. (a) Shade mode. (b) Texture mode.



(a) (b)

Figure 6.5: Morgantina Silver Treasure. Unity Platform. (a) Shade mode. (b) Texture mode.

are highlighted with well-noticeable markers, and when users select them a tooltip or a sided info-box shows the related info appear.

The proposed Unity web system is mainly intended to assist experts to explore 3D models and consult analytical reports. The system is able to work by using a simple internet browser with no other specific client application, but 3D models with lots of details require high bandwidth networks, in order to transfer the high amount of data. Moreover, the system can be accessible through internet networks to make the 3D artifacts available to researchers and scholars from all over the world.

The prototype has been developed by using Unity engine, version 5.0 [195]. It is an environment with an integrated game engine, provided by Unity Technologies, that is typically employed to produce digital games for different platform, such as PC, consoles, mobile devices, and websites. It handles 3D models and other kinds of assets, such as material, lights, image, and video. Unity 5.0 encodes algorithms in two different programming languages: C# and JavaScript. In this work, we employed C# and the Unity Integrated Development Environment (IDE) named Mono Develop. Although Unity is often used for digital game development, it can be employed for generic applications related to 3D modeling. The main advantage of Unity is the simplicity in managing multimedia resources and the user-friendly development GUI, as well as the multiplatform builder. To give the possibility to test our system, we provide a demo version available at [200].

The main aim of a mobile application (Android platform) is to follow user mobility. This leads us to develop a fruition system of Morgantina treasure to enrich users experience during museum visits, which is especially useful when the original artifacts are lent to other institutions. Nevertheless, we decide to keep the semantic annotation feature to give users historical information, as well as a further platform for research purpose.

The app has been developed 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 of the GUI. The 3D scene is drawn employing OpenGL ES (Open GL Embedded System), a subset of the standard OpenGL functions intended for embedded systems, such as 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 markers, we decide to use SQLite, the free database management system) adopted by Android. The developed application can be found as a demo version in [200]. To test it, we used a low-mid end device that mounts a CPU Intel Atom Z2520 Dual-core 1.2 GHz, a memory of 1 GB, a GPU PowerVR SGX544, and the OS Android 5.0. Currently, despite the good application portability, we cannot ensure the correct functioning of all the devices and Android OS version; a main requirement is an OS Android version 5.0 or higher.

6.5 An Archaic kouros from Leontinoi?

Greek Archaic sculpture is dominated by the production of statues of young naked boys, so called *kouroi* (plural of *kouros* meaning in Greek "boy"), and young girls with long vests, named *korai* (plural of *kore* meaning in Greek "girl"), having religious or funerary significance and for this reason generally offered as ex voto in sanctuaries or placed above or by tombs in cemeteries [201]. The statues were the symbolic representation of the worshippers consecrating their lives to the deities or idealized portraits of the dead. Their widespread distribution in the Greek Mediterranean between the end of 7^{th} and the early decades of 5th century BC testifies to the fortune of these iconographies which summarized the concept of kalokagathia, the combination of virtues - goodness and excellence - to which Greek civilization was devoted [202]. 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 [203]. However, very few life-size statues were found intact, as after the Classical age it became customary to detach the heads of Greek statues in order to create head-portraits. In fact, with few exceptions of statues found intact but in a smaller scale, this class of Greek statues in Sicily is represented just by heads without matching bodies, and headless bodies. A unique case is that of the "Biscari head" kept at the Museo Civico "Castello Ursino" of 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 6^{th} - beginning of 5^{th} century BC and almost unanimously believed to be part of the same life-size kouros [204]. The head was part of the private collection of Ignazio Paternó Castello, 5^{th} Prince of Biscari (1719-1786), the founding figure of early archaeological research and antiquarianism in 18^{th} century Sicily [205]. The head, also known as "Biscari head", retrieved 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 (Fig. 6.6) before being incorporated in the main collection of the Museo Civico "Castello Ursino" of Catania [206, 207]. In a rare picture taken around 1938

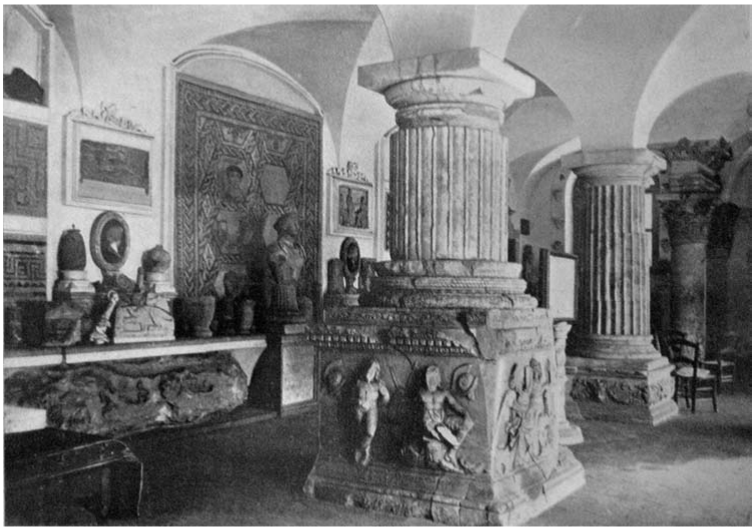
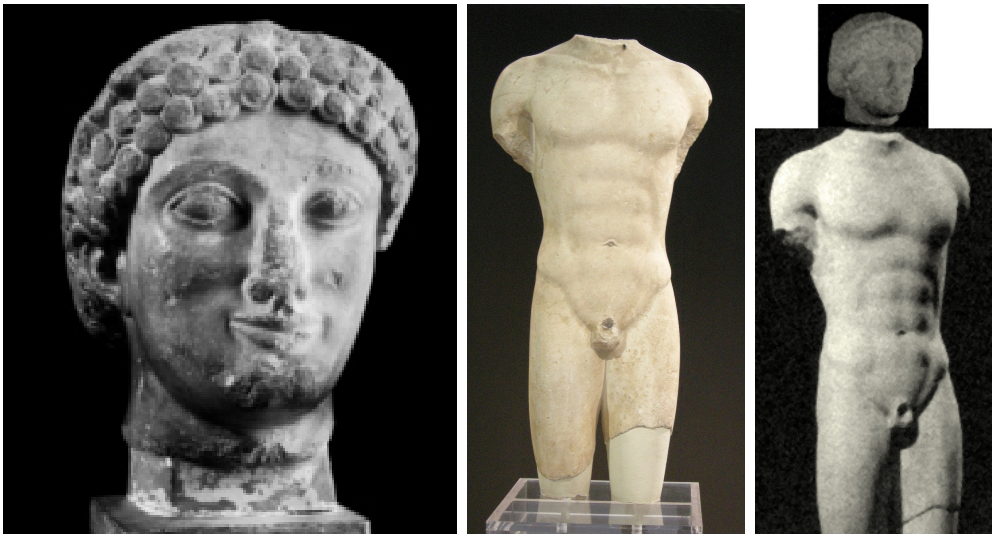


Figure 6.6: Catania, Museum of Palazzo Biscari alla Marina, Hall of marbles [205].

from the archive of Fratelli Alinari (Fig. 6.7(a)), the head appears set on gypsum base attached to a wooden pedestal, which was later removed.

The torso (Fig. 6.7(b)) was accidentally found in the country right outside the area of the ancient colony of Leontinoi and purchased in 1904 for 1,000 liras by Paolo Orsi from the Marquis of Castelluccio, who was another famous collector of antiquities. Due to the approximate context of provenance, the statue should have had funerary functions. As separated artefacts the two pieces were subject of several studies aimed to define their style, chronology and eventually also 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 try it on 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 for this experiment. Many decades after, Gino Vinicio Gentili reappraising the problem of the association of the two pieces published a photofit (Fig. 6.7(c)), where he matched the photographs of the head and the torso [208]. 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



(a) (b) (c)

Figure 6.7: Kouros from Leontinoi. (a) The Biscari head (Archivio Fratelli Alinari, Firenze, 1938). (b) The torso from Leontinoi (photo authors). (c) Photofit of the head and the torso [208].

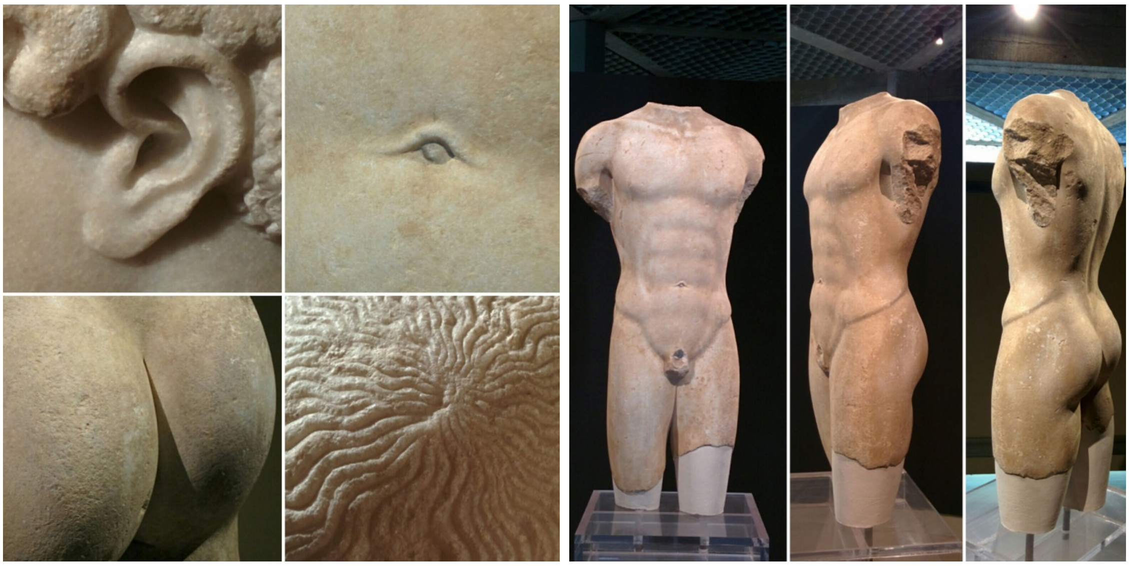
the first intact Sicilian kouros.

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 restoration of the kouros of Leontinoi.

6.5.1 Acquisition and data processing

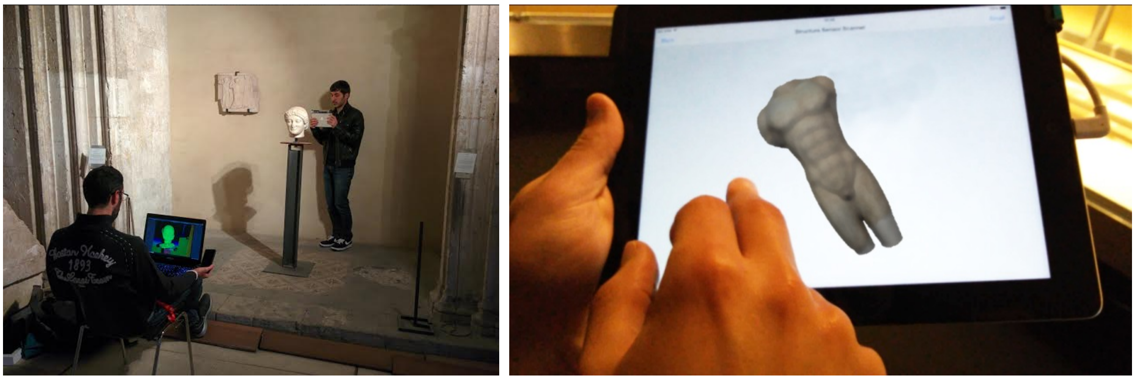
The acquisition was carried out with extreme care in order to properly capture the many anatomical details of the two pieces (Fig. 6.8). The scanning was performed using the Structure sensor connected through Wi-Fi to Skanect in Uplink mode. The scan volume was set to $0.6m^3$ for the head and to $1.2m^3$ for the torso (Fig. 6.9).

Both the artefacts are placed on a pedestal; in particular the head is fixed to the base onto a metal support. After digital acquisition the meshes were pre-processed deleting the vertices extraneous to the ones of the artefacts. The pedestals were cropped out from the acquired models. Then, these 3D models were manipulated with two popular software among archaeologists. Meshlab [182] was employed in



(a) (b)

Figure 6.8: Kouros from Leontinoi. Details of anatomical features of the kouros.

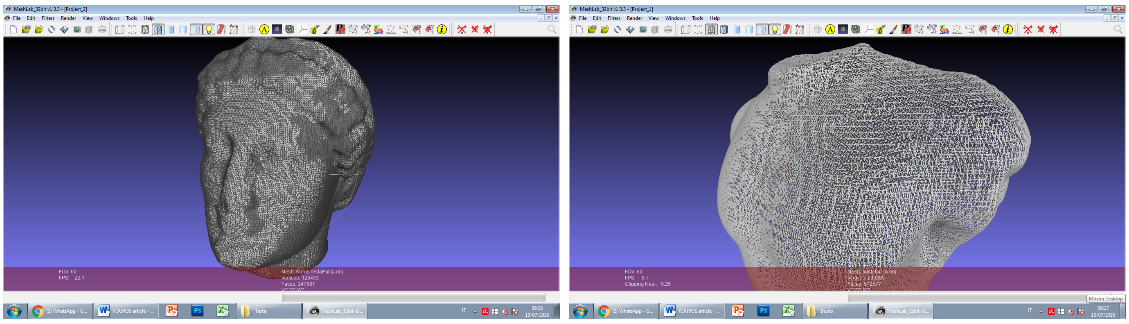


(a) (b)

Figure 6.9: Kouros from Leontinoi. (a) The Biscari head (Archivio Fratelli Alinari, Firenze, 1938). Acquisition of the head at the museum of Catania.

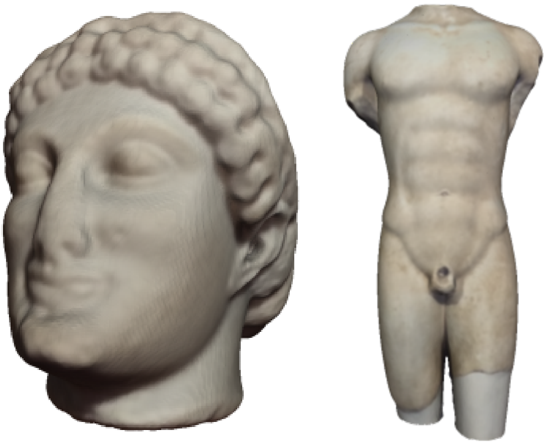
order to refine the models (Fig. 6.10) and after the process of gap filling and polishing results turned out to be rather satisfying (Fig. 6.11).

Subsequently the models were imported into Blender [209] to perform a manual alignment between the head and the torso, obtaining the results showed in Fig. 6.12.



(a) (b)

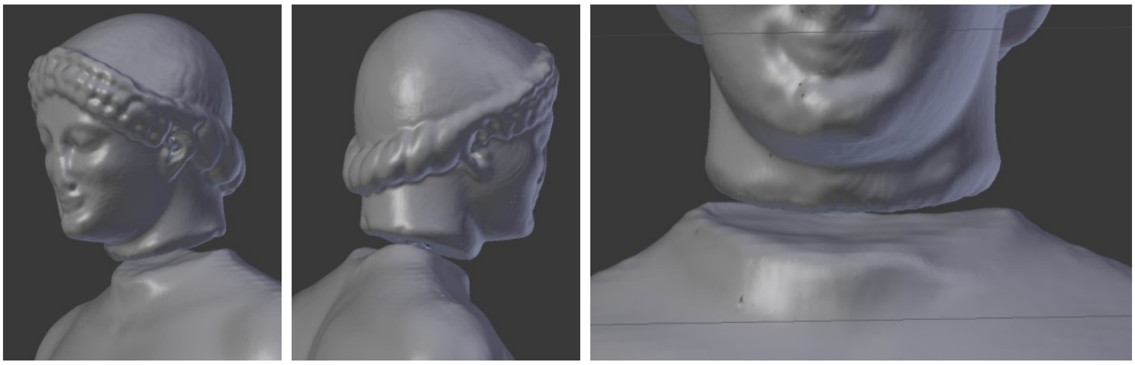
Figure 6.10: Kouros from Leontinoi. Processing on the 3D model of the head in Blender.



(a) (b)

Figure 6.11: Kouros from Leontinoi. Textured 3D model (a) of the head and (b) of the torso.

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 restoration has given back to the community of scholars the first realistic representation of the kouros of Leontinoi, the first life-size statue of Archaic kouros from Greek Sicily. A web platform has been properly arranged in order to share in a simple and effective way the results of this research [161, 210]. The aim of this tool is to provide a high quality visualization of the combined 3D models, linked with related metadata in order to provide an accurate archaeological and



(a) (b) (c)

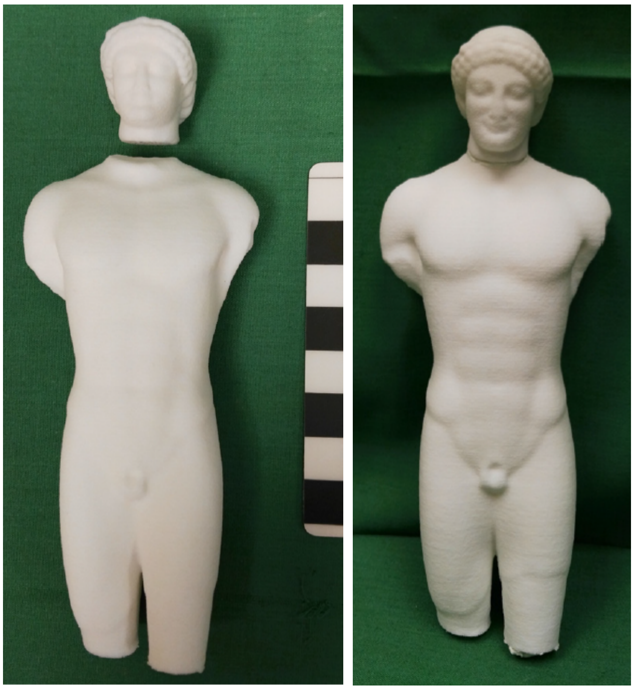
Figure 6.12: Kouros from Leontinoi. Manual alignment of the 3D models of the head and the torso in Blender.

historical context to the artefacts [171, 192]. Another advantage of the use of this web platform is the opportunity to upgrade the versions of the 3D models to monitor the conditions of the artefacts and to involve the community of world wide web users in the discussion [170, 211, 172].

6.5.2 Three-Dimensional Printing

The next step of our research effort was to create a physical copy of the statue in scale 1:10 through 3D printing (Fig. 6.13). 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. Total print time was 24 hours and consumed 170g of polymer. Post print work-up was kept to 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.

To touch a 6th century BC statue or to hold its polymeric replica do not evoke the same emotions, but the opportunity to hold the replica without fear of dropping it, damaging it, manipulating it or using it for its intended purpose, makes it more *authentic*. In fact, “the experiences elicited by touch in this context go beyond, but



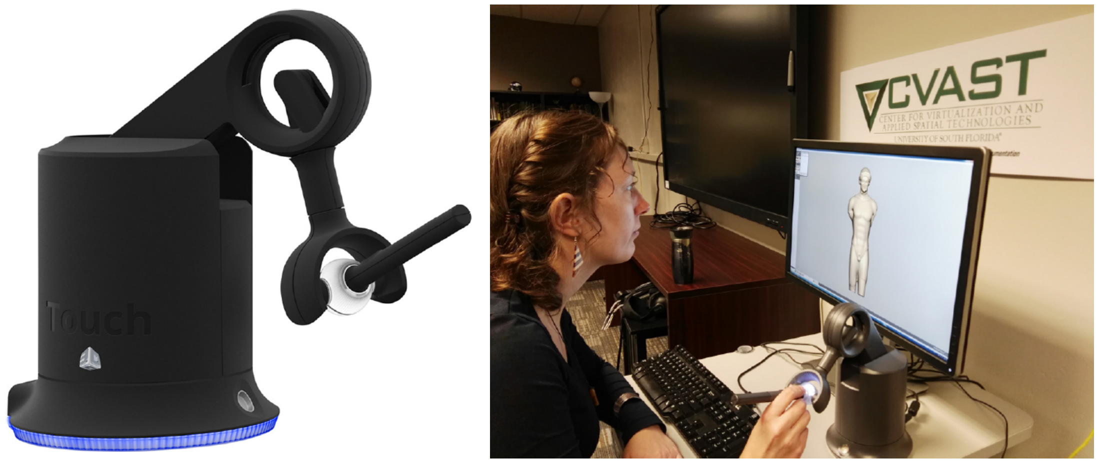
(a) (b)

Figure 6.13: Kouros from Leontinoi. 3D prints.

do not exclude, learning and enjoyment to include deep emotional responses stimulated by object handling" [212]. Although conservation and preservation remains among the top priorities for most museums, the need for the audience to see, touch and to feel the object of their interest must not be underestimated. In addition, it must be considered that "people now inhabit a multimedia world, with all the expectations that this brings and that museums need to become familiar with the languages of these technologies to stay relevant" [213].

6.5.3 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


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

(a)

(b)

Figure 6.14: Kouros from Leontinoi. Haptic technology. (a) Haptic device 3D Systems Touch 3D Stylus. (b) Students interacting with the digital model

the kouros of Leontini will represent another example of enhanced realization for the public with visual impairments. However, the process of 3D printing is still rather 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. Promising results come from the recent research on haptic technologies applied to museum studies may be able to offer an alternative perspective on this issue and show a new way of learning through objects [212].

An experimental test has been undertaken at the Center for Virtualization and Applied Spatial Technologies – CVAST [214] of the University of South Florida, 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. Using the haptic device 3D Systems Touch 3D Stylus (Fig. 6.14(a)) paired up with the proprietary software Geomagic Sculpt, a group of students were asked to interact with the digital model (Fig. 6.14(b)), and then to interact with the 3D print, and finally describe the feedback in a questionnaire (Table 6.3), inspired by the questionnaire designed for the experiment with the

Table 6.3: Kouros from Leontinoi. Questionnaire used to validate the touch interaction with the 3D print and the sensorial experience with the haptic device.

	SD	D	U	A	SA
Interpreting the museum objects through virtual or real tactile experience is a very inclusive approach					
I felt that the touch interaction provided space to add my own interpretation					
I felt more genuine the touch interaction with the 3D print than that with through the haptic device					
With the help of the haptic device, the computer interface seemed to vanish					
After interacting either with the 3D print and the digital models via the haptic device I felt I interacted directly with the statue					
Further Notes:					

SD = strongly disagree, D = disagree, U = undecided,
A = agree, SA = strongly agree

bronze bust of Sophocles at Northlight Gallery of Huddersfield in 2006 [215]. 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 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 the research agenda.

6.6 Conclusion

In this Chapter, we described several real case studies of 3D data analysis for Cultural Heritage: a doorway of the Monastery of Benedettini in Catania [164], the Morgantina Silver Treasure [162] and Kouros from Leontinoi [165].

In Section 6.3, we 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 6.2, 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.

In Section 6.4, we presented the results of a campaign of noninvasive diagnostic analysis (x-ray, UV, and XRF) and a 3D survey on the Morgantina silver treasure to collect useful data for a twofold aim: monitoring the conservation state over time (to check after 4 years) and guaranteeing the virtual visit of the collection during its absence. The acquired 3D models and diagnostic data have been organized for the first time, 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, while the archeometric analyses have made important contributions to the investigation of their composition and the processing techniques utilized in their creation. The ongoing web-oriented platform and the Android app consist of an active tool for managing metadata; it will gradually be implemented through knowledge acquired by specialists and at the same time contribute to the valorization of these archaeological findings to the public. All the findings of the archeometric analyses have been included in these digital platforms to enable scholars to access a lot of information in a fast and practical way. As future works, texture information could also be used as input for a Computer Vision algorithm to automatically detect and classify decorative patterns [216].

Finally, we shown how the research presented in Section 6.5 has clearly demonstrated that the hypothesis suggested in the first place by Libertini [206, 207] was correct: head and torso 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 volumetry. The virtual reassembly has in fact added a further level of information which was not present in the photofit produced by Gentili, the readability of which was also improved with an algorithm [211]. This research has elucidated how 3D scanning, virtual anastylosis and web sharing can contribute to the improvement of museum policies in the field of public outreach, showing how a case of limited accessibility, represented by the state of a Greek statue irremediably divided in pieces between two museums, can become the public's path to virtual discovery of a new masterpiece of Greek sculpture. Furthermore the employment of 3D printing and haptic technology has demonstrated the strategic importance of touch interaction in the learning process of archaeology not just for that public with visual or mental impairments but it has pointed out an alternative way to enhance the access to the cognitive process itself.

Acknowledgment

The authors are grateful to curators of the *Archaeological Museum of Siracusa* for the collaboration and to the director of the *Museo Civico "Castello Ursino"* for the authorization to carry out the 3D scanning of the Biscari head.

Chapter 7

3D Data Analysis for Medical Research

7.1 Introduction

In the last decade, 3D scanners have been employed in architecture, engineering, biology, cultural heritage as well as diagnostic medicine and reconstruction surgery [217, 218, 219, 220, 221, 166, 222, 223, 20]. These devices allow doctors to get a detailed virtual model of a human body. The opportunity to acquire body parts shape, including soft tissues like the female human breast, has motivated our conjunct study with medical specialists in breast reconstruction.

Our main aim is to find a discriminative parametrization of female breast shape (i.e., a small set of parameters to meaningfully describe it). This kind of mathematical representation gives the possibility to easily define accurate metric for breast difference evaluation. This result is very attractive for breast surgeon, since it can be used to develop new tools to assess the symmetry after a breast reconstruction. It could also be an effective strategy to create clear and well-defined breast shape categories.

Currently, the surgeons are routinely used to acquire pictures of patients (a 2D projection of the breast). The only way to evaluate the surgery is still based on a photographic comparison using pictures taken before and after the surgery. Nevertheless, 3D scanners capture and store more information, like volume estimation, curvature and so on. The analysis of 3D data of this kind would enable specialists to plan and asses the surgery in a more accurate way.

The 3D scanner acquisition of human body parts requires a certain time and skills. Long scanning time tends to increase the patient stress as well as the amount of noise due to breath and involuntary micro-movements. Modern hand-held scanners reduce these problems by allowing low acquisition time. Furthermore they guarantee sufficiently high quality of the data. Actually, extremely high resolution and accuracy are pointless to capture general shape. Moreover, dense points clouds would affect the processing time. For these reasons we propose to perform dataset acquisition with a fast and low-cost hand-held 3D scanner: Structure Sensor [167]. High portability of hand-held scanners simplifies the operator job, which can easily turn around the patient.

The 3D data have to be processed and simplified to capture just the information that surgeons need for their analysis. In the approach described in this Chapter, we consider normals orientation to build a compact representation of breast model. Principal Component Analysis (PCA) [224] has been employed to further summarize processed 3D data. PCA is a popular and valuable approach to reduce the high dimensionality of the datasets and capture just the most significant features. Feature reductions through PCA has already been used in the parametrization process of human body parts [225, 226]. Concerning breast shapes, other authors proposed to analyze them using linear measurements, stationary laser scanners, Magnetic Resonance Imaging (MRI), X-rays or thermoplastic moulding [227, 228, 229, 230, 231, 232]. In this Chapter, we present two approaches to 3D breast data analysis and summarization: Thin-Plate Splines (TPS) [20] and Bag-of-Normals (BoN) [21]. In the former, the 3D meshes are projected in 2D space and then TPS [22] is used to estimate the non-linear transformation that change each breast projection in the average one. In the latter, breasts are represented exploiting BoN representation, resulting in descriptor of 64 parameters (64-dimensions, or just 64-d). PCA is computed for both of the approaches and the obtained first 2 principal components are used to plot breasts shape into a 2D space.

Our contribution in the field can be summarized in the following points:

- The acquisition of 3D breast models to build a proper dataset and perform significant experiments. At the best of our knowledge there are not available dataset like this.

- The idea to exploit 3D normals to create a compact representation of 3D breast models.
- Time and cost optimization by employing a hand-held 3D scanner.
- Gain a compact description of a 3D model, easy to be transfered through the web.

The contents of this Chapter are based on our published papers [20, 21]. The structure of the Chapter is the following: in Section 7.2 we quickly recall the Thin-Plate Spline (TPS) method applied in biological and biomedical research field. Employed devices and proposed method are described in Section 7.3. Details on the dataset are provided in subsection 7.3.1. Shape parametrization is described in Section 7.4. PCA is defined in Section 7.5. Results are reported in Section 7.6. Discussion and conclusions end the Chapter in Section 7.7.

7.2 Thin-Plate Spline (TPS) in biomedical investigations

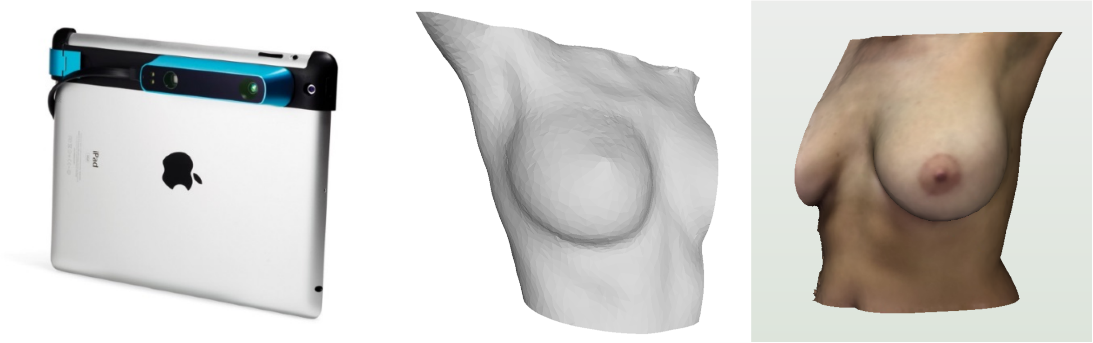
In many biological and biomedical investigations, the most effective way to analyze forms of whole biological organs or organisms is by recording geometric locations of landmark points. Methods of geometric morphometry, based on the analysis of landmark configurations, allow further in-depth investigation of morphological processes [233, 234]. The most widely applied method for landmark-based non-rigid registration is based on Thin-Plate Spline (TPS). TPS provides a mathematical framework to decompose into affine and non-affine components matching between shapes while minimizing a bending energy based on the second derivative of the mapping. This approach has been introduced into medical image registration by Bookstein [235] and has been widely applied in biological, medical and other applications. The method of TPS, allows a quantitative shape analysis taking into account a specific series of morphological variables, the so-called “principal warps”. At the first step, the mean configuration of landmarks is calculated as a reference shape. Then, the mean shape is morphed into the other landmark sets by using TPS. This transformation may be factorized in two parts, namely the affine part, which includes rotation, scaling and translation, and a nonlinear part, which can be obtained as a linear combination of a radial basis function $U(r)$. Function $U(r)$ depends on the dimension of the space where landmarks have been obtained. In 2D, $U(r) = -r^2 lg(r)$, while in 3D $U(r) = |r|$. Bookstein suggested to extract from the linear weights, which define the deformation in terms of $U(r)$'s functions, a special basis, that he has named "Principal Warps Basis" [22]. This basis provides a useful way to understand the nonlinear warping on a given landmark configuration. Researchers have used the space of Principal Warps as a direct parameter space. Usually, dimension reduction of the Principal Warps space is achieved using Principal Component Analysis (PCA), hence we have chose this approach in our method.

7.3 Tools and Dataset Acquisition

The study we conducted is mainly focused on digital shape analysis of breast models to assist breast surgeons in medical and surgical purposes. Our idea is based on three key points: minimally invasive for the patient, use of low cost devices, easy data visualization-&-understanding for people with a medical background.

We employed a 3D scanner with structured infrared light technology that allows us to acquire the information about depth of thousands of points at the same time. The Structure Sensor (Fig. 7.1(a)) is a hand-held scanner proved to be empirically able to acquire up to 12 meters, although it is recommended a distance in the range 0.4 and 3.5 meters [167, 163]. Its maximum accuracy is 0.5 mm, but worsens when the volume of the area scanned is large. Since the scanner uses infrared rays, it is recommended for indoor usage only. The device is calibrated, that means each 3D model will show its real size. The sensor itself is not able to acquire RGB colour mode information (Fig. 7.1(b)), however it is possible to plug into an iPad and uses the tablet camera to this purpose (Fig. 7.1(c)).

To acquire a breast model, we propose a clinical procedure in which the female patients hold hands behind and above the head. In this way, the operator can move around the breast with the Structure Sensor (which is clipped onto the iPad). Although textures have been acquired, these have not been used for the present investigation. An example of the model acquired with Structure Sensor is shown in Fig. 7.1.



(a) (b) (c)

Figure 7.1: (a) Structure Sensor used to acquire breasts 3D models. (b) A sample not post-processed 3D model acquired with Structure Sensor. (c) Example of a textured mesh, as it is acquired by the Structure Sensor.

Once the model is acquired, it is automatically pre-processed through a 3D processing software (Meshlab [182]), in order to remove noise, isolated vertices and faces. Mesh editing is followed by a manual definition, through cropping, of the Region of Interest (ROI). ROI extraction is a critical part of the proposed procedure. We adopt a simple approach that has been proved to be replicable and reasonable precise. We manually selected the ROI exploiting four anatomical reperees suggested by the breast surgeons (Fig 7.2). In our acquisitions, we scanned both left and right breasts but all of them have been, when needed, vertically mirrored in order to make the dataset right-left side invariant, as shown in Fig. 7.2. Note that we considered right breast the one corresponding to the right arm of the patient.

Each model is saved with the standard OBJ format, which describes the information on vertices, faces and face normals. The average number of vertices is $\sim 1,500$, while the average number of faces is $\sim 4,000$. This resolution is not extremely high but it is enough to capture information about breast shape, which is the point of this work.

7.3.1 3D Breast Dataset

After review of the study protocol and formal approval by the internal ethic committee of *Associazione Santantonese per la Lotta ai Tumori (ASLT)* we gathered a dataset with breasts acquired from different volunteers, aged between 25 and 65,

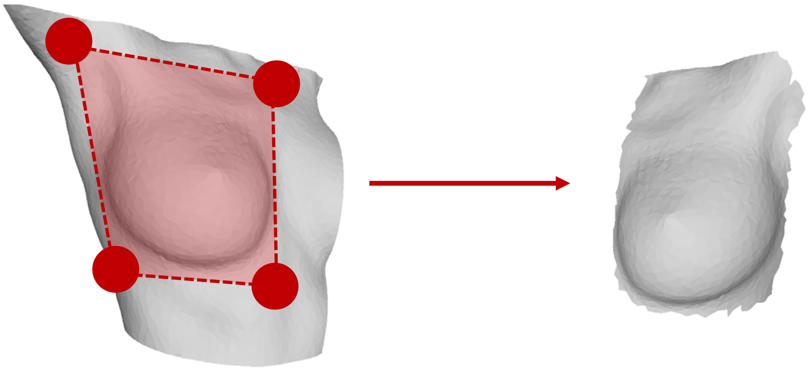


Figure 7.2: Definition of the Region of Interest (ROI) through 4 anatomical reperees suggested by the breast surgeons. After the cropping is performed, the right breasts are vertically mirrored.

with different shapes and volumes. Breast surgeons put a label on each model, describing size and ptosis of the breast. The severity of ptosis is characterized by evaluating the position of the nipple relatively to the infra-mammary fold. Supervised by doctors, we created a dataset in the following way:

- Main Dataset: is made up of 31 breasts, 17 left and 14 right. To guarantee a proper dataset variability, we have included breasts of different size and ptosis.

Then, in order to test the strenght of the proposed methodology, we selected a patient and acquired her breast several times in pre-operation and post-operation conditions. Hence, two more groups of meshes is distinguishable:

- Group 1: is made by 52 meshes, 26 left and 26 right. Notice that this set of meshes has been acquired by two operators, namely a junior and an expert one, so it can be used to investigate how the proficiency of the operator may change the parameterization.

- Group 2: is made by 16 breasts, 8 left and 8 right.

Hence, in our experiments we gathered and employed a total of 99 breast models.

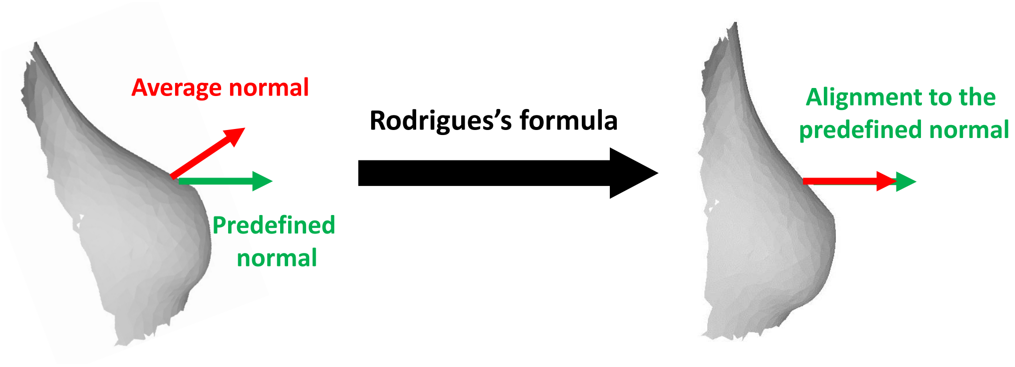


Figure 7.3: Alignment process using Rodrigues's formula [236].

7.4 Shape Parametrization

In this subsection, we define the method employed to process the 3D models, in order to parametrize the breast shape with a minimum number of parameters. Although the acquisition device is calibrated, it has no system to get the correct orientation into the real space (e.g., gravity sensors). Hence, the meshes have to be initially oriented along the same direction and its centroid moved to the origin of a 3D Cartesian coordinate system [20]. Subsequently, the average normal is computed and we find the rotation matrix, in order to align the average normal along the Z axis. We use the unit vector $(0, 0, 1)$ as reference in this procedure. Finally, to get the rotation matrix, a closed form named Rodrigues's rotation formula [236] is employed (Fig. 7.3). Specifically, given two vectors u_1 and u_2, the formula computes the rotation to align u_1 to u_2. In our case $u_1 = averageNormal$ and $u_2 = (0, 0, 1)$.

7.4.1 Shape Parametrization for TPS Approach

After breast registration, the 3D mesh is projected to a 2D space along the Z axis of the coordinate system. This leads to a grayscale depth-map representation. Each breast depth-map is quantized to extract seventeen landmarks which are used as reference points to apply the TPS algorithm. Finally, the transformation matrix obtained by TPS is used in Principal Component Analysis (PCA) to identify the most significant warps. Main steps are detailed in the following.

Breast Projection

Vertices coordinates are linearly normalized in $[0, 1]$ exploiting the global minimum (minimum coordinates among all the models) and the global maximum for both the components X and Y. In this way, the scale variability along X and Y axes for our models is preserved. The Z component is also normalized in $[0, 1]$: in this case, local minimum of the Z component for each model is translated to 0 and the global maximum is mapped to 1. This guarantees that at least one vertex for each mesh has coordinate value $Z = 0$. The alignment process is graphically summarized in Fig. 7.3 .

After that alignment and normalization have been performed, the model is quantized and the mesh is projected on the X-Y plane (frontal view). We quantized X and Y into 150 levels, in order to obtain depth-maps with a 150×150 resolution. For each point in X-Y, the Z component represents the depth value. Since folds may happen or more than 2 measured points may fall above the same quantized X-Y point, then the maximum Z value is chosen. In general, gap may arise (see Fig. 7.4(a)). Hence, as a last step, grayscale morphological closure [237] is performed to fill the empty regions inside the breast area (Fig. 7.4(b)). We used a circle with a diameter equals to 5 pixels as structural element for grayscale morphological closure, and then we applied a Gaussian smoothing with a kernel of size 3×3 pixels and standard deviation equals to 1 to regularize the data.

Landmarks extraction

TPS is able to compute a non-linear transformation to fit a set of landmarks to a reference set. Our idea is to use TPS to estimate the deformation to apply to a "standard" breast to morph it into another one. We need to extract a fixed number of landmarks from each one of the 3D models, and to fix a set of landmarks for a reference breast shape, to successfully apply TPS.

Landmarks extraction is performed by linearly normalizing each of the depth-maps in $[0, 1]$ and quantizing the Z values into 3 levels. Landmarks are extracted as follows: the first landmark is the maximum value of the original depth-map. Notice that this point may or may not correspond to the nipple. We consider the 8 lines which pass for the first landmark along 8 fixed directions $(0°, 45°, ..., 315°)$. The second set of 8 landmarks is identified by the crossing between the same 8 lines and

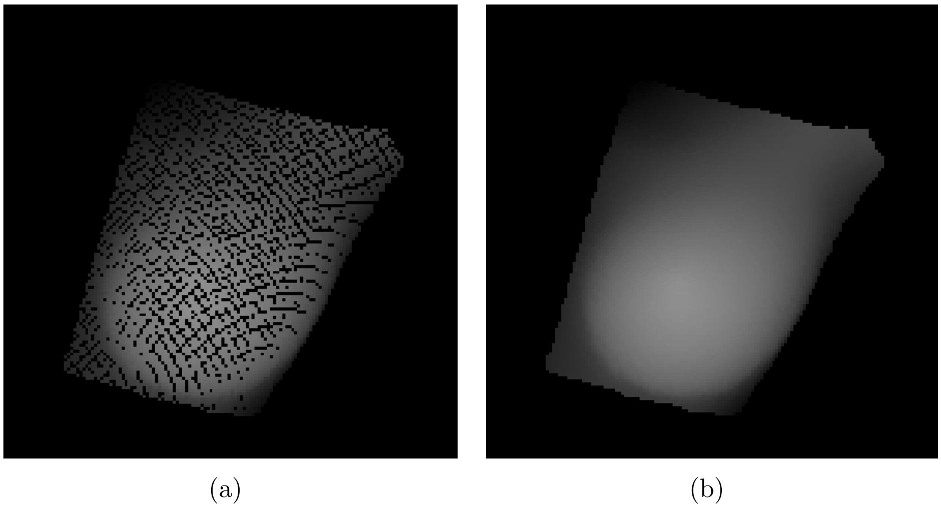

(a) (b)

Figure 7.4: (a) The projected 3D mesh. The empty regions in (x, y) point indicates that no vertices have been found there. (b) The depth-map (a) after the application of closure morphological operator.

the highest quantization level boundary. Finally, the last 8 landmarks are chosen accordingly with the previous strategy, but considering the second quantization level. Eventually, 17 landmarks are extracted. An example of landmarks extraction is shown in Fig. 7.5.

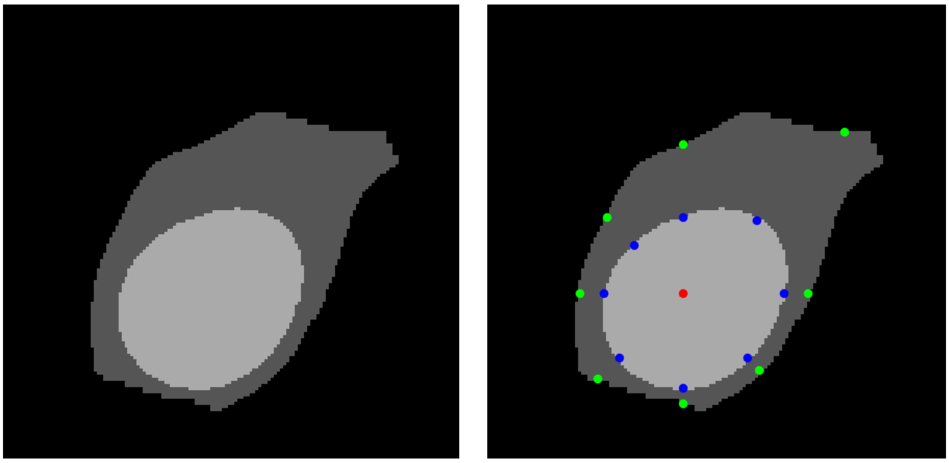
The set of reference landmarks is learnt by computing the average of the 41 sets of 17 landmarks.

Thin Plate Spline (TPS) Mathematical Definition

In this subsection, we quickly recall the mathematical theory of TPS, as introduced by Bookstein [22]. Let P and Q be two sets of points in the plane, such that:

$$P = \{p_i = (x_i, y_i), i = 1, 2, \dots, n\}$$
$$Q = \{q_i = (x_i^{'}, y_i^{'}), i = 1, 2, \dots, n\}$$

We are looking for a transformation $F : R^2 \rightarrow R^2$ such that $F(p_i) = q_i$ for $i = 1, 2, \dots, n$. This problem can be broken down into two interpolation problems



(a) (b)

Figure 7.5: (a) The 3-levels quantized depth-map of Fig. 7.4(b). (b) The landmarks extracted from depth-map of Fig. 7.4(b); red landmark is the maximum value of the original depth-map, blue and green landmarks are respectively the landmarks related to the first and the second quantized levels.

$f_x(p_i) = x'_i$ and $f_y(p_i) = y'_i$ in x and y directions, respectively.The TPS model for one direction interpolation function is

$$f(x, y) = a_0 + a_x x + a_y y + \sum_{i=1}^{n} w_i U(|p_i - (x, y)|) \tag{7.1}$$

where $U(r) = -r^2 \log(r)$ with $r = \sqrt{x^2 + y^2}$ is the basis function, $A = (a, a_x, a_y)$ defines the affine part and $W = (w_1, w_2, \cdots, w_n)$ describes an additional non-linear deformation.

Now, let

$$L = \left[\begin{array}{c|c} K & P \\ \hline P^T & O \end{array}\right],$$

where

$$K = \begin{bmatrix} 0 & U(r_{12}) & \cdots & U(r_{1n}) \\ U(r_{12}) & 0 & \cdots & U(r_{2n}) \\ \vdots & \vdots & \vdots & \vdots \\ U(r_{1n}) & U(r_{2n}) & \cdots & 0 \end{bmatrix} \qquad P = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{bmatrix}$$

$r_{ij} = \sqrt{|p_i - p_j|}$ and O is 3×3 matrix of zeros. Let $V = (x'_1, x'_2, \cdots, x'_n, 0, 0, 0)^T$, then the coefficient vector $\widetilde{W} = (w_1, w_2, \cdots, w_n, a, a_x, a_y)^T$ could be calculated by $\widetilde{W} = L^{-1}V$. The function f minimizes the nonnegative quantity

$$I_f = \int\int \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dxdy$$

which is called "bending energy". Deformation through the Y direction could be achieved similarly.

In our algorithm, TPS has been employed to compute non-linear transformation of a set of 17 landmarks into the set of reference ones. This allows to estimate the shape of a breast 3D model, where the aforementioned transformation matrix is used as descriptor. The affine description of the transformation takes into account issues about scale and rotations. For the clinical application scope, a descriptor of this kind is a relevant parameter that may be taken into account. The principal aim of this Chapter, however, is not concerned with size or volume, but wishes to address the shape variability of the female breast independently from the scale. For this reason, the successive analysis uses only the coefficients of the non linear shape warping.

Principal Warps

Bookstein in [22] suggests that a more meaningful analysis can be done using not the w_i's from Equation 7.1, but their combination along the so called "Principal Warps" that could be found as $P = WS$, where $W = (w_1, w_2, \cdots, w_n)$ and S is a $n \times n$ matrix that contains eigenvectors of matrix $L_n^{-1} K L_n^{-1}$ and L_n^{-1} is the upper left $n \times n$ subblock matrix of L^{-1} in Equation 7.1. Moreover, since W is a $2 \times n$ matrix then P results into a $2 \times n$ matrix.

The principal warps formalism allows to read the original data in term of a result of a set of relevant principal deformations along directions that are prescribed by the intrinsic geometry of the reference landmarks set. Unfortunately, the dimension of the principal warps is equal to the dimension of the original data. So, it may not be useful for clinicians when the original data has a high dimension. In our case, the dimension would be $2 \times 17 = 34$. Since it is not easy to analyze data with high

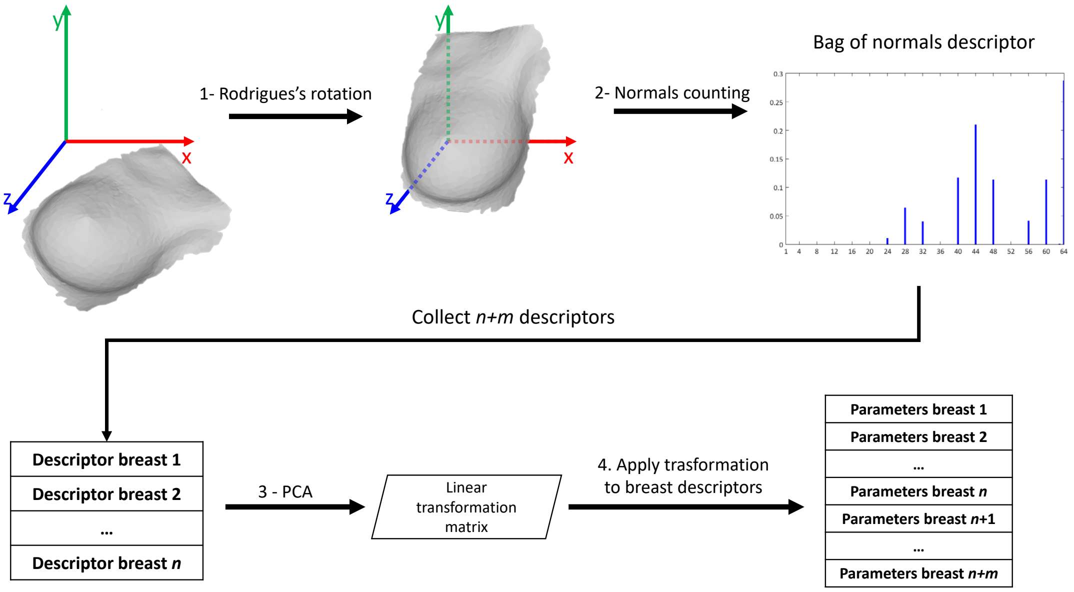


Figure 7.6: Pipeline of the proposed method. Note that PCA is applied on n breast descriptors. Then, the "learnt" transformation matrix is used as model to extract parameters of all the 3D meshes. Additional details are reported in Section 7.6.

dimension, we reduce their dimension applying the principal component analysis, as it is explained in Section 7.5.1.

7.4.2 Shape Parametrization for BoN Approach

In the Bag-of-Normal (BoN) approach, each 3D model is represented as histograms of normals. Since normal vectors define the orientation of each model vertex/face, the BoN-based algorithm starts with the initial alignment step. Then, the normal space is clustered and the occurrences for each cluster counted. This descriptor is finally reduced by PCA. The summarized pipeline of proposed method is shown in Fig. 7.6.

Once each mesh has been correctly oriented in our coordinate system, than we can proceed to obtain a representation of the normals distribution over a suitably quantized grid. Firstly, all the normals are normalized [1]. We divided each normal

[1] This is not a pun. *Normalized normals* are granted to be unit vectors.

u for $||u||$, in order to get a unit vector. By performing this process, the three components of normal vector (u_x, u_y, u_z) fall in the range $[-1, 1]$. We linearly quantize the space of each component into 4 levels, in order to obtain $4 \times 4 \times 4 = 64$ different clusters. Finally, each mesh is represented by counting the occurrences in each cluster. Then, this histogram with 64 bins is, in turn, normalized to get the final BoN descriptor.

7.5 Principal Component Analysis (PCA)

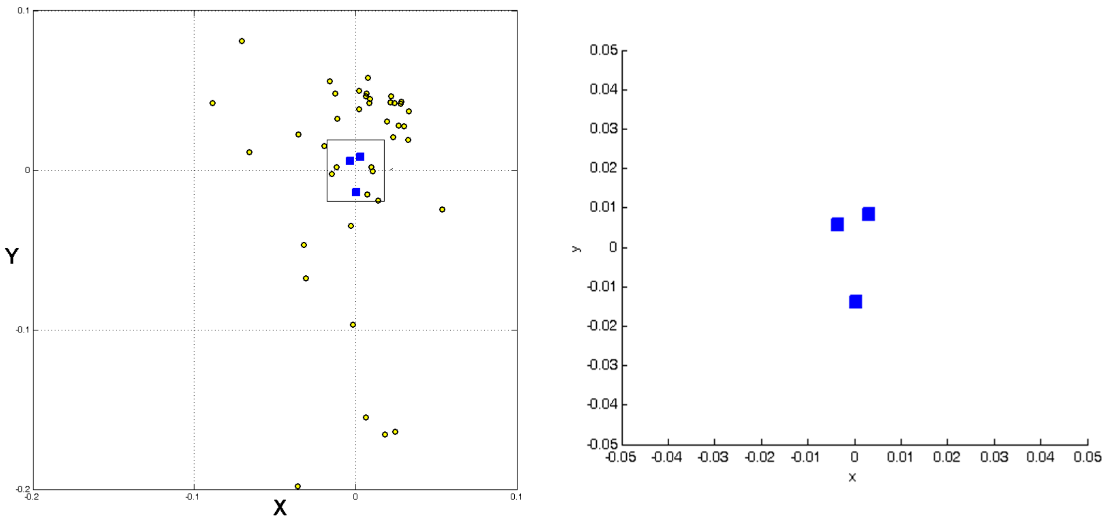
PCA is a popular statistical method that is commonly used for finding patterns in data of high dimension or reducing such dimensionality. This reduction is more interesting when one wants to extract the main characteristics from complex data. PCA is applied on datasets described by several attributes. It is able to find a linear transformation, which moves the data into another space where the transformed attributes are uncorrelated. The aim is to identify the "Principal Components" (a reduced set of attributes which represent the original data) [224].

7.5.1 PCA for TPS Approach

We applied PCA on Principal Warps of our dataset for TPS approach, which counts $m = 41$ observations (breasts) and $n = 34$ dimensions. The two first principal components that have largest variances can be easily extracted by seeking the largest eigenvalues. The amount of variance retained using only two principal components is 93%. In particular, the first one accounts for 78% and the second one for 15%.

7.5.2 PCA for BoN Approach

We applied PCA on the 64-d BoN descriptors, in order to describe each 3D breast with a very small set of parameters, namely 2. This procedure allows us to represent each 3D model as a point in 2D coordinate system, where axes are the first two Principal Components. This kind of representation, visually asses the results of a surgery intervention by observing the change of position of a breast in the 2D space. It Section 7.6.2 we report the results that show that 2 components are enough to represent breast shape.



(a) (b)

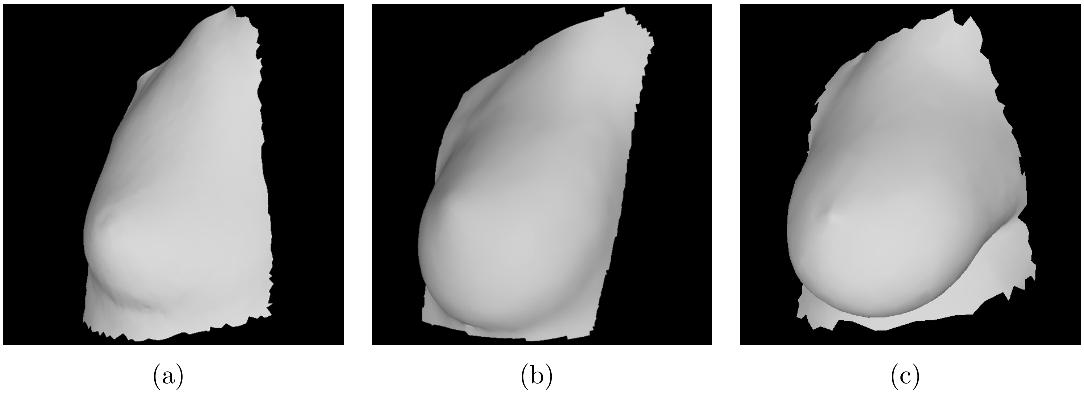
Figure 7.7: TPS approach. (a) Breasts landmark 2D representation. Axis are the two first principal components of principal warps. Yellow circle markers are the 41 observations of our dataset. (b) blue square markers are the representation of the inter-operator test.

7.6 Results

7.6.1 Results for TPS Approach

After the PCA application on the Principal Warps, we obtain two principal components, hence each breast can be identified by two coordinates. We plot each breast in a 2D space for a proper visual analysis (Fig. 7.7(a)). The breast specialist has carefully assessed the meaning of this parametrization, and come to the following asserts: the second principal component, plotted along X axis, seems not particulary meaningful, whilst the first one (along Y axis) captures size and ptosis variability of the breast. Smaller breasts with low level of ptosis tend to have lower Y values, similarly larger breasts with high level of ptosis assume the higher values. In Fig. 7.8 the breasts landmark visualization for minimum and maximum Y value (ptosis) are shown together with an intermediate one.

To validate the framework we conduct also an inter-operator experiment: three different operators have acquired the breast of the same patient. Our expectation is that the three new meshes will be represented close together in the parameter



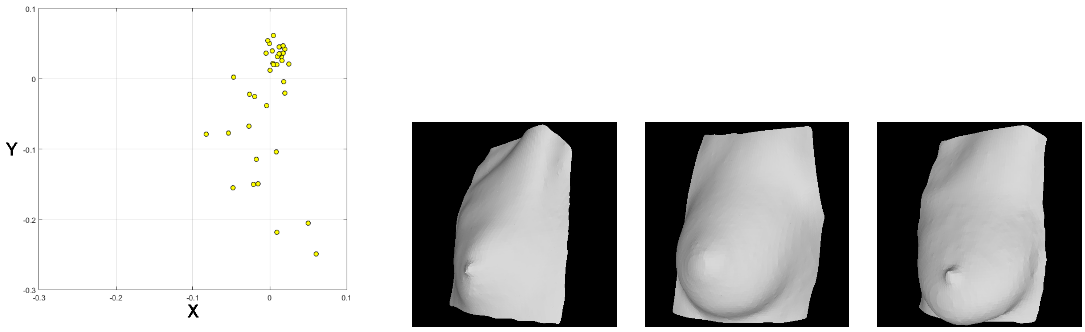
(a) (b) (c)

Figure 7.8: TPS approach. Coordinates are related to Fig. 7.7.(a) Breast mesh related to the minimum Y value (very low ptosis level), $y = -0.198$. (b) An intermediate breast, $y = -0.047$. (c) Breast mesh related to the maximum Y value (very high ptosis level), $y = 0.081$.

space. With this test, we want to demonstrate the small inter operator-invariance of the descriptors. We projected the 3D meshes in 2D by using the proposed breast projection method described in Section 7.4.1. In the normalization step along Z axis, we exploit the global maximum computed on proposed dataset. Then, we extract the principal components of these acquisitions accordingly to the PCA model previously built. Results are shown in Fig. 7.7(b): although the landmark representations are not exactly the same, due to noise introduced by involuntary micro-movements and cropping, the final representation in the 2D space is sufficiently quite similar, as expected. To further confirm the proposed method we have acquired a second dataset with 40 new breasts models and we have processed them by using the model learnt with the first dataset. Results of this experiment have shown in Fig. 7.9.

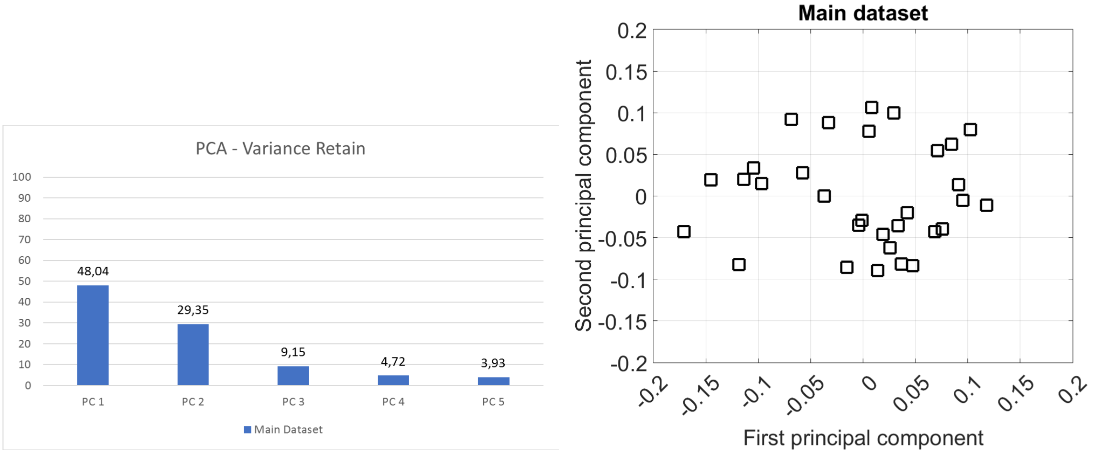
7.6.2 Results for BoN Approach

We computed PCA on the 31 models in the main dataset. Exploiting only the first 2 principal components we obtained a variance retain of $48.04 + 29.35 = 77.39\%$ (Fig. 7.10(a)) and models can be represented in a chart, as shown in Fig. 7.10(b). Breast surgeons confirmed us the evidence of Fig. 7.10(b): the first 2 principal components seem enough to distinguish characteristic traits of the labelled models, since models are clearly separated in the obtained result. However, since there are



(a) (b) (c) (d)

Figure 7.9: TPS approach. (a) Breasts landmark 2D representation of the second dataset. Axis are the two first principal components of principal warps. Yellow circle markers are the 40 observations of our dataset. (b) Breast mesh related to the minimum Y value (very low ptosis level), $y = -0.249$. (c) An intermediate breast, $y = -0.067$. (d) Breast mesh related to the maximum Y value (very high ptosis level), $y = 0.061$.



(a) (b)

Figure 7.10: BoN approach. PCA computed on the Main Dataset. (a) Variance Retain of the first 5 principal components. The sum of the first 2 principal components is 77.39%. (b) Plot of the 31 models in the Main Dataset using the first 2 principal components.

not official metrics to describe breast shape, we currently cannot associate each component to a specific geometrical property.

We plotted the models of Group 1, exploiting the PCA computed only on the main dataset (Fig. 7.11(a)), in order to further assess the soundness of the proposed

Table 7.1: BoN approach. Mean and Standard Deviation of models in Group 1 and 2. L stands for Left, R for Right. Each entry is a pair in which the values are related to the first and second principal component, respectively.

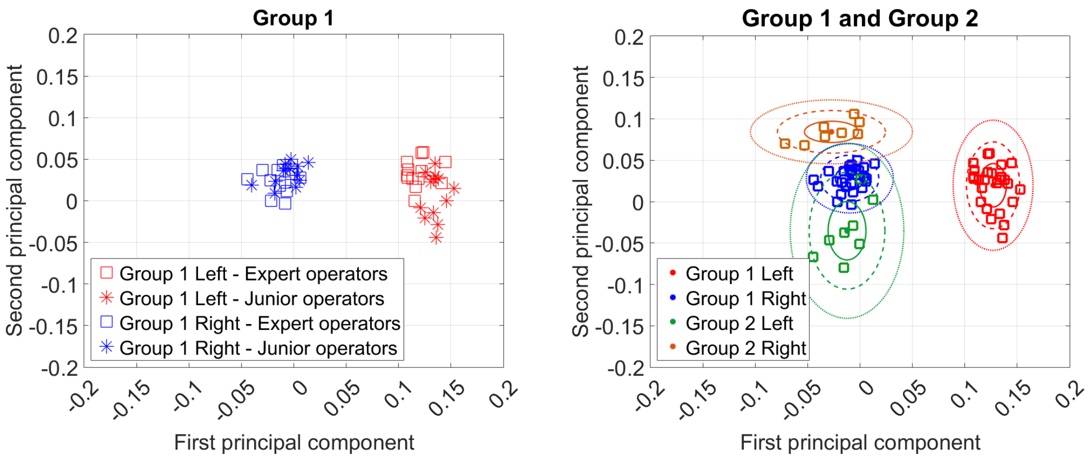
	Group 1		Group 2	
	L	R	L	R
Mean	(0.1269, 0.0198)	(−0.0098, 0.0281)	(−0.0124, −0.0352)	(−0.0273, 0.0843)
Stand. Dev.	(0.0127, 0.0262)	(0.0138, 0.0139)	(0.0181, 0.0353)	(0.0258, 0.0129)

method. The left breast is clearly distinguishable from the right one, as expected. Once more, using the same principal components, we plotted also models from Group 2 (Fig. 7.11(b)). We remark that 3D models in Groups 1 and 2 include the right and left breast of the same patient, before and after a surgery, respectively. Mean and standard deviation of models in Groups 1 and 2 have been reported in Table 7.1. Error ellipses including the 68% (σ), 95% (2σ) and 99% (3σ) of the data are contextually shown in Fig. 7.11(b).

Clusters positions of left and right breasts between Group 1 and Group 2 are clearly changed. Accordingly to the PCA variance retain values, the weigthed distances for Group 1 and Group 2 are $\sqrt{(0.136 \cdot 0.48)^2 + (0.008 \cdot 0.29)^2} = 0.0653$ and $\sqrt{(0.014 \cdot 0.48)^2 + (0.119 \cdot 0.29)^2} = 0.0352$, respectively. These results confirm that, after the surgeon, the two breasts become more similar, mainly proved by the minor distance among the first principal component.

So, results shown in Table 7.1 and Fig. 7.11(b) are a confirmation that right left breast, after the surgery (meshes from Group 2), have now first principal components that have pretty similar mean and variance values, while before the surgery (Group 1) they were different.

The comparative chart with components of all the digitized breasts is shown in Fig. 7.12. Some significant cases from Main Dataset are shown in Figs. 7.13(a), 7.13(b), 7.13(c), while the patient scanned in Groups 1 and 2 is shown in Figs. 7.13(d) and 7.13(e). The breast surgeons confirmed us that the positions of models from these latter sets are coherent with respect to the one of models from Main Dataset. These results show that the first principal component is strong enough to characterize the shape of a breast, and through standard deviation computations on Group 1 and 2 we can also give a cue about the error in this estimation.



(a) (b)

Figure 7.11: BoN approach. Plots of models in Group 1 and Group 2 using the first 2 principal components of PCA computed on the Main Dataset. (a) Visual comparison of the principal components of Group 1 between models acquired by the two groups of operators, properly juniors and experts. (b) Comparison of the principal components between Group 1 (pre-surgery) and 2 (post-surgery). Error in the parametrization has been highlighted through error ellipses added on each set of models. Starting from the ellipsis centroid (the mean value of the set), each concentric error ellipsis contains the 68% (σ), the 95% (2σ) and the 99% (3σ) of the elements, respectively.

7.7 Conclusions

In this Chapter, we have focused on digital shape analysis of breast models to assist breast specialists for medical and surgical purposes. We fixed three key points for our proposed solution: minimally invasive for the patient approach, use of low cost devices, easy data visualization-&-understanding for people with a medical background. We proposed a clinical procedure in which the female patients hold the hands behind and above the head, while an operator can digitize her breast with a 3D scanner. After a manual ROI definition through cropping, meshes are automatically processed. We presented two approaches to 3D breast data analysis and summarization: Thin-Plate Splines (TPS) [20] and Bag-of-Normals (BoN) [21]. A reference dataset has been used to compute PCA on a set of discriminative and different breasts, and the obtained first 2 principal components have been used to plot the breasts into a 2D space. We empirically proved that breasts subjected

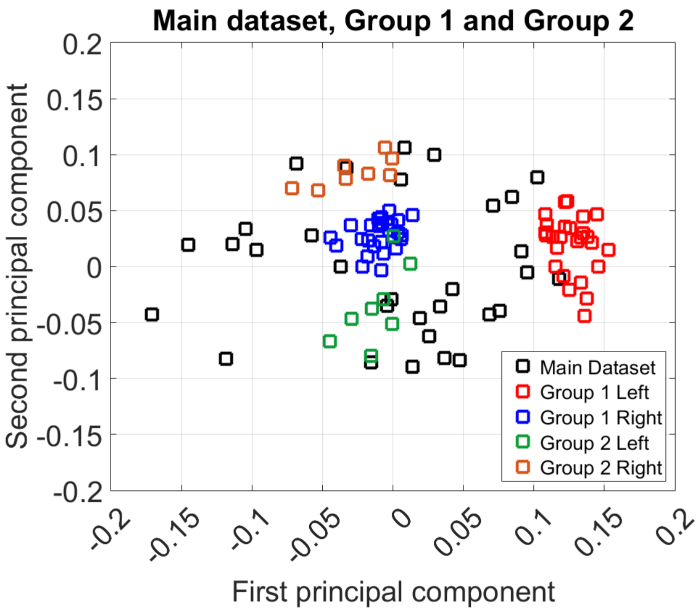


Figure 7.12: BoN approach. Comparison of the first 2 principal components (X and Y axis, respectively) between different datasets. PCA computed on the main dataset, comparison between the main dataset, Group 1 and Group 2.

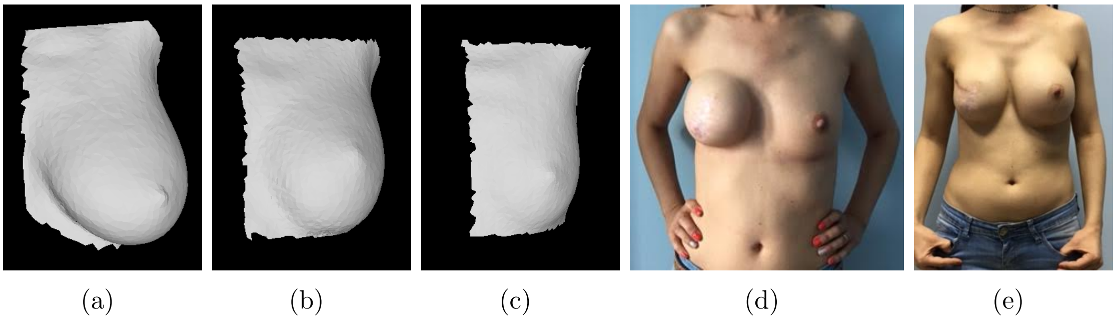


Figure 7.13: BoN approach. Coordinates are related to Fig. 7.10(b). Significant acquired models. (a-c) Models from Main Dataset with principal components $(-0.17; -0.04)$, $(0; -0.02)$ and $(0.11; -0.01)$, respectively. They are in the most left, central e right position of the plot of Fig. 7.10(b). A clear difference about the shape of the breasts can be noticed. (d-e) Patient of Group 1 (pre-surgery) and Group 2 (post-surgery), respectively; note that we considered right breast the one corresponding to the right arm of the patient.

to a surgery change their representation in this space, and through the variance computations on Group 1 and 2 we also gave a cue about the error in this estimation.

We believe that the proposed procedure, assessed by the surgeon, represents a valid solution to evaluate results of surgeries, since one of the most important goal of surgeons is to symmetrically reconstruct breasts, but an objective tool to measure the result is currently missing. Moreover, thanks to the parametrization of 3D breast meshes, we are able to represent 3D data with just 2 parameters, gaining a compact descriptor easier to be transfered and browsed through hospital networks, if compared to a complex 3D mesh. As future works, we planned to augment the ROI extraction phase, which is a critical part of the proposed procedure and requires professionals with a proper know-how of 3D object editing.

Acknowledgment

The authors would like to thank the *Azienda Ospedaliera Cannizzaro*, the *Associazione Santantonese per la lotta ai tumori (ASLT)* and the female volunteers for their contribution as models.

Conclusion and Appendices

Chapter 8

Dissertation Final Discussion, Remarks and Future Works

This dissertation collected all the research work done by the PhD candidate in the Joint Open Lab for Wireless Applications in multi-deVice Ecosystems (JOL WAVE) of TIM Telecom Italia, which sponsored his doctoral fellowship. With this dissertation we wanted to contribute to exponential growth of LTE-based multimedial services experienced in the last decade. We presented real use-cases in which a big amount of multimedia data is analyzed and summarized. These applications may be the enabling technology for LTE-based multimedial services. Three main categories of media have been treated in this dissertation: images, videos, and 3D data. For images and videos we realized two frameworks: The Social Picture and RECfusion.

In The Social Picture (TSP) [3, 6] an huge amount of crowdsourced social images can be collected and explored. With this purpose, TSP represents a well suited enabling technology for LTE networks. The framework embeds a number of advanced Computer Vision algorithms, able to capture the visual content of images and organize them in a semantic way. We employed VisualSFM (VSFM) to compute a 3D sparse reconstruction of a collection within TSP. Starting from the NVM file computed by VSFM, we were able to build two important relationships: the one between cameras and points and the one between cameras themselves. Using these relationships, we implemented two advanced Image Analysis applications. Through the scene summarization, we were able to create a video with a set of canonical images representing the visual content of a selected collection. This kind of summarizing-video is hardly evaluable with an objective metric. Instead, it is usually evaluated with a subjective consensus. We defined a score for each image in the scene and stated how it may be exploited to discard the most unrelated images during the traversal. We demanded the assessment of a valid threshold for score values to a wider future experimentations. Then, we shown how density-maps can be used together with the Structure-from-Motion (SfM) technique to highlight parts of the image with robust visual features. Several types of density-maps have been defined with different aims. Eventually, we shown that Social-Weigthed-Density (SWD) maps represent a good tool to stress the presence of visual features even when a strong occlusion is present in the image. Currently, SWD-maps shown in the dissertation have not been publicly released yet, but they could be shared under request.

RECfusion [7, 8, 9] is a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Given a set of video streams as input, RECfusion grouped video streams by means of similarity and popularity, then it automatically suggests a video stream to be used as output, acting like a “virtual director”. With our best approach described in [9], RECfusion reaches very good results employing vote-based procedure, which is totally automatic and independently by a hyperparameter fine tuning phase. However, the vote-based procedure has a tradeoff: it is unable to create and track an unlimited number of clusters. Once video streams have been tracked, users are able to extract all scenes coherent and with the same context from the collection of video. RECfusion represents an enabling technology for LTE networks, since videos are usually very huge files to be transfered on network and also because the suggested real use-case is suitable for real-time applications. As future works and possible applications, we are planning to augment the framework with features specifically focused on Assistive Technology or Security issues (i.e., highlight/track bad behaviour in the life style, log the visited places, search something or someone that appears in the scene).

Through TSP and RECfusion frameworks the following topics have been discussed: Image Matching and Saliency Estimation, Video and Scene Summarization, . We shown use-cases of multi-device images and videos analysis and summarization tasks.

Focusing on the images, we treated in detail image matching employing advanced techniques (e.g., Compact Descriptors for Visual Search - CDVS). We detailed Content Based Image Retrieval (CBIR) methods, and we described how to compute the heatmap, which represent a valuable tool for analysis and summarization of large images collections. We compared 4 CBIR descriptors: Speeded-Up Robust Features (SURF) [63], Maximally Stable Extremal Regions (MSER) [58], Scale Invariant Feature Transform (SIFT) [61] from VLFeat library [70] and the Compact Descriptors for Visual Search (CDVS [11]) implementation by Telecom Italia R&D team [52]. We shown how our approach described in [3], in which CDVS is improved with Back-Projection Verification and Query Expansion, outperforms other tested method. We employed heatmaps in TSP, with an incremental approach: once image collection is created, then heatmap is computed for the first time. From that moment on, new images added to the collection are compared and matched only with images in the Query Expansion structure. This is a faster and incremental way for heatmap updating.

Social-Weigthed-Density (SWD) maps have been used for reach a novel definition of a saliency model we that we named *Social Saliency*. We named our saliency metric as "social" because the model employed to learn and understand what are the salient parts of an image is based on a crowd-sourced consensus. The model of attention is obtained querying image databases of social media (e.g., Flickr, Instagram, Panoramio), that definitely represent a "social" environment. We tested two methods to learn the saliency model: one is based on Support Vector Regression (SVR) and the other one on a Counting Convolutional Neural Network (C-CNN) named *Hydra C-CNN* [92]. We compared results of linear and non-linear SVR with the ones obtained by Hydra CNN. We found that non-linear SVR obtained higher Pearson Correlation Coefficient (COR) and Area Under Curve (AUC) values than Hydra model. Linear SVR gained the lowest COR and AUC values, and its saliency maps are too sparse and segmented. The Gaussian SVR also generated segmented maps, but they are more similar to the corresponding ground truth SWD-maps. Although Hydra model reached lower COR and AUC values compared to non-linear SVR, saliency maps obtained with this former method seems to be able to highlight salient parts better than the other two SVR models, and are very similar to Salient Object Segmentation (SOS) maps.

A deeper experimentation with the Social Saliency is needed, in order to increase the soundness of this novel method. More effort in standardization of our dataset and employment of experimentally meaningful tools, like the metric evaluation code of MIT, will be put in the future updates of the system. Hydra-based architecture is the most promising one. We are planning to acquire images collections with different environments and contexts, in order to allow the learning of a more generic visual attention model and reduce the overfitting coming from a single collection.

In the third part of this dissertation, we treated 3D data media. We report in the appendices a benchmark of the 3D web viewers currently available. This kind of viewers is an enabling technology to allow online fruition of 3D contents for as many users and scholars as possible. We described real use-cases in two different contexts: Cultural Heritage [166, 163, 161, 162, 164, 165] and Medical [20, 21].

We addressed the issue of Cultural Heritage digitization through 3D scanning (Digital Archaeology [15]). This topic is important for preservation and conservation of Cultural Heritage sites. We described several real case studies of 3D data analysis for Cultural Heritage: a doorway of the Monastery of Benedettini in Catania [164], the Morgantina Silver Treasure [162] and Kouros from Leontinoi [165]. Through these test cases, we gained several outcomes. We demonstrated that low cost hand-held sensor can obtain high quality 3D models of architectural details, useful to integrate expensive Time-of-Flight scans and make the digitalization of the cultural heritage easier and faster with affordable economical efforts. The digitized models and diagnostic data have been organized for the first time, in an integrated way, within a web-oriented platform and an Android application. We specifically designed and implemented both platform and application to increase the existing archaeological knowledge and to obtain referenced information of the conservation state. They consist of an active tool for managing metadata; it will gradually be implemented through knowledge acquired by specialists and at the same time contribute to the valorization of these archaeological findings to the public. All the findings of the archeometric analyses have been included in these digital platforms to enable scholars to access a lot of information in a fast and practical way. As future works, texture information could also be used as input for a Computer Vision algorithm to automatically detect and classify decorative patterns. We shown how 3D scanning and web sharing can contribute to the improvement of museum policies in the field of public outreach, showing how artifacts with limited accessibility (i.e., the Kourous from Leontinoi) can become the public's path to virtual discovery of a new masterpiece of Greek sculpture. Furthermore, the employment of 3D printing and haptic technology has demonstrated the strategic importance of touch interaction in the learning process of archeology not just for that public with visual or mental impairments but it has pointed out an alternative way to enhance the access to the cognitive process itself.

As regards the Medical context, we have focused on digital shape analysis of breast models to assist breast specialists for medical and surgical purposes. We fixed three key points for our proposed solution: minimally invasive for the patient approach, use of low cost devices, easy data visualization-&-understanding for people with a medical background. We proposed a clinical procedure in which the female patients hold the hands behind and above the head, while an operator can digitize her breast with a 3D scanner. After a manual ROI definition through cropping, meshes are automatically processed. We presented two approaches to 3D breast data analysis and summarization: Thin-Plate Splines (TPS) [20] and Bag-of-Normals (BoN) [21]. A reference dataset has been used to compute PCA on a set of discriminative and different breasts, and the obtained first 2 principal components have been used to plot the breasts into a 2D space. We empirically proved that breasts subjected to a surgery change their representation in this space, and we also estimated a confidence degree. The outcomes of this research are very attractive for breast surgeons, since it can be useful for pre/post surgeon patients monitoring and in performance and quality assessment of surgeons. Moreover, thanks to the parametrization of 3D breast meshes, we were able to represent 3D data with just 2 parameters, gaining a compact descriptor easier to be transfered and browsed through hospital networks, if compared to a complex 3D mesh. As future works, we planned to augment the ROI extraction phase, which is a critical part of the proposed procedure and requires professionals with a proper know-how of 3D object editing.

Many real use cases have been shown in this dissertation. In all of them, we described analysis and summarization methods for several kind of media, properly videos, images and 3D data, to be used in high bandwidth connected environments. Definitively, the major project reported in this dissertation is the framework The Social Picture (TSP), in which we are able to process collections of thousands of images. Through this procedure we analyzed and summarized images collections. We use all the images within a collection to train a model of attention and learn the novel kind of saliency that we named Social Saliency. TSP employs a crowdsourcing paradigm that finds value on each image uploaded in the framework or gathered by the social media. Every image matters, even a funny selfie done during a vacation. We shown how cues about what is perceived as *social salient* may rise from a collection of images. And it is not all: we described how media collections can be employed for 3D reconstruction from images, scene and context tracking from videos, parametrization from 3D medical data and preservation and restoration from 3D Cultural Heritage data. In this way, *the tourists become unintended keepers of a virtual cultural heritage, that will never be lost, making meaningful even our most meaningless selfies* [238].

Appendix A

ARCA: Automatic Recognition of Color for Archaeology

A.1 Introduction

At the beginning of the 20th century, Albert H. Munsell [239] established a system for specifying colors more precisely and showing the relationships among them. The Munsell color order system is based on the color-perception attributes of hue, value and chroma. Munsell defined numerical scales with visually uniform steps for each of these attributes. Hue is that attribute of a color by which we distinguish blue from red, yellow from green, and so on. Hues are naturally ordered in this scale: red (R), yellow-red (YR), yellow (Y), green-yellow (GY), green (G), blue-green (BG), blue (B), purple-blue (PB), purple and red-purple (RP). Black, white and the grays between them are called "neutral colors" (N). Value indicates the lightness of a color in a scale of value ranges from 0 (pure black) to 10 (pure white). Chroma is the degree of departure of a color from the neutral color of the same value. The scale starts from 0, for neutral colors, but there is no arbitrary end to the scale, as new pigments gradually become available. However, limits for representable chroma values have been defined by the so called MacAdam limits [240]. Specifying color by the Munsell system is a practice limited to opaque objects, such as soils or painted surfaces. This practice provides a simple visual method as an alternative to the more complex and precise method based on the CIE system and on spectrophotometry. For this reason, the Munsell system is adopted in contexts in which the recording or identification of colors of specimens (i.e., flowers, minerals, soils) is required [241]. The Munsell charts are appropriate for almost all jobs requiring color specification by visual means, as stated by specific neurobiological researches that demonstrated how that system has successfully standardized color in order to match the reflectance spectra of Munsell's color chips with the sensitivity of the cells in the lateral geniculate nucleus (LGN cells), responsible for color specification [242]. In archaeology Munsell charts are widely used as the standard for color specification of organic materials, colored glass, soil profiles, rock materials, textiles, metals, colored glasses, paintings and principally pottery. Archaeologists are used to employing Munsell Soil Charts directly on a cultural heritage or excavation site to identify the colors of the soils and of the artifacts retrieved. Indeed, it is very useful in the examination, classification and genesis analysis of soils [243, 244, 245]. For which regards the interpretation of pottery, the precise color specification of parts like treated surfaces, clay body, core, and outer layers like painting and slip, it is fundamental for defining its stylistic and technical features. Color specification might be exploited to bind the artifacts to a specific culture, society or civilization or even to a certain period of time [246].

As previously mentioned, the standard practice of Munsell estimation exploiting the Soil Charts is by visual means. The two adjacent constant-hue charts or chips between which the hue of the specimen lies have to be chosen. Then, by moving the masks from chip to chip to find the most similar one to the specimen, one can estimate its value, chroma and hue [241]. As can be seen, this procedure is error prone, time consuming and very subjective. The process described above should be repeated more than once and possibly also by other users, in order to obtain a more accurate estimation, since colors might not be perceived uniformly by different people [247]. Hence, an objective and automatic Munsell estimation method would be a valuable improvement to the field of archaeology.

Digital cameras have been used before to acquire pictures of soil specimens in a laboratory with controlled lighting conditions. Then, the Munsell notation has been exploited to estimate the mineral and organic composition of the specimens [248, 249, 250, 251, 252, 253]. However, all of these works still require a strictly controlled environment for the digital acquisition of the images. Suggested controls for a perfect estimation are related to artificial and natural lightning conditions, specimen and camera positions, angle of view, setting of the working plane and background with proper opaque and black materials to avoid light reflection [241].

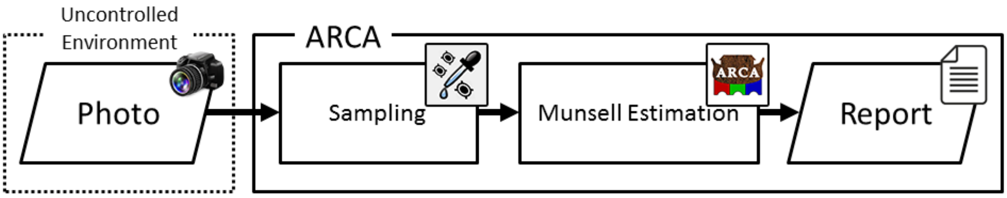


Figure A.1: Pipeline of the proposed ARCA application.

Prepare a perfectly controlled environment is difficult, time-consuming and potentially expensive. With the spread of smartphones with ever more sensors onboard, particularly high resolution cameras, new methods exploiting the Munsell system have been developed. In [254] a mobile phone application for Munsell estimation under strictly controlled illumination conditions is presented. In [255] a similar setting is discussed, but focused on a Complementary Metal Oxide Semiconductor (CMOS) sensor assembled on a smartphone. Also in these cases, a controlled environment is required. In the past we have performed several experiments about this topic in a such environment [256, 211, 250, 252, 253]. However, to the best of our knowledge, a method using an uncontrolled setting for image acquisition is still missing. In this Appendix, we describe ARCA: Automatic Recognition of Color for Archaeology, a desktop application for Munsell estimation. ARCA is the core of the pipeline of our method, consisting in the image acquisition of specimens, manual sampling of the image in a user-friendly way for archaeologists, Munsell estimation of the sampled points and creation of a sampling report (Fig. A.1). We focused on the need of archaeologists to have a practical and tested application that might help them in the color specification task during an excavation. Through the pipeline of ARCA, archaeologists do not need expensive tools (i.e., spectophotometer, Munsell Soil Charts, color checker) or a laboratory with a controlled environment for the acquisition in order to perform color estimation. They just need to take a picture of the specimen, and moreover, no strict constraints need to be applied in advance. Then, from the ARCA application, they are able to select multiple samples at once and the system will estimate the Munsell notation for them in an objective and deterministic way.

A dataset of 108 images, named for this reason ARCA108, consisting of a total of 22, 848 samples, have been gathered, in order to evaluate, in an uncontrolled environment, what the best configuration in which the image acquisition should be done. This dataset represents a new valuable asset for color specification research purposes. The Munsell system is usually exploited to establish and evaluate the color and gloss tolerance of specimens [257, 258]. We compared all the samples with Munsell reference values exploiting the CIEDE2000 (ΔE_{00}) color difference definition [257]. Several accuracy problem have been reported for the color specification task [259], so to be comparable with other Munsell estimation methods we will consider mean values and standard deviations from the evaluation phase.

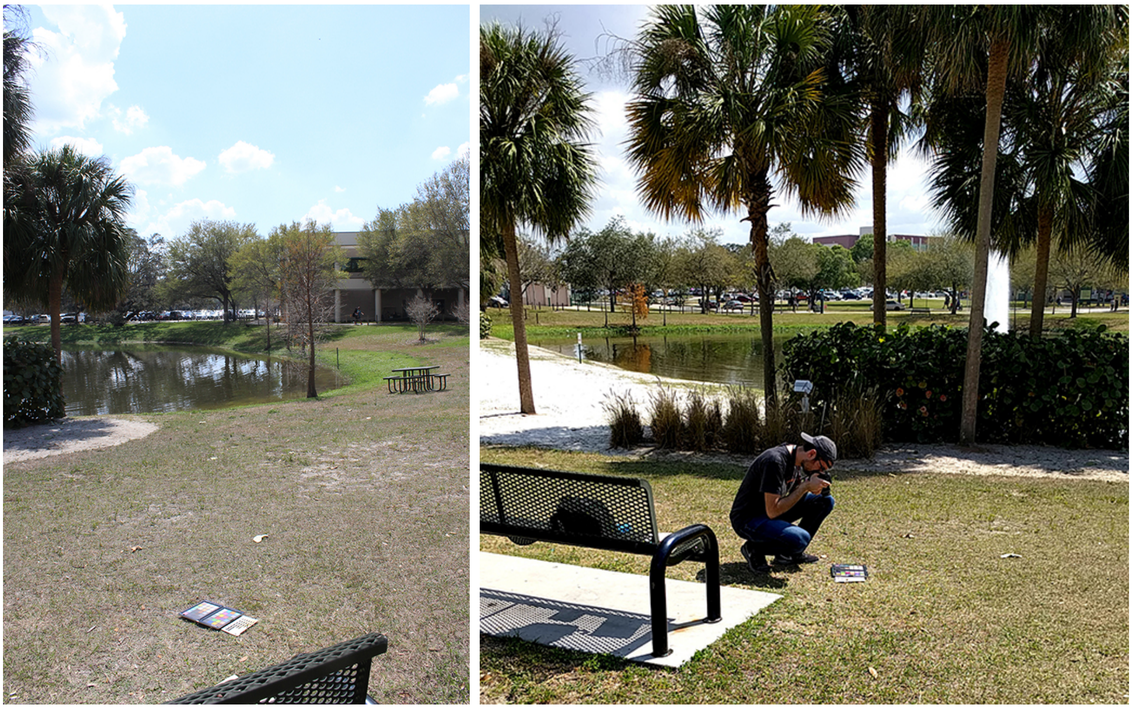
The contents of this Appendix are based on our published paper [260]. The structure of the Appendix is the following: in Section A.2 the acquisition phase, validation phase and ARCA desktop application will be described. The experimental results are given in Section A.3. Discussion and conclusions end the Chapter in Section A.4.

A.2 Material and methods

Two main phases can be distinguished in our experiments: acquisition and validation. In the former, we wanted to simulate the most common situation of Munsell field-estimation, while in the latter the main aim was to validate the proposed system, in order to prove its reliability in the Munsell estimation process. In the following subsections, the acquisition phase, ARCA desktop application and validation phase are detailed. The order in which they are presented is coherent with the proposed pipeline: acquire, sample and estimate.

A.2.1 Acquisition phase

No strict constraints have been added in the acquisition phase, in order to allow an easy replicability of the process shown in this work. Two kinds of devices have been employed in our experiments: a professional reflex and a common smartphone. The reflex model was a Canon EOS 1200D (mounting an EFS 18-55mm zoom lens model) with a resolution of 18 megapixels, while the smartphone model was a Nexus 5X with a main camera resolution of 12.2 megapixels. The subjects of the taken pictures were the following Munsell Soil Color Charts (Year 2000 Revised Washable



(a) (b)

Figure A.2: (a) The day in which acquisition phase has been performed was a sunny day with minor cloud cover. (b) Photos have been taken with an unguided approach.

Edition): *GLEY1*, *GLEY2*, *10R*, *2.5YR*, *5YR*, *7.5YR*, *10YR*, *2.5Y*, *5Y*. A Gretag-Macbeth color checker has been also employed, in order to evaluate the gains of have reference colors during photos acquisition.

Our acquisition was set in Tampa, Florida (US), in GPS coords 283’47.9”N 824’40.9”W, on March 8, that was an almost sunny day, with some cloud cover (Fig. A.2(a)). It was performed from 10:30 am to 12:30 pm and with an unguided approach (Fig. A.2(b)), so without any fixed positions or angles of view for the camera or subjects. We acquired the 9 charts of the Munsell Soil Color Charts, with the following possible settings:

- 2 kinds of devices: professional DSLR (Digital Single Reflex Camera) and common smartphone;

- 3 automatic white balancing algorithms (executed by the devices in the image capture phase): automatic, sunny (corresponding to standard illuminant

D65: $\sim 6,500K^{\circ}$) and cloudy (corresponding to standard illuminant D75: $\sim 7,500K^{\circ}$);

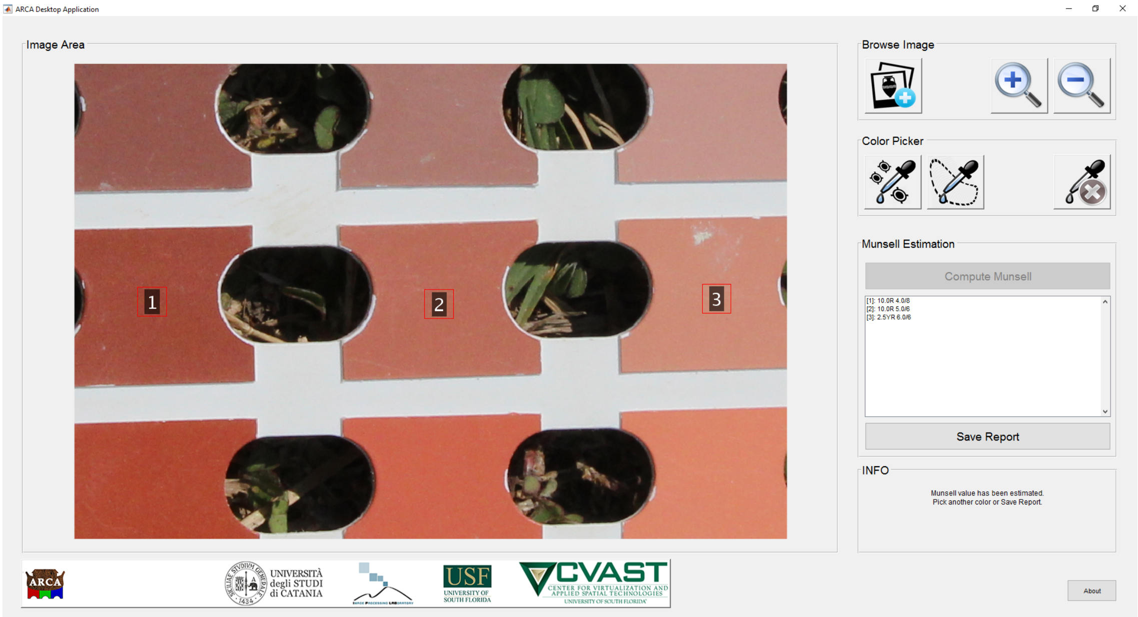
- 1 fluorescence presetting: direct sunlight;
- 1 ISO setting: 400 ISO;
- 1 focus setting: autofocus;
- 2 kind of subject: the chart itself and the chart with a Gretag-Macbeth color checker nearby.

In this way, we obtained a total of 12 configurations for each Munsell chart, gaining a total of 108 images. The resolution of the images is 5184×3456 pixels and 3840×2160 pixels for pictures taken by a DSLR camera and smartphone, respectively. All the images were saved in the standard JPG format, with a lossless setting for the quality (the highest possible).

The gathered dataset has been publicly released with the name ARCA108 and it is freely available at [261].

A.2.2 ARCA Desktop Application

The current version of the ARCA desktop application has been developed in Matlab Graphical User Interface Design Environment (GUIDE). From the GUI the user is able to perform several actions: open an image, zoom in/out to focus on a detail of an image, sample the image (through pick-point or draw-a-region-by-freehand), remove the current sample, estimate the Munsell notation of the sample and save the report. The ARCA application GUI is shown in Fig. A.3. Multiple samples can be selected through the pick-point tool; their Munsell notation estimation will be done at once when launched. After every estimation, indexed markers are added on the image, so the user can track all the samples with their own Munsell estimation. Samples are also highlighted with a red border (this is particularly useful when draw-a-region tool is used). Munsell conversion and ΔE_{00} computations for a validation phase are performed exploiting the publicly available Matlab toolbox by P. Centore [262, 263], that has been proved to be comparable with other not open-source conversion methods [264, 265, 266, 267]. Finally, when the report of the estimation is going


ARCA Desktop Application
Image Area
1
2
3
Browse Image
Color Picker
Munsell Estimation
Compute Munsell
Save Report
INFO
About
ARCA
UNIVERSITÀ degli STUDI di CATANIA
USF UNIVERSITY OF SOUTH FLORIDA
CVAST CENTER FOR VIRTUALIZATION AND APPLIED SPATIAL TECHNOLOGIES UNIVERSITY OF SOUTH FLORIDA

```
[1]: 10.0R 4.0/8
[2]: 10.0R 5.0/6
[3]: 2.5YR 6.0/6
```

Munsell value has been estimated.
Pick another color or Save Report.

Figure A.3: Screenshot of the ARCA desktop application GUI. Three samples have been taken on the current image; marks are visible on the image so the user can visually track the estimated Munsell values. Now the user can keep sampling the image (adding new Munsell estimations) or save a report of the current estimation.

to be created, the user must provide a name for the report and a directory will be created with that name. The report is made up of three elements: the starting image with the indexed markers on it, a Matlab file and a textual report containing the list of Munsell estimations.

A.2.3 Validation Phase

We evaluated the system comparing the expected Munsell value of each chip in the Munsell charts with its observed one. We performed the sampling from the charts importing the images in our ARCA desktop application and manually picking points that were visually near to the centroid of each chip. We considered a patch of 49 × 49 pixels around the picked centroid, for a total of 2, 401 pixels per chip. As done in [254], the Munsell charts labeled as *GLEY1* and *GLEY2* have not been evaluated, since they contain neutral colors very similar to one another's and with very low chroma values. We sampled 238 chips (for each one of the 12 configurations), and for each sampled chip we computed mean, median and mode of the extracted patch.

So, by also taking into account the RGB value in the centroid, we obtained 4 RGB values for each sampled chip. Using the Munsell toolbox by P. Centore [263] the sampled RGB values have been converted to the Munsell color space. We have also considered a discretized version of the converted RGB values, computed by rounding the converted values to the closest Munsell reference values in the Munsell charts. In this way, we obtained a total of $22,848$ Munsell observed values to be compared with the 238 expected ones.

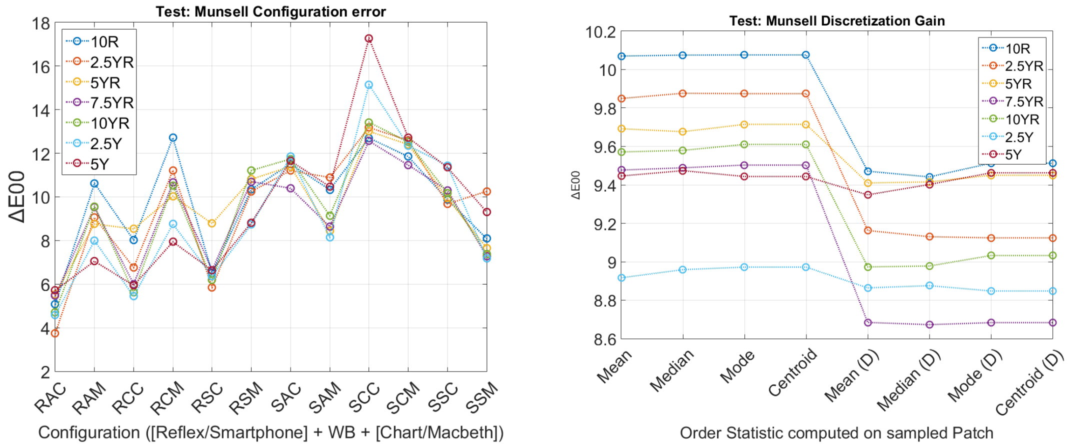
A.3 Results

In the experimental setting, 12 possible configurations were defined (Section A.2.1). We repeat that a "configuration" is one of the possible combination of the following settings: Device:[Reflex/Smartphone] + WhiteBalancing:[Auto/Sunny/Cloudy] + Subject:[Solo_Chart/With_Macbeth]. Moreover, for each sampled chip, 4 order statistics were investigated: mean, median, mode and centroid value have been exploited in the Munsell computation (Section A.2.3). Since Munsell references are a discrete set of values, it is also possible to apply a discretization to the continuous Munsell values obtained after the conversion, so the order statistics to be taken into account become 8. Hence, several questions can be raised, and will be answered in the following subsections:

1. What is the best configuration, among the 12 defined?
2. What is the best order statistic, among the 8 defined?
3. How much is worthwhile the application of the discretization?
4. Is the error in the Munsell notation estimation *acceptable*?

A.3.1 Best configuration

For each one of the 12 possible configurations, 7 Munsell charts were acquired. The average value of the ΔE_{00} between the Munsell reference chips and the 8 order statistics from every chip in the acquired charts has been computed. Results are shown in Fig. A.4(a). From this chart it is possible to assess that the best configuration is [Reflex, Auto White Balancing, Solo Chart]. Instead, among the configurations



(a) (b)

Figure A.4: Validation plots. (a) Investigation of the best configuration among the 12 tested. (b) Investigation of the best order statistic to be used on the patch during the Munsell estimation. Note how the discretization decreases the ΔE_{00} in almost the totality of the cases, as expected.

that exploit the smartphone as device, the best configuration is [Smartphone, Sunny White Balancing, With Macbeth]. It is interesting, and almost surprising, to notice how the use of a color checker together with a reflex professional camera increases the ΔE_{00} distance, while together with a general purpose smartphone it has a positive influence decreasing the distance. Hence, in our best configuration none expensive color checker is needed.

A.3.2 Best order statistic

The average value of the ΔE_{00} between the Munsell reference chips and the 8 order statistics from every chip in the whole dataset has been computed. Results are shown in Fig. A.4(b). The values of the order statistics, respectively with and without quantization, is almost similar, besides for 2.5Y and 5Y Munsell charts where it is almost the same in both the cases. The mean slightly outperforms the other order statistics. Additional evidence coming from this chart is that quantization decrease the ΔE_{00} distance in almost the totality of the cases. This state directly brings to the successive question.

A.3.3 Discretization

Munsell Soil Charts contain a discrete set of reference values, but conversion from RGB to Munsell System generates values in a continuous system. Archaeologists are used to employ only the discrete values, not the continuous ones, so a discretization is needed. Since we consider all the values near to a reference one as the same, the discretization to the closest Munsell reference value will matter in the ΔE_{00} computation. We counted how many times discretization actually decreased or increased the initial ΔE_{00}. In the 59.42% of the cases a positive gain has been obtained. Moreover, the negative gain is usually obtained with low chroma values, which are the most ambiguous to be classified. From the result shown in Fig. A.4(b) and this other cue it is possible to assess that it is worthwhile to apply discretization, as expected.

A.3.4 Color Tolerance

The issue related to the amount of acceptable error on a Munsell estimation is known as color tolerance. The tolerance ranges change with respect to the application for which the estimation is made for. The standard definition is that same colors should have $\Delta E_{00} = 1$ [257]. In the industrial field two colors can be considered the same (imperceptible differences) only if the ΔE_{00} is lesser than 2. However, this strong criteria is usually relaxed introducing "tolerable" ranges: until 3-4 CIELAB units can be considered the same colors, until 5-6 CIELAB units the colors are hardly distinguishable, higher than 6 CIELAB units classification performance starts to decrease [254]. Moreover, the colors printed in the Munsell reference Soil Charts are usually affected by an intrinsic error from ~ 1 to ~ 4 CIELAB units, where higher error is found in elder Charts [259]. Related works employing smartphones during acquisition phase in a controlled environment have reported an error in the estimation of 3.75 ± 1.8 CIELAB units [254, 255]. In Table A.1 the mean and standard deviation values of ΔE_{00} computed during the validation phase have been reported. As previewed by Fig. A.4(a), the best configuration is [Reflex, Auto White Balancing, Solo Chart], which has 4.95 ± 2.89 CIELAB units of error. Performances drastically drop with other configurations. The best configuration for smartphone, that is [Smartphone, Sunny White Balancing, With Macbeth], has 8.20 ± 2.71 CIELAB

Table A.1: Mean and Standard Deviation of Munsell estimation for each one of the 12 defined configurations.

	Configurations ([Reflex/Smartphone] + WB + [Chart/Macbeth])											
	RAC	RAM	RCC	RCM	RSC	RSM	SAC	SAM	SCC	SCM	SSC	SSM
Mean	***4.950***	8.992	6.626	10.338	6.691	10.169	11.381	9.436	13.786	12.264	10.352	8.198
St. D.	2.887	2.667	2.807	3.191	***2.562***	2.914	2.888	3.266	3.220	2.756	2.998	2.719

units of error. To summarize, taking into account all the previous considerations about intrinsic error of Munsell Charts and the unconstrained experimental setting, the error obtained with our best configuration (employing the reflex) seems reasonable.

A.4 Conclusion

In this Appendix, ARCA: Automatic Recognition of Color for Archaeology, a desktop application for Munsell estimation, has been presented. We focused on the need of archaeologists to have a practical and tested application that might help them in the color specification task during an excavation. The following pipeline for Munsell notation estimation, aimed at archaeologists, has been proposed: image acquisition of specimens, manual sampling of the image in the ARCA desktop application, automatic Munsell estimation of the sampled points and creation of a sampling report. Differently from our previous works [250, 252, 253], we performed the whole experiments in an uncontrolled environment. A dataset of $22,848$ samples has been gathered under the uncontrolled environment assumption and evaluated with respect to the Munsell reference Soil Charts. This dataset has been named ARCA108 and it represents a valuable asset for color specification research purposes. We defined 8 possible order statics for characterize the samples and 12 possible configurations during the acquisition phase. Experimental results shown that the defined order statistics reach very similar results, and that discretization of the converted Munsell notation decreases the error of ~ 1 CIELAB unit. The best configuration among the tested ones is [Reflex, Auto White Balancing, Solo Chart], with 4.95 ± 2.89 CIELAB units of error. Compared to other related works, taking into account intrinsic error of Munsell reference Soil Charts and the uncontrolled experimental setting, this result is encouraging and reasonable. We proved that ARCA can represent for archaeologists a valid tool for color specification. ARCA allows archaeologists to select multiple samples and estimate the corresponding Munsell notation at once, in a fast, objective and deterministic way, avoiding the error-prone and time-consuming procedure of Munsell Estimation by visual means and without any expensive tool like spectophotometer, Munsell Soil Charts or Gretag-Macbeth color checker. For future works, we are planning to improve the ARCA application (i.e., image processing algorithms for noise reduction, deployment of a mobile version), to expand the validation phase acquiring other Munsell Soil Charts from Tropical Soils edition and, most of all, to conduct a color specification test-case on archaeological soils and pottery.

Acknowledgments

The method described in this work is part of the patent USF Ref. No.: 17A055 - Automatic digital method for classification of colors in Munsell Color System.

Appendix B

3D Web Viewers Benchmark

B.1 Introduction

Nowadays there is a growing interest in the digitization of cultural heritage site and artifacts. A digital file can be easily duplicated and sent all over the world without the harm of damage the original file and loose potentially critical details. The spread of media contents results useful both for the general audience and the specialized scholars. Indeed, cultural heritage experts might study a digital version of an artifact inaccessible due to political, historical or geographical reasons. The artifact itself might even been destroyed or modified due to weather effects or reckless looters. In this scenario web viewers represent a smooth and easy way to explore the digital content of a collection directly from the web browser. Currently, a plethora of available web viewers exist. The main aim of this report is to assess if someone of the analyzed web viewers fits the needing of scholars specialized in Cultural Heritage digitization. In particular, this benchmark has been drafted in Spring 2017 within the Center for Virtualization and Applied Spatial Technologies (CVAST) in University of South Florida (USF) [214]. Activities of CVAST include digitization of ancient cultural heritage artifacts and sites, indeed it acquired an huge amount of 3D data. This collection should be made accessible to the public through a proper web viewer. The viewer should implement advanced exploration tools (e.g., measurement, marking, annotation, . . .), in order to augment the data exploration and make it useful for experts in the archaeological and historical fields.

Analyzed web viewers will be described in the remainder of the benchmarking report. Finally, in Table B.1 a comparison highlighting pros and cons of the reported frameworks is shown.

B.2 Benchmarking Report

In this benchmarking report the following web viewers have been analyzed: Sketchfab, Autodesk solutions, TMS, Unity, 3DHOP. Details are given in the proper subsections. Refer to the comparative Table B.1 at the end of the report for a comparison at a glance of all reported viewers.

B.2.1 Sketchfab

Sketchfab is a website hosting collections of 3D objects [197]. In its community users created collections for very different purposes: gaming, sports, technologies and also cultural heritage. It supports a wide range of viewable formats [268], which includes among the others the OBJ, PLY and STL formats. Uploaded models can be set to be "unlisted" (Private), hence URL must be known in order to view the content. However, anyone with the URL can view and embed the private model, so a password may be set to restrict the access only to authorized users. Download of the meshes can be disabled. No measurement or annotation tools are available in the viewer, but annotations can be added statically on the model by the owner. Models can be shared via URL or html embedding code. Users cannot customize the viewer, but frequent updates are released through the time.

B.2.2 Autodesk Solutions

Autodesk provides a set of different solutions to handle and manage a digital collection. For educational organizations, account registration and some software are freely available using an official e-mail address during registration. It is possible to distinguish between softwares, viewers and galleries. Among the many Autodesk softwares, only *ReMake* has been included in this report. Among the viewers, *A360 Viewer*, *Project Play* and *ReCap* have been reported. Finally, galleries represent a way to publish the collections on the web, so they have been reported at the end of this section.

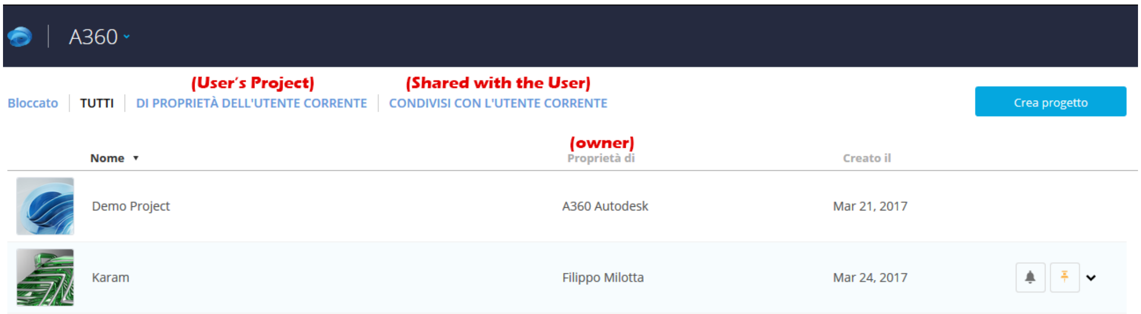
Autodesk A360

A360 is a solution of Autodesk thought to provide storage space for individual users and teams. Different features can be distinguished within A360.

BIM Team: this feature provides a project-based collaboration file-system hosted in cloud, for design teams in the Building Interactive Model (BIM) field [269]. This is clearly out of the scope of Cultural Heritage digitization, but the framework itself has some potentials and it is worthy to be reported in this benchmark. As can be seen on the official website [269], its annual cost is variable and depends on the number of users in the team: $1,200$\$ per 10 users (small teams), $1,800$\$ per 25 users (large teams) and $4,800$\$ per 100 users (organizations). The number of users really matters, since the space available in the cloud for data storage is strictly related to the number of users: 500GB per user. So, basically, it is possible to compute the annual pricing with respect to the storage needed: $1,200$\$ for 5TB ($5,000$GB), $1,800$\$ for 12.5TB ($12,500$GB) and $4,800$\$ for 50TB ($50,000$GB). In addition, each user has 25GB of personal storage space on *A360 Drive* (freely given with educational account, where free-registration applies).

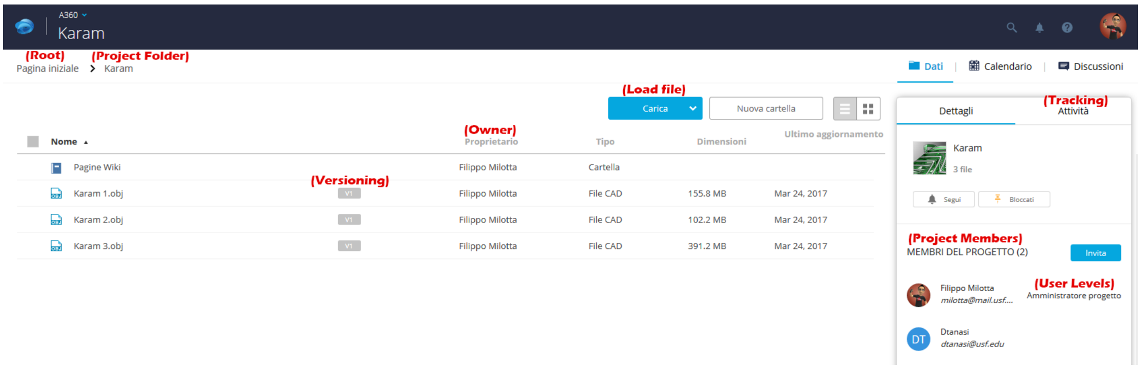
There is no limitation on the number of "Projects" that can be created in *A360 BIM Team*. A project is nothing more than a folder in which meshes can be stored. Privacy and administration settings can be individually set for each project. As in any file-system, files can be copied, moved, deleted and, as an additional feature, substituted with a more recent version (Figs. B.1 and B.2). Hence, versioning is supported, this means that in any time it is possible to rollback (or simply take a look) to a previous version of the same file. File can be shared via URL, e-mail or html embedding. The sharing feature allows to enable or disable the download of the file and it is possible to set a password in order to obtain access to the data. The meshes can be browsed through the *Viewer* feature (detailed below).

Viewer: it implements the classical feature of 3D object visualization (i.e., orbit, pan, zoom, ...- Fig. B.3). There is no tool for measurements. The list of compatible file formats is reported in the website [270]. In particular, OBJ is supported, while PLY not. A critical issue is that textures seem to be not supported at all. We opened a thread on the official Autodesk forum [271], but no answers have been



Nome	Proprietà di	Creato il
Demo Project	A360 Autodesk	Mar 21, 2017
Karam	Filippo Milotta	Mar 24, 2017

Figure B.1: A360 BIM Team: Projects view.



Nome		Proprietario	Tipo	Dimensioni	Ultimo aggiornamento
Pagine Wiki		Filippo Milotta	Cartella		
Karam 1.obj	V1	Filippo Milotta	File CAD	155.8 MB	Mar 24, 2017
Karam 2.obj	V1	Filippo Milotta	File CAD	102.2 MB	Mar 24, 2017
Karam 3.obj	V1	Filippo Milotta	File CAD	391.2 MB	Mar 24, 2017

MEMBRI DEL PROGETTO (2)

Filippo Milotta
milotta@mail.usf....
Amministratore progetto

Dtanasi
dtanasi@usf.edu

Figure B.2: A360 BIM Team: files within a Project.

given yet. Our guess is that only materials are supported in this viewer, so textures cannot be added, currently. Annotations are supported, but with a strong limitations (Fig. B.4): they are stored in the *BIM Team* environment. This means that, once the mesh has been shared, than no annotations will be linked to the model and the users may only add locally-valid annotations (that will be lost when the viewer is closed, eventually).

Drive: is really similar to the file-system of *BIM Team*. The differences are that the storage space in *Drive* is personal and limited to 25GB. The rationale of having this feature is that users can locally work on a mesh, storing it in their personal storage space, and then uploading or commit their changes or new meshes in the project's folder in *BIM Team*. Indeed, files stored in *Drive* are visible and synchronized with other Autodesk desktop applications, like *ReMake*.

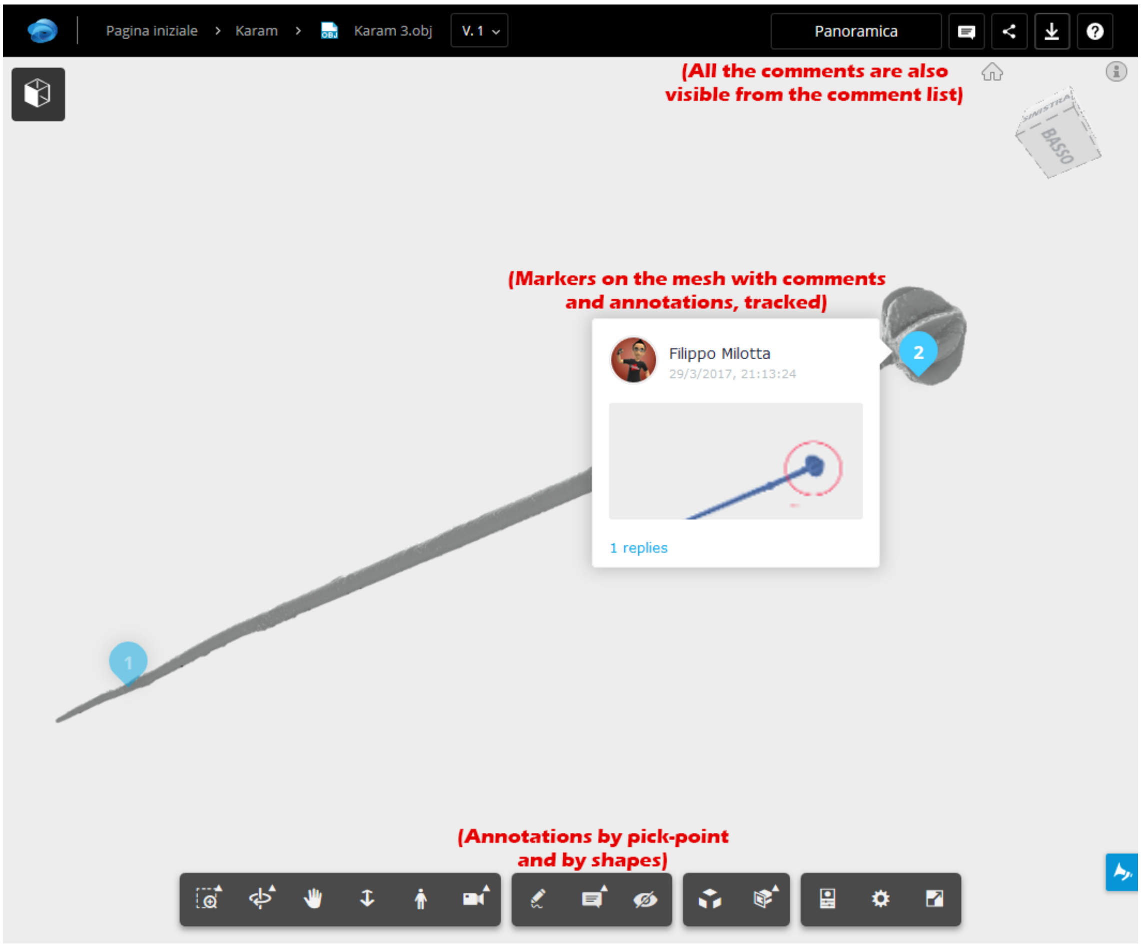

Pagina iniziale
Karam
Karam 3.obj
V. 1
Panoramica
(All the comments are also visible from the comment list)
BASSO
(Markers on the mesh with comments and annotations, tracked)
Filippo Milotta
29/3/2017, 21:13:24
1 replies
2
1
(Annotations by pick-point and by shapes)

Figure B.3: A360 Viewer: a 3D model with two annotations.

Design Graph: this feature is claimed to be a shape-based Machine Learning framework to recognize and understand parts, assemblies and entire designs of 3D models. Unfortunately, it is currently in a beta version, and does not work yet. This feature has been reported here just as a secondary remark, since in a near future it could become useful for tasks like meshes recognition (e.g., pots, coins, . . .).

ReMake - Project Play:

ReMake (former Memento) is a desktop application for 3D object editing [273]. Models imported in ReMake are converted in RCM (ReMake) format. Models can also be imported from the synchronized *A360 Drive* account of the user. They must

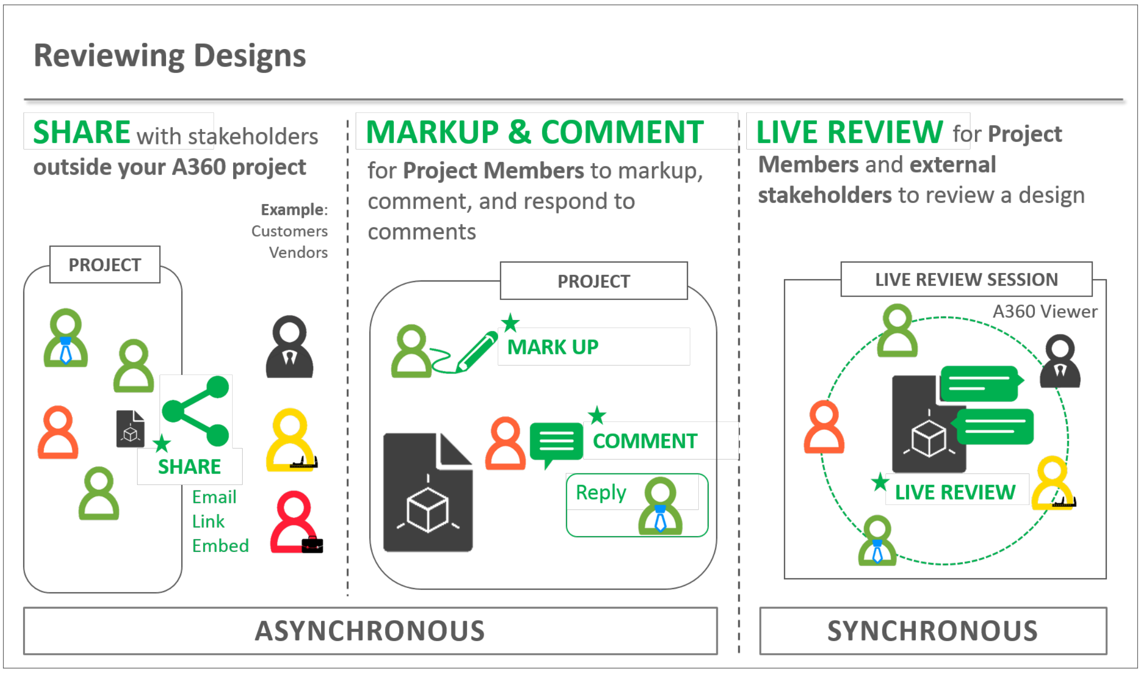

Reviewing Designs
SHARE with stakeholders outside your A360 project
Example: Customers Vendors
PROJECT
SHARE
Email
Link
Embed
MARKUP & COMMENT for Project Members to markup, comment, and respond to comments
PROJECT
MARK UP
COMMENT
Reply
LIVE REVIEW for Project Members and external stakeholders to review a design
LIVE REVIEW SESSION
A360 Viewer
LIVE REVIEW
ASYNCHRONOUS
SYNCHRONOUS

Figure B.4: A360 Viewer annotation system [272].

have color vertices (no textured colors) in order to obtain a RCM textured file. Since it is an object modeling and editing software, many tools for change the geometry of the meshes are available in the toolbars B.5. Once edited the object, it is possible to export it via screenshot, video or file (converting it into OBJ, FBX, STL, PLY, XYZ, and PTS formats). It is also possible to directly publish the model into the *Autodesk ReMake Gallery*.

This process of "publishing into gallery" passes through the *Autodesk Project Play* viewer. Indeed, models are normalized in order to be explored in this viewer. Three templates for the viewer are available: classic (just visualization options), advanced (with measurement tools, Fig. B.6) and comparative (two cameras, so the models can be viewed from two different points of view at the same time). The classical visualization modes are available: solid, textured, wireframe. None annotation tools is implemented.

The *Autodesk Project Play* viewer is currently in a development phase and under testing. We have applied as beta developer/tester and our application has been accepted; in this way we have obtained access to the tutorials and the supplementary

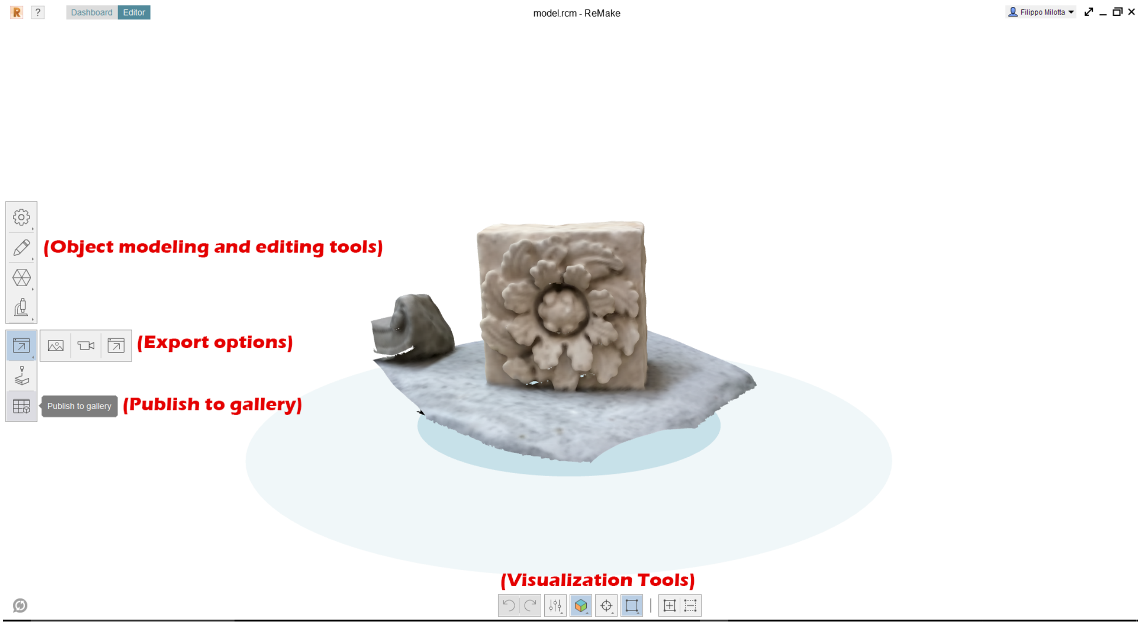


Figure B.5: ReMake GUI of the desktop application.

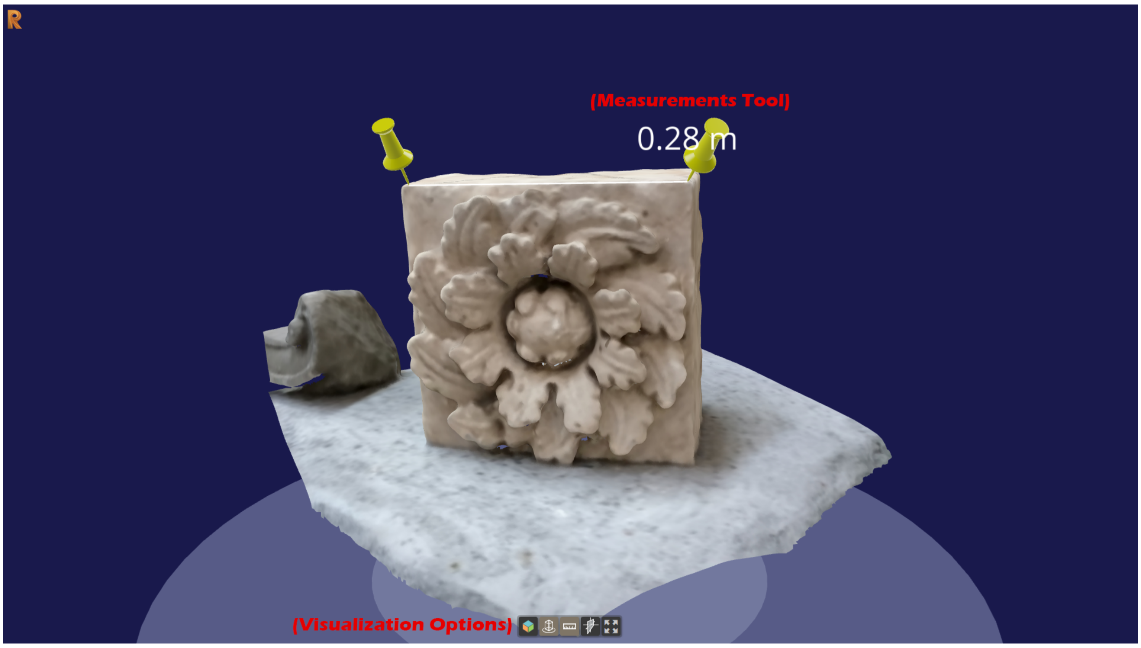


Figure B.6: Project Play viewer.

materials about the implementation of this viewer. It is meaningful, since *Autodesk Project Play* viewer is the starting point of the Smithsonian X 3D viewer [196]. The Smithsonian's viewer represent a sort of state-of-the-art for the web viewers adopted by the museums. Moreover, in an our recent publication on Journal of Electronic Imaging [162], where we presented our Unity-based viewer, one of the reviewer has stated the following during the reviewing process: "*See the work of Autodesk for the Smithsonian Museum: I am not saying you must have the same level of polishness (or the same time and effort investment), but all the features useful for a technical public are there (sections, measurement, lighting and shading control)*". So, the *Autodesk Project Play* has the potentials to become a good web viewer, but it needs an effort for the customization of the base viewer, since each component/tool/feature has to be programmed ad-hoc. Moreover, the system is based on a node-graph system development that requires a proper knowledge to be used.

ReCap:

ReCap is a software/viewer for SfM acquisition and laser scan [274]. The viewer has very basic functionalities like orbit and visualization modes (Fig. B.7). There is none annotation, editing or measurement tools. As in *ReMake*, models processed in *ReCap* can be retrieved from the personal *A360 Drive* storage space and can be published to the *Autodesk ReCap Gallery*.

Autodesk Galleries:

the Autodesk Galleries are the place where the collections (called projects) of the users are meant to be published. There, many different galleries exist (Fig. B.8), distinguished with relation to the interest field of the collections published within (Fig. B.8). The best fitting gallery for the digital archaeology is the *ReMake Gallery*. However, the management of the collection is the same in any gallery, so the choice about which one should be used is mainly dependent on the implied softwares. When a digital object is published to a gallery, a confirmation from the Autodesk administrator is need sometimes; it usually requires about 1 or 2 business days. Once a project has been published it is... published. This means that it is accessible to everyone. It is possible to set the privacy of an object in the collection and allow

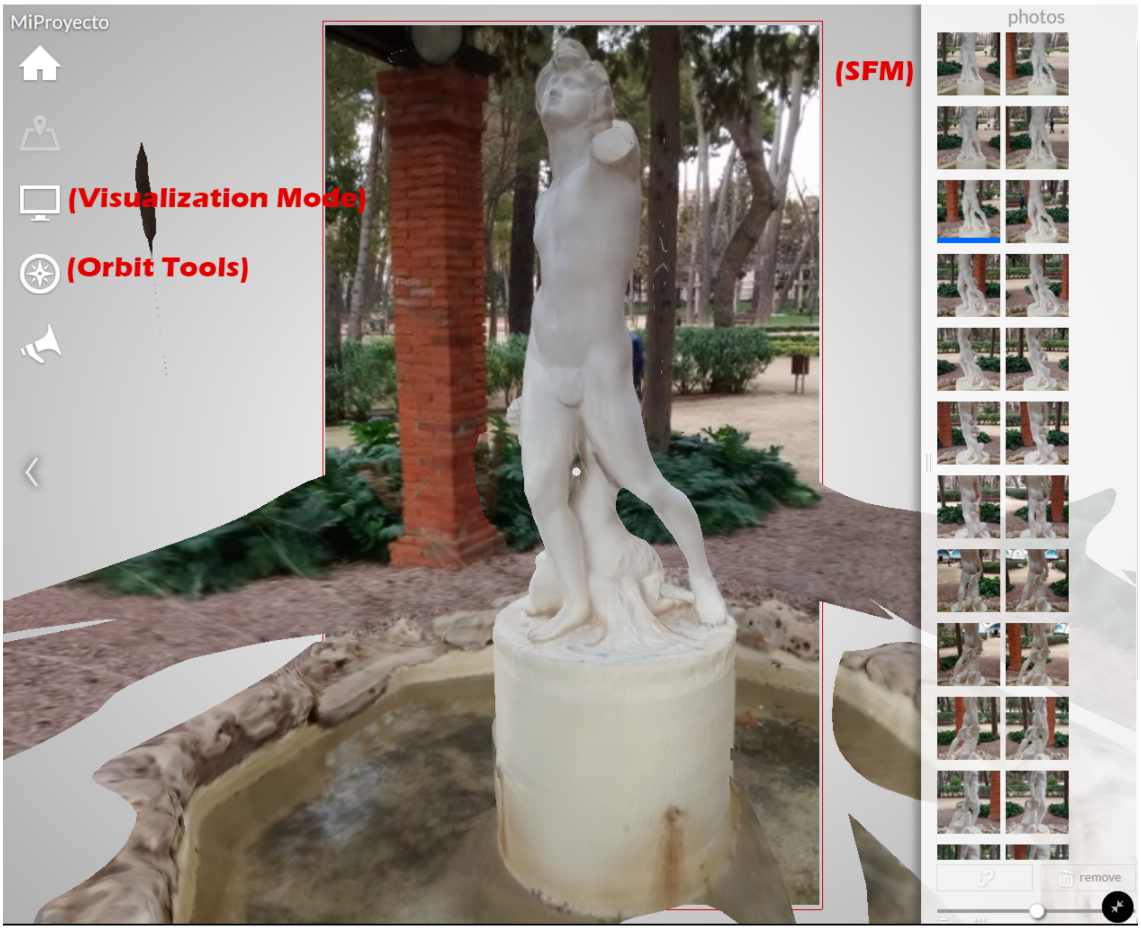


Figure B.7: ReCap viewer.

access to the object through a password. However, the password cannot be set by the user: it is automatic generated by the system and should be shared by the owner of the collection to any authorized user. Each object in the collection has a different password. It is possible to enable or disable the download of the objects in the collection. A tag system to label the data, simplifying the queries of the users, is provided.

Collections can contain different type of media files, like text, image, video or 3D object. As said before, three kind of viewers are available in the Galleries: *A360 Viewer*, *Project Play Viewer* and *ReCap Viewer*. If an object in the collection can be browsed with one of this viewer, then an icon will appear in the collection cover: a cube for *A360 Viewer*, a triangle for *Project Play* and a stylized *R* for *ReCap*

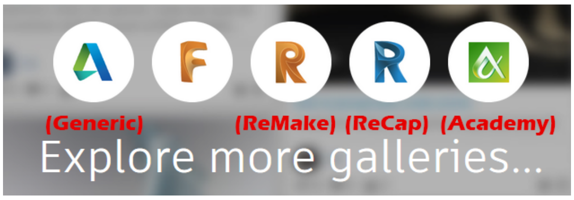


Figure B.8: List of the most common Autodesk Galleries.

Viewer (Figs. B.9 and B.10).

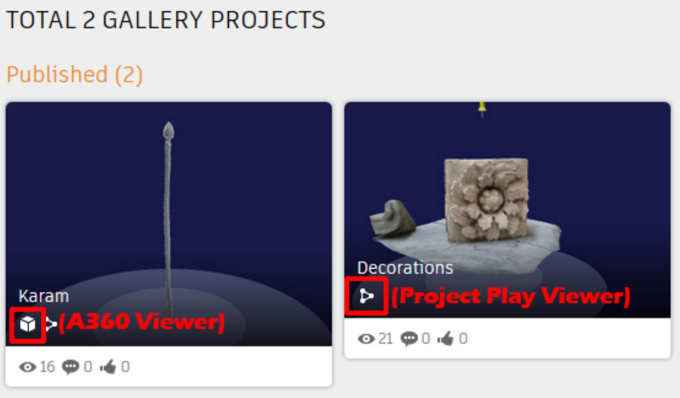


Figure B.9: Collections with objects that can be browsed with A360 and Project Player Viewers.

B.2.3 The Museum System (TMS)

The Museum System (TMS) is a digital collection management system currently adopted by several museums (e.g., the Ringling Museum) [275]. We have contacted the developers of TMS in order to know if their system is capable of handle 3D data, but they answered that it cannot.

B.2.4 Unity Engine

Unity is a framework mainly adopted to develop videogames [195]. We have participated in the development of a web viewer based on Unity [162, 166]. It is highly customizable, but lot of effort should be putted on it to obtain a meaningful result. However, Unity is not able to handle meshes with an ultra high resolutions and for

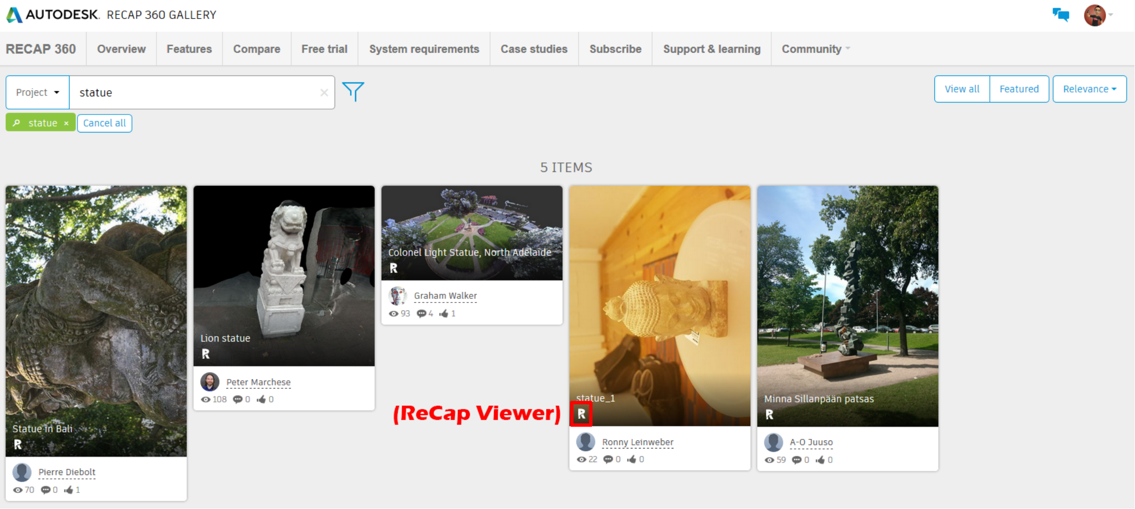


Figure B.10: Collections with objects that can be browsed with ReCap Viewer.

this reason tends to become unstable. When a mesh is loaded into Unity, it is automatically splitted into sub-meshes with at most $\sim 64,000$ vertices; this can results in a drop of the quality of the mesh.

B.2.5 3DHOP

As stated in the website, 3DHOP allows the creation of interactive visualization of 3D models directly inside a standard web page, just by adding some HTML and JavaScript components in the HTML code [276]. This means that, given the source code of the original 3DHOP (freely available in the website), it is possible to customize the code and potentially create new functionalities for the viewer. The system is based on the JavaScript language. It supports only PLY format (and Nexus format for multi-resolution meshes). It implements the classical feature of 3D object visualization (i.e., orbit, pan, zoom, . . .). There is a tool for measurements (but it does not taken into account the original metric unit, that must be manually redefined in the code). It supports textured models (only color vertex). It partially supports annotations by pick-point or regions, currently just statically (must be defined on the model and cannot be defined dinamically by the users).

B.3 Comparison Table

Table B.1: comparative table highlighting pros and cons of the reported frameworks. Each column is respectively filled with viewer names, place in which data are stored, supported formats, supported color capability, measurement tool implementation, annotation tool implementation, customizable option.

VIEWER	**STORAGE**	**FORMATs**	**COL.**	**MEAS.**	**ANNO.**	**CUST.**
Sketchfab	Sketchfab website	OBJ, STL, PLY, . . .	Yes	No	Static	No
A360	A360 BIM Team	OBJ, STL, PLY, . . .	No	No	Only on A360	No
Project Play	ReMake Gallery	RMC	Yes	Yes	Only if cust.	Yes
ReCap	ReCap Gallery	[SfM, 3D Scanner]	Yes	No	No	No
Unity	Locally	OBJ	Yes	Yes	Static	Yes
3DHOP	Locally	PLY	Yes	Yes	Static	Yes

Bibliography

[1] H. Chesbrough, W. Vanhaverbeke, and J. West. *Open innovation: Researching a new paradigm.* Oxford University Press on Demand, 2006.

[2] S. Sesia, M. Baker, and I. Toufik. *LTE-the UMTS long term evolution: from theory to practice.* John Wiley & Sons, 2011.

[3] S. Battiato, G. M. Farinella, F. L. M. Milotta, A. Ortis, L. Addesso, A. Casella, V. D'Amico, and G. Torrisi. "The Social Picture". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval.* ACM. 2016, pp. 397–400.

[4] C. Wu. "Towards linear-time incremental structure from motion". In: *3DTV-Conference, 2013 International Conference on.* IEEE. 2013, pp. 127–134.

[5] Website: Visual Structure from Motion (VSFM). *VSFM GUI application.* http://ccwu.me/vsfm/. Online; last visited on June 26, 2017.

[6] F. L. M. Milotta, M. Bellocchi, and S. Battiato. "The Social Picture: Advanced Image Analysis Applications". In: *STAG: Smart Tools and Applications in Graphics (2017).* 2017.

[7] A. Ortis, G. M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, and S. Battiato. "RECfusion: Automatic Video Curation Driven by Visual Content Popularity". In: *ACM Multimedia.* ACM MM 2015. 2015, pp. 1179–1182. DOI: 10.1145/2733373.2806311. URL: http://www.recfusionproject.altervista.org/.

[8] F. L. M. Milotta, S. Battiato, F. Stanco, V. D'Amico, G. Torrisi, and L. Addesso. "RECfusion: Automatic Scene Clustering and Tracking in Video from Multiple Sources". In: *EI – Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications.* 2016. URL: http://recfusionproject.altervista.org/clustertracking.htm.

[9] S. Battiato, G. M. Farinella, F. L. M. Milotta, A. Ortis, F. Stanco, V. D'Amico, L. Addesso, and G. Torrisi. "Organizing Videos Streams for Clustering and Estimation of Popular Scenes". In: *Proceedings of Image Analysis and Processing - ICIAP 2017 19th International Conference*. Springer. 2017, pp. 51–61. DOI: 10.1007/978-3-319-68560-1.

[10] L. Duan, F. Gao, J. Chen, J. Lin, and T. Huang. "Compact descriptors for mobile visual search and MPEG CDVS standardization". In: *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE. 2013, pp. 885–888.

[11] L. Duan, J. Lin, J. Chen, T. Huang, and W. Gao. "Compact descriptors for visual search". In: *IEEE MultiMedia* 21.3 (2014), pp. 30–40.

[12] C. Koch and S. Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry". In: *Matters of intelligence*. Springer, 1987, pp. 115–141.

[13] A. Borji and L. Itti. "State-of-the-art in visual attention modeling". In: *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2013), pp. 185–207.

[14] K. Duncan and S. Sarkar. "Saliency in images and video: a brief survey". In: *IET Computer Vision* 6.6 (2012), pp. 514–523.

[15] T. L. Evans and P. T. Daly. *Digital archaeology: bridging method and theory*. Psychology Press, 2006.

[16] T. Nakamura, S. Suzuki, H. Sadeghi, S. M. Fatemi Aghda, T. Matsushima, Y. Ito, S. K. Hosseini, A. J. Gandomi, and M. Maleki. "Source fault structure of the 2003 Bam earthquake, southeastern Iran, inferred from the aftershock distribution and its relation to the heavily damaged area: Existence of the Arg-e-Bam fault proposed". In: *Geophysical research letters* 32.9 (2005).

[17] D. M. Seekins. "State, society and natural disaster: cyclone Nargis in Myanmar (Burma)". In: *Asian Journal of Social Science* 37.5 (2009), pp. 717–737.

[18] J. Curtis. "Archaeology and cultural heritage in war zones". In: *History for the taking* (2011), pp. 55–76.

[19] W. Wahbeh, S. Nebiker, and G. Fangi. “Combining Public Domain and Professional Panoramic Imagery for the Accurate and Dense 3d Reconstruction of the Destroyed Bel Temple in Palmyra.” In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 3.5 (2016).

[20] G. Gallo, D. Allegra, Y. G. Atani, F. L. M. Milotta, F. Stanco, and G. Catanuto. “Breast Shape Parametrization Through Planar Projections”. In: *International Conference on Advanced Concepts for Inteligent Vision Systems*. Springer. 2016, pp. 135–146.

[21] D. Allegra, F. L. M. Milotta, D. Sinitò, F. Stanco, G. Gallo, T. Wafa, and G. Catanuto. “Description of Breast Morphology through Bag of Normals Representation”. In: *Proceedings of Image Analysis and Processing - ICIAP 2017 19th International Conference*. Springer. 2017, pp. 511–521. DOI: 10.1007/978-3-319-68548-9.

[22] F. L. Bookstein. “Principal warps: thin-plate splines and the decomposition of deformations”. In: *Transactions on Pattern Analysis and Machine Intelligence* 11.6 (1989), pp. 567–585.

[23] Y. Rui, T. S. Huang, and S. Chang. “Image retrieval: Current techniques, promising directions, and open issues”. In: *Journal of visual communication and image representation* 10.1 (1999), pp. 39–62.

[24] N. Marz and J. Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[25] J. M. Wolfe. “Guided search 2.0 a revised model of visual search”. In: *Psychonomic bulletin & review* 1.2 (1994), pp. 202–238.

[26] A. Mikulík, O. Chum, and J. Matas. “Image retrieval for online browsing in large image collections”. In: *International Conference on Similarity Search and Applications*. Springer. 2013, pp. 3–15.

[27] A. Mikulík, F. Radenović, O. Chum, and J. Matas. “Efficient image detail mining”. In: *Asian Conference on Computer Vision*. Springer. 2014, pp. 118–132.

[28] Website: heatmap.js. *Dynamic Heatmaps for the Web*. http://www.patrick-wied.at/static/heatmapjs/. Online; last visited on July 21, 2017.

[29] Website: Google Maps APIs. *Heatmap Layer*. `https://developers.google.com/maps/documentation/javascript/heatmaplayer`. Online; last visited on July 21, 2017.

[30] Website: Matlab Documentation. *Colormap*. `https://www.mathworks.com/help/matlab/ref/colormap.html`. Online; last visited on July 26, 2017.

[31] J. Huang, S. R. Kumar, M. Mitra, W. Zhu, and R. Zabih. "Image indexing using color correlograms". In: *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE. 1997, pp. 762–768.

[32] N. Sebe and M. S. Lew. "Robust color indexing". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM. 1999, pp. 239–242.

[33] C. Schmid and R. Mohr. "Local grayvalue invariants for image retrieval". In: *IEEE transactions on pattern analysis and machine intelligence* 19.5 (1997), pp. 530–535.

[34] X. S. Zhou, Y. Rui, and T. S. Huang. "Water-Filling: a novel way for image structural feature extraction". In: *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*. Vol. 2. IEEE. 1999, pp. 570–574.

[35] Y. Liu, D. Zhang, G. Lu, and W. Ma. "A survey of content-based image retrieval with high-level semantics". In: *Pattern recognition* 40.1 (2007), pp. 262–282.

[36] M. Rehman, M. Iqbal, M. Sharif, and M. Raza. "Content based image retrieval: survey". In: *World Applied Sciences Journal* 19.3 (2012), pp. 404–412.

[37] C. Wang, L. Zhang, and H.-J. Zhang. "Learning to reduce the semantic gap in web image retrieval and annotation". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, pp. 355–362.

[38] Y. Chen, J. Z. Wang, and R. Krovetz. "CLUE: cluster-based retrieval of images by unsupervised learning". In: *IEEE transactions on Image Processing* 14.8 (2005), pp. 1187–1201.

[39] D. Doermann. “The indexing and retrieval of document images: A survey”. In: *Computer Vision and Image Understanding* 70.3 (1998), pp. 287–298.

[40] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*. Vol. 463. ACM press New York, 1999.

[41] M. Carrillo and A. López-López. “Concept Based Representations as Complement of Bag of Words in Information Retrieval.” In: *AIAI*. Springer. 2010, pp. 154–161.

[42] J. Mukherjee, J. Mukhopadhyay, and P. Mitra. “A survey on image retrieval performance of different bag of visual words indexing techniques”. In: *Students' Technology Symposium (TechSym), 2014 IEEE*. IEEE. 2014, pp. 99–104.

[43] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. “Visual categorization with bags of keypoints”. In: *Workshop on statistical learning in computer vision, ECCV*. Vol. 1. 1-22. Prague. 2004, pp. 1–2.

[44] C. Grana, D. Borghesani, M. Manfredi, and R. Cucchiara. “A fast approach for integrating ORB descriptors in the bag of words model”. In: *Proc. SPIE*. Vol. 8667. 2013, pp. 866709–866709.

[45] R. Ji, L. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao. “Location discriminative vocabulary coding for mobile landmark search”. In: *International Journal of Computer Vision* 96.3 (2012), pp. 290–314.

[46] Y. Yang and S. Newsam. “Bag-of-visual-words and spatial extensions for land-use classification”. In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM. 2010, pp. 270–279.

[47] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. “CHoG: Compressed histogram of gradients a low bit-rate feature descriptor”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 2504–2511.

[48] B. Girod, V. Chandrasekhar, D. M. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham. “Mobile visual search”. In: *IEEE signal processing magazine* 28.4 (2011), pp. 61–76.

[49] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod. "Tree histogram coding for mobile image matching". In: *Data Compression Conference, 2009. DCC'09.* IEEE. 2009, pp. 143–152.

[50] R. Ji, L. Duan, J. Chen, H. Yao, Y. Rui, S. Chang, and W. Gao. "Towards low bit rate mobile visual search with multiple-channel coding". In: *Proceedings of the 19th ACM international conference on Multimedia.* ACM. 2011, pp. 573–582.

[51] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. "Aggregating local image descriptors into compact codes". In: *IEEE transactions on pattern analysis and machine intelligence* 34.9 (2012), pp. 1704–1716.

[52] S. Lepsøy, G. Francini, G. Cordara, and P. P. B. de Gusmao. "Statistical modelling of outliers for fast visual search". In: *Multimedia and Expo (ICME), 2011 IEEE International Conference on.* IEEE. 2011, pp. 1–6.

[53] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. "A comparison of affine region detectors". In: *International journal of computer vision* 65.1-2 (2005), pp. 43–72.

[54] K. Mikolajczyk and C. Schmid. "A performance evaluation of local descriptors". In: *IEEE transactions on pattern analysis and machine intelligence* 27.10 (2005), pp. 1615–1630.

[55] K. Mikolajczyk and C. Schmid. "An affine invariant interest point detector". In: *Computer Vision—ECCV 2002* (2002), pp. 128–142.

[56] K. Mikolajczyk and C. Schmid. "Scale & affine invariant interest point detectors". In: *International journal of computer vision* 60.1 (2004), pp. 63–86.

[57] F. Schaffalitzky and A. Zisserman. "Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?"" In: *Computer Vision—ECCV 2002* (2002), pp. 414–431.

[58] 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 (2004), pp. 761–767.

[59] T. Tuytelaars and L. Van Gool. “Content-based image retrieval based on local affinely invariant regions”. In: *Visual Information and Information Systems*. Springer. 1999, pp. 656–656.

[60] T. Tuytelaars and L. J. Van Gool. “Wide Baseline Stereo Matching based on Local, Affinely Invariant Regions.” In: *BMVC*. Vol. 412. 2000.

[61] D. G. Lowe. “Object Recognition from Local Scale-Invariant Features”. In: *Proceedings of the International Conference on Computer Vision* 2 (2001), pp. 1150–1157.

[62] D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60.2 (2004), pp. 91–110. DOI: 10.1023/B:VISI.0000029664.99615.94.

[63] H. Bay, T. Tuytelaars, and L. Van Gool. “Surf: Speeded up robust features”. In: *Computer vision–ECCV 2006* (2006), pp. 404–417.

[64] 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 (1981), pp. 381–395.

[65] O. Chum and J. Matas. “Matching with PROSAC-progressive sample consensus”. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE. 2005, pp. 220–226.

[66] T. Sattler, B. Leibe, and L. Kobbelt. “SCRAMSAC: Improving RANSAC's efficiency with a spatial consistency filter”. In: *Computer vision, 2009 ieee 12th international conference on*. IEEE. 2009, pp. 2090–2097.

[67] R. I. Hartley. “In defense of the eight-point algorithm”. In: *IEEE Transactions on pattern analysis and machine intelligence* 19.6 (1997), pp. 580–593.

[68] E. Parzen. “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[69] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. “Total recall: Automatic query expansion with a generative feature model for object retrieval”. In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE. 2007, pp. 1–8.

[70] Website: VLFeat.org. *Documentation Matlab API*. http://www.vlfeat.org/matlab/matlab.html. Online; last visited on June 28, 2017.

[71] Website: Matlab Documentation. *MSER*. https://www.mathworks.com/help/vision/ref/detectmserfeatures.html. Online; last visited on July 28, 2017.

[72] Website: Matlab Documentation. *RANSAC*. https://www.mathworks.com/help/vision/ref/estimategeometrictransform.html. Online; last visited on July 28, 2017.

[73] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. "Movimash: online mobile video mashup". In: *ACM International Conference on Multimedia*. 2012, pp. 139–148. DOI: 10.1145/2393347.2393373.

[74] S. Bano and A. Cavallaro. "ViComp: composition of user-generated videos". In: *Multimedia Tools and Applications* 75 (12 2016), pp. 7187–7210. DOI: 10.1007/s11042-015-2641-2.

[75] T. Weyand and B. Leibe. "Visual landmark recognition from internet photo collections: A large-scale evaluation". In: *Computer Vision and Image Understanding* 135 (2015), pp. 1–15.

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

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

[78] L. V. D. Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605.

[79] A. Karpathy and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137.

[80] E. Zheng and C. Wu. "Structure from motion using structure-less resection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2075–2083.

[81] A. Oliva and A. Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope". In: *International Journal of Computer Vision* 42.3 (2001), pp. 145–175.

[82] I. Simon, N. Snavely, and S. M. Seitz. "Scene summarization for online image collections". In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* IEEE. 2007, pp. 1–8.

[83] G. Kim, L. Sigal, and E. P. Xing. "Joint summarization of large-scale collections of web images and videos for storyline reconstruction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2014, pp. 4225–4232.

[84] J. B. Kruskal. "On the shortest spanning subtree of a graph and the traveling salesman problem". In: *Proceedings of the American Mathematical society* 7.1 (1956), pp. 48–50.

[85] Website: Youtube. *Scene Summarization.* `https://youtu.be/blrA10H06TM`. Online; last visited on August 2, 2017.

[86] D. Zhang, J. Han, C. Li, and J. Wang. "Co-saliency detection via looking deep and wide". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015, pp. 2994–3002.

[87] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han. "A self-paced multiple-instance learning framework for co-saliency detection". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2015, pp. 594–602.

[88] Y. Luo, M. Jiang, Y. Wong, and Q. Zhao. "Multi-camera saliency". In: *IEEE transactions on pattern analysis and machine intelligence* 37.10 (2015), pp. 2057–2070.

[89] N. Bruce and J. Tsotsos. "Saliency based on information maximization". In: *Advances in neural information processing systems.* 2006, pp. 155–162.

[90] H. Soo Park and J. Shi. "Social saliency prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2015, pp. 4777–4785.

[91] D. Basak, S. Pal, and D. C. Patranabis. “Support vector regression”. In: *Neural Information Processing-Letters and Reviews* 11.10 (2007), pp. 203–224.

[92] D. Onoro-Rubio and R. J. López-Sastre. “Towards perspective-free object counting with deep learning”. In: *European Conference on Computer Vision.* Springer. 2016, pp. 615–629.

[93] A. Borji, M. Cheng, H. Jiang, and J. Li. “Salient object detection: A benchmark”. In: *IEEE Transactions on Image Processing* 24.12 (2012), pp. 5706–5722.

[94] L. Itti, C. Koch, and E. Niebur. “A model of saliency-based visual attention for rapid scene analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998), pp. 1254–1259.

[95] L. Itti and C. Koch. “Computational modelling of visual attention”. In: *Nature reviews neuroscience* 2.3 (2001), pp. 194–203.

[96] A. M. Treisman and G. Gelade. “A feature-integration theory of attention”. In: *Cognitive psychology* 12.1 (1980), pp. 97–136.

[97] A. Furnari, G. M. Farinella, and S. Battiato. “An experimental analysis of saliency detection with respect to three saliency levels”. In: *European Conference on Computer Vision.* Springer. 2014, pp. 806–821.

[98] S. Frintrop, T. Werner, and G. Martin Garcia. “Traditional saliency reloaded: A good old model in new shape”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2015, pp. 82–90.

[99] X. Hou, J. Harel, and C. Koch. “Image signature: Highlighting sparse salient regions”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.1 (2012), pp. 194–201.

[100] J. Harel, C. Koch, and P. Perona. “Graph-based visual saliency”. In: *Advances in neural information processing systems.* 2007, pp. 545–552.

[101] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. “Saliency from hierarchical adaptation through decorrelation and variance normalization”. In: *Image and Vision Computing* 30.1 (2012), pp. 51–64.

[102] X. Hou and L. Zhang. “Saliency detection: A spectral residual approach”. In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE. 2007, pp. 1–8.

[103] S. Goferman, L. Zelnik-Manor, and A. Tal. “Context-aware saliency detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.10 (2012), pp. 1915–1926.

[104] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. “Frequency-tuned salient region detection”. In: *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE. 2009, pp. 1597–1604.

[105] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. “Automatic salient object segmentation based on context and shape prior.” In: *BMVC*. Vol. 6. 7. 2011, p. 9.

[106] L. Mai, Y. Niu, and F. Liu. “Saliency aggregation: A data-driven approach”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 1131–1138.

[107] L. Wang, H. Lu, X. Ruan, and M. Yang. “Deep networks for saliency detection via local estimation and global search”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3183–3192.

[108] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. “Minimum barrier salient object detection at 80 fps”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1404–1412.

[109] R. Zhao, W. Ouyang, H. Li, and X. Wang. “Saliency detection by multi-context deep learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1265–1274.

[110] G. Li and Y. Yu. “Visual saliency based on multiscale deep features”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5455–5463.

[111] Z. Wang, G. Xu, Z. Wang, and C. Zhu. “Saliency detection integrating both background and foreground information”. In: *Neurocomputing* 216 (2016), pp. 468–477.

[112] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. "Multi-level Net: A Visual Saliency Prediction Model". In: *European Conference on Computer Vision*. Springer. 2016, pp. 302–315.

[113] Z. Hu, Z. Zhang, Z. Sun, and S. Zhao. "Saliency Detection based on Global Color Distribution and Active Contour Analysis." In: *KSII Transactions on Internet & Information Systems* 10.12 (2016).

[114] L. Zhang, J. Li, and H. Lu. "Saliency detection via extreme learning machine". In: *Neurocomputing* 218 (2016), pp. 103–112.

[115] B. Zou, Q. Liu, Z. Chen, S. Liu, and X. Zhang. "Saliency detection using boundary information". In: *Multimedia Systems* 22.2 (2016), pp. 245–253.

[116] H. Li, J. Chen, H. Lu, and Z. Chi. "CNN for saliency detection with low-level feature integration". In: *Neurocomputing* 226 (2017), pp. 212–220.

[117] R. Arya, N. Singh, and R. K. Agrawal. "A novel combination of second-order statistical features and segmentation using multi-layer superpixels for salient object detection". In: *Applied Intelligence* 46.2 (2017), pp. 254–271.

[118] Y. Zhang, F. Zhang, and L. Guo. "Saliency detection by selective color features". In: *Neurocomputing* 203 (2016), pp. 34–40.

[119] Q. Duan, T. Akram, P. Duan, and X. Wang. "Visual saliency detection using information contents weighting". In: *Optik-International Journal for Light and Electron Optics* 127.19 (2016), pp. 7418–7430.

[120] X. Huang, C. Shen, X. Boix, and Q. Zhao. "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 262–270.

[121] Website: mit saliency benchmark. *About*. http://saliency.mit.edu/home.html. Online; last visited on August 9, 2017.

[122] S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravi. "Saliency-based selection of gradient vector flow paths for content aware image resizing". In: *IEEE Transactions on Image Processing* 23.5 (2014), pp. 2081–2095.

[123] Y. Fang, Z. Chen, W. Lin, and C. Lin. “Saliency detection in the compressed domain for adaptive image retargeting”. In: *IEEE Transactions on Image Processing* 21.9 (2012), pp. 3888–3901.

[124] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, and T. Liu. “Visual saliency based object tracking”. In: *Asian conference on computer vision*. Springer. 2009, pp. 193–203.

[125] Q. Li, Y. Zhou, and J. Yang. “Saliency based image segmentation”. In: *Multimedia Technology (ICMT), 2011 International Conference on*. IEEE. 2011, pp. 5068–5071.

[126] L. Itti and C. Koch. “Feature combination strategies for saliency-based visual attention systems”. In: *Journal of Electronic imaging* 10.1 (2001), pp. 161–169.

[127] T. Judd, K. Ehinger, F. Durand, and A. Torralba. “Learning to predict where humans look”. In: *Computer Vision, 2009 IEEE 12th international conference on*. IEEE. 2009, pp. 2106–2113.

[128] M. Jiang, S. Huang, J. Duan, and Q. Zhao. “Salicon: Saliency in context”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1072–1080.

[129] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. “Learning to detect a salient object”. In: *IEEE Transactions on Pattern analysis and machine intelligence* 33.2 (2011), pp. 353–367.

[130] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. Hu. “Salient object detection and segmentation”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (2011), pp. 1–1.

[131] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang. “Saliency detection via graph-based manifold ranking”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3166–3173.

[132] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen. “A data-driven metric for comprehensive evaluation of saliency models”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 190–198.

[133] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. “What do different evaluation metrics tell us about saliency models?” In: *arXiv preprint arXiv:1604.03605* (2016).

[134] J. Benesty, J. Chen, Y. Huang, and I. Cohen. “Pearson correlation coefficient”. In: *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[135] T. Mauthner, H. Possegger, G. Waltner, and H. Bischof. “Encoding based saliency detection for videos and images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 2494–2502.

[136] Z. Jiang, Z. Lin, and L. S. Davis. “Learning a discriminative dictionary for sparse coding via label consistent K-SVD”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE. 2011, pp. 1697–1704.

[137] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. “Solving the multiple instance problem with axis-parallel rectangles”. In: *Artificial intelligence* 89.1 (1997), pp. 31–71.

[138] C. Zhang, H. Li, X. Wang, and X. Yang. “Cross-scene crowd counting via deep convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 833–841.

[139] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. “Caffe: Convolutional architecture for fast feature embedding”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 675–678.

[140] Website: Matlab Documentation. *fitrlinear*. `https://www.mathworks.com/help/stats/fitrlinear.html`. Online; last visited on August 11, 2017.

[141] S. J. Wright and J. Nocedal. “Numerical optimization”. In: *Springer Science* 35.67-68 (1999), p. 7.

[142] Website: Matlab Documentation. *fitrsvm*. `https://www.mathworks.com/help/stats/fitrsvm.html`. Online; last visited on August 11, 2017.

[143] Website: Matlab Documentation. *predict*. `https://www.mathworks.com/help/stats/compactlinearmodel.predict.html`. Online; last visited on August 11, 2017.

[144] Website: Github. *Counting CNN and Hydra CNN*. `https://github.com/gramuah/ccnn`. Online; last visited on September 7, 2017.

[145] M. Mancuso and S. Battiato. "An introduction to the digital still camera technology". In: *ST Journal of System Research* 2.2 (2001).

[146] G. Finlayson and G. Schaefer. "Colour indexing across devices and viewing conditions". In: *International Workshop on Content-Based Multimedia Indexing*. 2001.

[147] G. Finlayson, S. Hordley, G. Schaefer, and G. Y. Tian. "Illuminant and device invariant colour using histogram equalisation". In: *Pattern recognition* 38.2 (2005), pp. 179–190. DOI: `10.1016/j.patcog.2004.04.010`.

[148] I. Arev, H. S. Park, Y. Sheikh, J. Hodgins, and A. Shamir. "Automatic editing of footage from multiple social cameras". In: *ACM Transactions on Graphics* 33.81 (4 2014). DOI: `10.1145/2601097.2601198`.

[149] H. S. Park, E. Jain, and Y. Sheikh. "3D social saliency from head-mounted cameras". In: *Advances in Neural Information Processing Systems*. 2012, pp. 431–439.

[150] Y. Hoshen, G. Ben-Artzi, and S. Peleg. "Wisdom of the crowd in egocentric video curation". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, pp. 587–593. DOI: `10.1109/CVPRW.2014.90`.

[151] A. Mittal, A. K. Moorthy, and A. C. Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.

[152] A. Nagasaka and T. Miyatake. "Real-time video mosaics using luminance-projection correlation". In: *Trans. IEICE* (1999), pp. 1572–1580.

[153] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, and S. Battiato. "Representing Scenes for Real–Time Context Classification on Mobile Devices". In: *Pattern Recognition* 48.4 (2015), pp. 1086–1100. DOI: `10.1016/j.patcog.2014.05.014`.

[154] G. M. Farinella and S. Battiato. "Scene classification in compressed and constrained domain". In: *IET computer vision* 5.5 (2011), pp. 320–334. DOI: `10.1049/iet-cvi.2010.0056`.

[155] F. Naccari, S. Battiato, A. Bruna, A. Capra, and A. Castorina. “Natural scenes classification for color enhancement”. In: *IEEE Transactions on Consumer Electronics* 51.1 (2005), pp. 234–239.

[156] L. Dohyoung and K. N. Plataniotis. “A Taxonomy of Color Constancy and Invariance Algorithm.” In: *Advances in Low-Level Color Image Processing*. 2014, pp. 55–94.

[157] J. Domke and Y. Aloimonos. “Deformation and viewpoint invariant color histograms”. In: *British Machine Vision Conference*. 2006, pp. 509–518.

[158] Website: RECfusion. *RECfusion Dataset 2015*. http://www.recfusionproject.altervista.org/. Online; last visited on August 1, 2017.

[159] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys. “Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos.” In: *ACM Transactions on Graphics*. 2010, pp. 1–11.

[160] Website: RECfusion. *Cluster Tracking*. http://iplab.dmi.unict.it/recfusionICIAP17. Online; last visited on August 1, 2017.

[161] M. F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, F. L. M. Milotta, S. Schiavone, and F. Stanco. “Archaeometric characterization and 3D survey: new perspectives for monitoring and valorization of Morgantina silver Treasure (Sicily)”. In: (2015).

[162] M. F. Alberghina, F. Alberghina, D. Allegra, F. Di Paola, L. Maniscalco, G. Milazzo, F. L. M. Milotta, L. Pellegrino, S. Schiavone, and F. Stanco. “Integrated three-dimensional models for noninvasive monitoring and valorization of the Morgantina silver treasure (Sicily)”. In: *Journal of Electronic Imaging* 26.1 (2017), pp. 011015–011015.

[163] D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F. L. M. Milotta, C. Santagati, and F. Stanco. “Hand Held 3D Scanning for Cultural Heritage: Experimenting Low Cost Structure Sensor Scan”. In: *Handbook of Research on Emerging Technologies for Architectural and Archaeological Heritage*. IGI Global, 2017, pp. 475–499.

[164] D. Allegra, G. Gallo, L. Inzerillo, M. Lombardo, F. L. M. Milotta, C. Santagati, and F. Stanco. “Low cost handheld 3D scanning for architectural elements acquisition”. In: *Proceedings of the Conference on Smart Tools and Applications in Computer Graphics*. Eurographics Association. 2016, pp. 127–131.

[165] F. Stanco, D. Tanasi, D. Allegra, and F. L. M. Milotta. “3D digital imaging for knowledge dissemination of Greek archaic statuary”. In: *Proceedings of the Conference on Smart Tools and Applications in Computer Graphics*. Eurographics Association. 2016, pp. 133–141.

[166] F. Stanco, D. Tanasi, D. Allegra, F. L. M. Milotta, G. Lamagna, and G. Monterosso. “Virtual anastylosis of Greek sculpture as museum policy for public outreach and cognitive accessibility”. In: *Journal of Electronic Imaging* 26.1 (2017), pp. 011025–011025. DOI: 10.1117/1.JEI.26.1.011025.

[167] Website: structure.io. *Home Page*. http://structure.io/. Online; last visited on July 29, 2017.

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

[169] G. Gallo, F. Milanese, E. Sangregorio, 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 (2010), pp. 93–97.

[170] F. Stanco, D. Tanasi, M. Buffa, and B. Basile. “Augmented perception of the past: The case of the telamon from the greek theater of syracuse”. In: *Multimedia for Cultural Heritage* (2012), pp. 126–135.

[171] F. Stanco, D. Tanasi, G. Gallo, M. Buffa, and B. Basile. “Augmented Perception of the Past-The Case of Hellenistic Syracuse.” In: *Journal of Multimedia* 7.2 (2012).

[172] F. Stanco and D. Tanasi. “Beyond virtual replicas: 3D modeling and maltese prehistoric architecture”. In: *Journal of Electrical and Computer engineering* 2013 (2013), p. 11.

[173] C. Santagati, L. Inzerillo, and F. Di Paola. “Image-based modeling techniques for architectural heritage 3D digitalization: limits and potentialities”. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 5.w2 (2013), pp. 555–560.

[174] F. Remondino. “Heritage recording and 3D modeling with photogrammetry and 3D scanning”. In: *Remote Sensing* 3.6 (2011), pp. 1104–1138.

[175] A. Bandiera, J. A. Beraldin, and M. Gaiani. “[ITA] Nascita ed utilizzo delle tecniche digitali di 3D imaging, modellazione e visualizzazione per l’architettura ei beni culturali”. In: *Ikhnos”. Lombardi editore, Siracusa* (2011), pp. 81–134.

[176] B. Benedetti, M. Gaiani, and F. Remondino. *[ITA] Modelli digitali 3D in archeologia: il caso di Pompei.* Edizioni della Normale, 2010.

[177] B. Curless. “From range scans to 3D models”. In: *ACM SIGGRAPH Computer Graphics* 33.4 (1999), pp. 38–41.

[178] N. Pears, Y. Liu, and P. Bunting. *3D imaging, analysis and applications.* Vol. 3. Springer, 2012.

[179] G. Vosselman and H. Maas. *Airborne and terrestrial laser scanning.* CRC Press, 2010.

[180] E. Cappelletto, P. Zanuttigh, and G. M. Cortelazzo. “3D scanning of cultural heritage with consumer depth cameras”. In: *Multimedia Tools and Applications* 75.7 (2016), pp. 3631–3654.

[181] I. Toschi, A. Capra, L. De Luca, J. 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* 2.5 (2014), p. 371.

[182] 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 (2008), pp. 129–136.

[183] L. Inzerillo and C. Santagati. “[ITA] Il progetto del rilievo nell’utilizzo di tecniche di modellazione dense stereo matching”. In: (2013).

[184] 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 (2016), p. 178.

[185] M. Russo and A. M. Manferdini. “Integrated multi-scalar approach for 3d cultural heritage acquisitions”. In: *Handbook of Research on Emerging Digital Tools for Architectural Surveying, Modeling, and Representation* 337.3 (2015).

[186] P. Cignoni, C. Rocchini, and R. Scopigno. “Metro: Measuring error on simplified surfaces”. In: *Computer Graphics Forum*. Vol. 17. 2. Wiley Online Library. 1998, pp. 167–174.

[187] Website: Leica Geosystems. *Leica HDS3000*. `http://hds.leica-geosystems.com/en/5574.htm`. Online; last visited on August 3, 2017.

[188] M. Callieri, P. Cignoni, M. Dellepiane, and R. Scopigno. “Pushing Time-of-Flight Scanners to the Limit.” In: *VAST*. 2009, pp. 85–92.

[189] M. Bell. *[ITA] La provenienza ritrovata: cercando il contesto di antichità trafugate*. 1997.

[190] M. Bell. “[ITA] Il tesoro di argenteria e la casa di Eupolemos a Morgantina”. In: *Sacri agli dei, Argenti della casa di Eupolemos a Morgantina. Il rientro* (2012), pp. 23–25.

[191] P. G. Guzzo. “A group of Hellenistic silver objects in the Metropolitan Museum”. In: *Metropolitan Museum Journal* 38 (2003), pp. 45–8.

[192] G. Marchand, E. Guilminot, S. Lemoine, L. Rossetti, M. Vieau, and N. Stephant. “Degradation of archaeological horn silver artefacts in burials”. In: *Heritage Science* 2.1 (2014), p. 5.

[193] D. Allegra, E. Ciliberto, P. Ciliberto, F. L. M. Milotta, G. Petrillo, F. Stanco, and C. Trombatore. “Virtual unrolling using x-ray computed tomography”. In: *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE. 2015, pp. 2864–2868.

[194] F. Stanco, S. Battiato, and G. Gallo. *Digital imaging for cultural heritage preservation: Analysis, restoration, and reconstruction of ancient artworks*. CRC Press, 2011.

[195] Website: Unity. *Home Page*. https://unity3d.com/. Online; last visited on July 28, 2017.

[196] Website: Smithsonian Institution. *Digitization — 3D*. https://3d.si.edu/. Online; last visited on July 28, 2017.

[197] Website: Sketchfab. *Home Page*. https://sketchfab.com/. Online; last visited on July 28, 2017.

[198] Y. Liao, M. Lezoche, H. Panetto, and N. Boudjlida. "Why, where and how to use semantic annotation for systems interoperability". In: *1st UNITE Doctoral Symposium*. 2011, pp. 71–78.

[199] E. Oren, K. Möller, S. Scerri, S. Handschuh, and M. Sintek. "What are semantic annotations". In: *Relatório técnico. DERI Galway* 9 (2006), p. 62.

[200] Website: Morgantina Silver Treasure. *Unity Viewer*. http://yoda.dmi.unict.it/morgantinaJournal/index.htm. Online; last visited on October 21, 2016.

[201] G. M. A. Richter and I. A. Richter. *Kouroi: archaic Greek youths: a study of the development of the Kouros type in Greek sculpture*. Phaidon Press, 1970.

[202] A. Anderson. "Beauty and Truth". In: *Boston: Agora* (2007).

[203] E. De Miro. "[ITA] La scultura siceliota in Sicilia nell'età classica". In: *G. Pugliese Carratelli (a cura di), I Greci in Occidente. Catalogo, II edizione, Milano*. 1996.

[204] R. Panvini and L. Sole. *[ITA] La Sicilia in età arcaica: dalle apoikiai al 480 a. c.* Centro regionale per l'inventario, la catalogazione e la documentazione, 2009.

[205] S. Pafumi. "[ITA] Le antichità del principe di Biscari: scelte e criteri espositivi di un collezionista tra antiquaria e nuova scienza archeologica". In: *Oggetti, uomini, idee, a cura di G. Giarrizzo e S. Pafumi, Pisa Roma*. 2009, pp. 87–116.

[206] G. Libertini. *[ITA] Il Museo Biscari*. Vol. 1. Casa editrice d'arte Bestetti e Tumminelli, 1930.

[207] G. Libertini and Museo Comunale (Catania, Italy). *[ITA] Il Castello Ursino e le raccolte artistiche comunale di Catania*. Tip. Zuccarello & Izzi, 1937.

[208] G. V. Gentili. “[ITA] I due kouroi da Osimo e i tre kouroi del vecchio museo Archeologico di Siracusa nello studio e ricordo di Luigi Bernabò Brea”. In: *M. Cavalier, M. Bernabò Brea (a cura di), In memoria di Luigi Bernabò Brea, Palermo*. 2002, pp. 67–106.

[209] F. Delli Ponti, A. Guidazzoli, S. Imboden, and M. C. Liguori. “A Blender open pipeline for a 3D animated historical short film”. In: (2011).

[210] Website: Kouros. *Web Viewer*. https://yoda.dmi.unict.it/kouros/. Online; last visited on August 3, 2017.

[211] V. Bruni, A. Crawford, D. Vitulano, and F. Stanco. “Visibility based detection and removal of semi-transparent blotches on archived documents.” In: *VISAPP (1)*. 2006, pp. 64–71.

[212] H. J. Chatterjee. *Touch in museums: Policy and practice in object handling*. Berg, 2008.

[213] D. Romanek and B. Lynch. “Touch and the value of object handling: Final conclusions for a new sensory museology”. In: *Touch in museums: Policy and practice in object handling* (2008), pp. 275–286.

[214] Website: Center for Virtualization and Applied Spatial Technologies (CVAST). *Home*. https://cvast.usf.edu/. Online; last visited on July 29, 2017.

[215] I. Onol. “Tactual Explorations: A tactile interpretation of a museum exhibit through tactile artworks and Augmented Reality”. In: (2006).

[216] F. Stanco, D. Tanasi, G. C. Guarnera, and G. Gallo. “Automatic Classification of Decorative Patterns in the Minoan Pottery of Kamares Style”. In: *Pattern Recognition and Signal Processing in Archaeometry: Mathematical and Computational Solutions for Archaeology*. IGI Global, 2012, pp. 186–211.

[217] D. Huber, B. Akinci, P. Tang, A. Adan, B. Okorn, and X. Xiong. “Using laser scanners for modeling and analysis in architecture, engineering, and construction”. In: *Conference on Information Sciences and Systems (CISS)*. 2010, pp. 1–6. DOI: 10.1109/CISS.2010.5464818.

[218] J. Stoll, P. Novotny, R. Howe, and P. Dupont. “Real-time 3D ultrasound-based servoing of a surgical instrument”. In: *International Conference on Robotics and Automation (ICRA)*. 2006, pp. 613–618. DOI: 10.1109/ROBOT.2006.1641778.

[219] A. Bottino, M. De Simone, A. Laurentini, and C. Sforza. “A New 3-D Tool for Planning Plastic Surgery”. In: *IEEE Transactions on Biomedical Engineering* 59.12 (2012), pp. 3439–3449. DOI: 10.1109/TBME.2012.2217496.

[220] P. Treleaven and J. Wells. “3D Body Scanning and Healthcare Applications”. In: *Computer* 40.7 (2007), pp. 28–34. DOI: 10.1109/MC.2007.225.

[221] Y. Dai, J. Tian, D. Dong, G. Yan, and H. Zheng. “Real-Time Visualized Freehand 3D Ultrasound Reconstruction Based on GPU”. In: *IEEE Transactions on Information Technology in Biomedicine* 14.6 (2010), pp. 1338–1345. DOI: 10.1109/TITB.2010.2072993.

[222] R. Laing, M. Leon, and J. Isaacs. “Monuments Visualization: From 3D Scanned Data to a Holistic approach, an Application to the City of Aberdeen”. In: *International Conference on Information Visualisation*. 2015, pp. 512–517. DOI: 10.1109/iV.2015.91.

[223] C. V. Nguyen, J. Fripp, D. R. Lovell, R. Furbank, P. Kuffner, H. Daily, and X. Sirault. “3D Scanning System for Automatic High-Resolution Plant Phenotyping”. In: *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2016, pp. 1–8. DOI: 10.1109/DICTA.2016.7796984.

[224] K. Pearson. “On lines and planes of closest fit to systems of points in space”. In: *Philosophical Magazine* 2.6 (1901), pp. 559–572.

[225] B. Allen, B. Curless, and Z. Popović. “The space of human body shapes: reconstruction and parameterization from range scans”. In: *International Conference on Computer Graphics and Interactive Techniques*. 2003, pp. 587–594.

[226] G. Gallo, G. C. Guarnera, and G. Catanuto. “Human Breast Shape Analysis using PCA.” In: *BIOSIGNALS*. 2010, pp. 163–167.

[227] D. J. Smith Jr, W. E. Palin Jr, V. L. Katch, and J. E. Bennett. "Breast volume and anthropomorphic measurements: normal values." In: *Plastic and reconstructive surgery* 78.3 (1986), pp. 331–335.

[228] G. M. Farinella, G. Impoco, G. Gallo, S. Spoto, and G. Catanuto. "Unambiguous Analysis of Woman Breast Shape for Plastic Surgery Outcome Evaluation." In: *eurographics Italian chapter conference*. 2006, pp. 255–261.

[229] G. Catanuto, G. Gallo, G. M. Farinella, G. Impoco, M. Nava, A. Pennati, and A. Spano. "Breast shape analysis on three-dimensional models". In: *Third European Conference on Plastic and Reconstructive Surgery of the Breast*. 2005.

[230] G. M. Galdino, M. Nahabedian, M. Chiaramonte, J. Z. Geng, S. Klatsky, and P. Manson. "Clinical applications of three-dimensional photography in breast surgery". In: *Plastic and Reconstructive Surgery* 110.1 (2002), pp. 58–70.

[231] M. Y. Nahabedian and G. Galdino. "Symmetrical breast reconstruction: is there a role for three-dimensional digital photography?" In: *Plastic and reconstructive surgery* 112.6 (2003), pp. 1582–1590.

[232] H. Y. Lee, K. Hong, and E. A. Kim. "Measurement protocol of women's nude breasts using a 3d scanning technique". In: *Applied Ergonomics* 35 (2004), 353—360.

[233] F. L. Bookstein. "Thin-plate splines and the atlas problem for biomedical images". In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 1991, pp. 326–342.

[234] A. Rosas and M. Bastir. "Thin-plate spline analysis of allometry and sexual dimorphism in the human craniofacial complex". In: *American Journal of Physical Anthropology* 117.3 (2002), pp. 236–245.

[235] F. L. Bookstein. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge, 1997.

[236] Website: Wolfram Mathworld. *Rodrigues' Rotation Formula*. http://mathworld.wolfram.com/RodriguesRotationFormula.html. Online; last visited on July 29, 2017.

[237] R. C. Gonzalez and R. E. Woods. "Digital image processing prentice hall". In: *Upper Saddle River, NJ* (2002).

[238] Website: RAI (Radio Televisione Italiana) - RaiPlay. *Superquark, St.2017, Puntata del 21/06/2017*. http://www.raiplay.it/video/2017/06/SuperQuark-del-21-06-2017-69861f8d-2f15-455c-ab17-af8a49b18c62.html. Online; created June 21, 2017; last visited on July 20, 2017.

[239] A. H. Munsell. *Atlas of the Munsell color system*. Wadsworth, Howland & Company, Incorporated, Printers, 1915.

[240] D. L. MacAdam. "The theory of the maximum visual efficiency of colored materials". In: *JOSA* 25.8 (1935), pp. 249–252.

[241] ASTM. "Standard Practice for Specifying Color by the Munsell System". In: *ASTM International D 1535-14* (2014).

[242] B. R. Conway and M. S. Livingstone. "A different point of hue". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.31 (2005), pp. 10761–10762.

[243] Soil Survey Staff USA. *Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys*. US Government Printing Office, 1975.

[244] E. Michéli, P. Schad, and O. Spaargaren. *World Reference Base for Soil Resources 2006: A Framework for International Classification, Correlation and Communication*. Food and agriculture organization of the United nations (FAO), 2006.

[245] M. Sánchez-Marañón, M. Soriano, M. Melgosa, G. Delgado, and R. Delgado. "Quantifying the effects of aggregation, particle size and components on the colour of Mediterranean soils". In: *European Journal of Soil Science* 55.3 (2004), pp. 551–565. ISSN: 1365-2389. DOI: 10.1111/j.1365-2389.2004.00624.x. URL: http://dx.doi.org/10.1111/j.1365-2389.2004.00624.x.

[246] R. R. Gerharz, R. Lantermann, and D. R. Spennemann. "Munsell color charts: A necessity for archaeologists?" In: *The Australian Journal of Historical Archaeology* (1988), pp. 88–95.

[247] C. Goodwin. “Practices of color classification”. In: *Mind, culture, and activity* 7.1-2 (2000), pp. 19–36.

[248] S. Aydemir, S. Keskin, and L. R. Drees. “Quantification of soil features using digital image processing (DIP) techniques”. In: *Geoderma* 119.1 (2004), pp. 1–8.

[249] R. A. V. Rossel, Y. Fouad, and C. Walter. “Using a digital camera to measure soil organic carbon and iron contents”. In: *Biosystems Engineering* 100.2 (2008), pp. 149–159.

[250] F. Stanco, D. Tanasi, A. Bruna, and V. Maugeri. “Automatic Color Detection of Archaeological Pottery with Munsell System”. In: *In proceedings of 16th International Conference Image Analysis and Processing ICIAP 2011.* Springer. 2011, pp. 337–346.

[251] T. K. O’Donnell, K. W. Goyne, R. J. Miles, C. Baffaut, S. H. Anderson, and K. A. Sudduth. “Determination of representative elementary areas for soil redoximorphic features identified by digital image processing”. In: *Geoderma* 161.3 (2011), pp. 138–146.

[252] F. Stanco, D. Tanasi, A. M. Gueli, and G. Stella. “Computer graphics solutions for dealing with colors in archaeology”. In: *Conference on Colour in Graphics, Imaging, and Vision.* Vol. 2012. 1. Society for Imaging Science and Technology. 2012, pp. 97–101.

[253] F. Stanco and A. M. Gueli. “Computer graphics solutions for pottery colors specification”. In: *IS&T/SPIE Electronic Imaging.* International Society for Optics and Photonics. 2013, 86600S–86600S.

[254] L. Gómez-Robledo, N. López-Ruiz, M. Melgosa, A. J. Palma, L. F. Capitán-Vallvey, and M. Sánchez-Marañón. “Using the mobile phone as Munsell soil-colour sensor: an experiment under controlled illumination conditions”. In: *Computers and Electronics in Agriculture* 99 (2013), pp. 200–208.

[255] P. Han, D. Dong, X. Zhao, L. Jiao, and Y. Lang. “A smartphone-based soil color sensor: For soil type classification”. In: *Computers and Electronics in Agriculture* 123 (2016), pp. 232–241.

[256] F. Stanco, L. Tenze, G. Ramponi, and A. De Polo. “Virtual restoration of fragmented glass plate photographs”. In: *Electrotechnical Conference, 2004. MELECON 2004. Proceedings of the 12th IEEE Mediterranean*. Vol. 1. IEEE. 2004, pp. 243–246.

[257] ASTM. “Standard Practice for Calculation of Color Tolerances and Color Differences from Instrumentally Measured Color Coordinates”. In: *ASTM International D 2244-05* (2005).

[258] ASTM. “Standard Practice for Establishing Color and Gloss Tolerances”. In: *ASTM International D 3134-97* (1997).

[259] M. Sánchez-Marañón, R. Huertas, and M. Melgosa. “Colour variation in standard soil-colour charts”. In: *Soil Research* 43.7 (2005), pp. 827–837.

[260] F. L. M. Milotta, F. Stanco, and D. Tanasi. “ARCA (Automatic Recognition of Color for Archaeology): a Desktop Application for Munsell Estimation”. In: *Proceedings of Image Analysis and Processing - ICIAP 2017 19th International Conference*. 2017, pp. 661–671. DOI: 10.1007/978-3-319-68548-9.

[261] Website: Automatic Recognition of Color for Archaeology (ARCA). *Dataset ARCA108*. http://iplab.dmi.unict.it/ARCA108/. Online; last visited on August 4, 2017.

[262] P. Centore. “An open-source inversion algorithm for the Munsell renotation”. In: *Color Research & Application* 37.6 (2012), pp. 455–464.

[263] Website: The Munsell and Kubelka-Munk Toolbox. *Documentation*. http://www.munsellcolourscienceforpainters.com/MunsellAndKubelkaMunkToolbox/MunsellAndKubelkaMunkToolbox.html. Online; last visited on July 29, 2017.

[264] W. C. Rheinboldt and J. P. Menard. “Mechanized conversion of colorimetric data to Munsell renotations”. In: *JOSA* 50.8 (1960), pp. 802–807.

[265] F. T. Simon and J. A. Frost. “A new method for the conversion of CIE colorimetric data to munsell notations”. In: *Color Research & Application* 12.5 (1987), pp. 256–260.

[266] Website: Wallkill Color. *Home Page*. http://www.wallkillcolor.com/. Online; last visited on July 29, 2017.

[267] Website: BabelColor. *Products*. http://www.babelcolor.com/products.htm#PRODUCTS_PT. Online; last visited on July 26, 2017.

[268] Website: Sketchfab Help Center. *3D File Formats*. https://help.sketchfab.com/hc/en-us/articles/202508396-3D-File-Formats?utm_source=website&utm_campaign=upload_hints. Online; last visited on July 28, 2017.

[269] Website: Autodesk BIM 360 Team. *Pricing*. https://team.bim360.com/pricing/index.html. Online; last visited on July 28, 2017.

[270] Website: Autodesk A360. *A360 viewable file formats*. http://help.autodesk.com/view/ADSK360/ENU/?guid=GUID-488804D0-B0B0-4413-8741-4F5EE0FACC4A. Online; last visited on July 28, 2017.

[271] Website: A360 Forum. *Textures in A360 Viewer*. https://forums.autodesk.com/t5/a360-forum/textures-in-a360-viewer/td-p/6971615. Online; last visited on July 28, 2017.

[272] Website: Autodesk A360. *Reviewing Designs*. http://help.autodesk.com/view/ADSK360/ENU/?guid=GUID-0DAE226A-54EA-48CC-91C4-28F61DF70A59. Online; last visited on July 28, 2017.

[273] Website: Autodesk. *ReMake*. https://remake.autodesk.com/about. Online; last visited on July 28, 2017.

[274] Website: Autodesk. *ReCap*. http://www.autodesk.com/products/recap/overview. Online; last visited on July 28, 2017.

[275] Website: GallerySystems. *Solutions*. http://www.gallerysystems.com/products-and-services/. Online; last visited on July 28, 2017.

[276] Website: 3DHOP. *About*. http://vcg.isti.cnr.it/3dhop/. Online; last visited on July 28, 2017.