



UNIVERSITÀ  
degli STUDI  
di CATANIA

**UNIVERSITÀ DEGLI STUDI DI CATANIA**  
DIPARTIMENTO DI MATEMATICA E INFORMATICA  
DOTTORATO DI RICERCA IN MATEMATICA E INFORMATICA XXXI CICLO

---

*Visual Behavior Analysis in Retail Scenario*

Vito Santarcangelo

---

TESI DI DOTTORATO DI RICERCA

---

Supervisor: Prof. Sebastiano Battiato

---

*“Anche al Sud si può”*

Alberto Camporesi

## *Abstract*

The retail world is today highly competitive and has seen its logics completely revolutionized by the introduction of e-commerce that have prompted a reaction from the retail market, requiring greater attention to the consumer. We therefore moved from the world of traditional marketing (generic flyer) to that of 1to1 marketing (specific attention to the customer, profiling and personalization of the assortment offer). In this context the need arises to introduce innovative tools that can allow the physical sales spaces to be kept competitive, interacting more with the customer in order to create a more relevant commercial proposal. As a consequence, the computer vision represented one of the possible means to carry out the behavioral analysis of the consumer useful for dynamically adapting the assortment proposal. DOOH (*Digital Out Of Home*) in its most widespread form of interactive point-of-sale kiosks is one of the best tools to get in touch with the customer, create a synergy with him, listen to his needs in order to improve the offer, the level of service and therefore customer satisfaction. Next to DOOH, it is necessary to introduce further and time-continuous monitoring tools, which map the entire customer's shopping experience into the point of sale. For this purpose the egocentric vision is introduced through the use of cam narratives on board the trolleys, which allow a timely story of the consumer, called *Visual Market Basket Analysis* (evolution of Market Basket Analysis), which generates process functional alerts to the improvement of the service offered. The story of these approaches is provided in this PhD thesis, which tells the three-year course carried out, its experiments and possible future developments. This study has been conducted thanks to the support of Centro Studi S.r.l., a sister company of a privately owned consumer goods distribution company called Orizzonti Holding Group, located in southern Italy. The study has been implemented through an industrial application approach, in a real context (Futura Supermarkets). Consequently, the PhD thesis has considered the typical difficulties of a challenging environment, starting from the creation and acquisition of a dataset to the integration of the approach in the current business processes.

## *Acknowledgements*

A sincere thank you goes to my tutor Professor Sebastiano Battiato, a fantastic person and a model to be followed. He made this PhD unforgettable due to his professionalism and humanity. In addition, I want to say thank you to Professor Giovanni Maria Farinella, who has always been available and willing to offer his support. These two people made the environment of the University of Catania and the IPLab laboratory unique. The organized summer schools are something that I will always carry in my heart. Thanks also to Professor Alberto Camporesi (special supporter), to Costantino Di Carlo and Valerio Di Carlo, who have financed the PhD and who represent a very high-quality company in southern Italy. I thank my family and Stefania who also always supported me during my PhD. A special thank you is for my scientific guide, Professor Egidio Cascini, who has always supported my passion for research. A further thank you goes to Filippo Stanco, Dario Allegra and Diego Sinitò, good friends who made my days in Catania more pleasant. A hug and a final thank you goes to all those people I met at the department of mathematics and computer science of the University of Catania and in the IPLab laboratory. They are another confirmation of the fact that the people of Sicily are special.



# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Scope of the thesis . . . . .	1
1.1.1 Context . . . . .	2
1.1.2 Marketing 3.0 . . . . .	3
1.1.3 DOOH . . . . .	4
1.1.4 Customer Tracking . . . . .	7
1.2 Papers and Patents . . . . .	10
<b>2 BEHAVIOR ANALYSIS IN RETAIL CONTEXT</b>	<b>13</b>
2.1 DOOH and Retail . . . . .	13
2.2 Face Detection . . . . .	14
2.3 Datasets . . . . .	16
2.4 Gender Recognition . . . . .	20
2.4.1 Face Detection . . . . .	20
2.4.2 Face Representation . . . . .	20
2.4.3 Gender Classification . . . . .	21
2.4.4 State of the Art Results . . . . .	22
2.5 Age Estimation . . . . .	25
2.6 Research Proposal . . . . .	26
2.7 Behavioral Analysis System . . . . .	27
2.8 Benchmark Analysis . . . . .	31
2.9 Google Cloud Platform . . . . .	32
2.9.1 Google Cloud Platform: Image Analysis . . . . .	34
2.9.2 Google Cloud Face and Emotion Detection . . . . .	37

2.10	Embedded Device Tests . . . . .	40
2.11	Thermal Cameras . . . . .	42
2.12	GDPR and DOOH . . . . .	43
<b>3</b>	<b>VISUAL MARKET BASKET ANALYSIS</b>	<b>50</b>
3.1	Market Basket Analysis . . . . .	50
3.1.1	Terminology . . . . .	50
3.1.2	How it works . . . . .	51
3.2	MBA technical approach . . . . .	52
3.2.1	Apriori . . . . .	52
3.2.2	Eclat . . . . .	53
3.2.3	Pattern-growth . . . . .	53
3.2.4	Observation . . . . .	55
3.3	Egocentric vision for Visual Market Basket Analysis . . . . .	56
3.3.1	Methods for VMBA . . . . .	58
3.3.2	Actions . . . . .	59
3.3.3	Location . . . . .	60
3.3.4	Scene Context . . . . .	60
3.3.5	Classifications . . . . .	61
3.3.6	VMBA15 Dataset . . . . .	63
3.3.7	Experimental settings and results . . . . .	64
3.4	Patented System . . . . .	67
3.5	Deep Learning Overview . . . . .	73
3.5.1	CNN structure . . . . .	74
3.5.2	CNN training . . . . .	76
3.5.3	Deep Learning Frameworks . . . . .	82
3.5.4	Code and test . . . . .	84
3.5.5	R-CNN . . . . .	87
3.5.6	Retail CNN Approach . . . . .	90
3.6	VMBA with 14 classes . . . . .	92
3.6.1	Proposed Method . . . . .	94
3.6.2	Experimental Settings and Results . . . . .	97
3.6.3	Overall Classification . . . . .	100
3.7	Conclusion . . . . .	102

<b>4</b>	<b>FUTURE DEVELOPMENT</b>	<b>104</b>
4.1	Introduction . . . . .	104
4.2	DOOH on infrared spectrum . . . . .	104
4.2.1	Face extraction . . . . .	105
4.2.2	Gender Recognition . . . . .	106
4.2.3	Infrared Analysis . . . . .	106
4.3	VMBA Evolution . . . . .	108
4.4	Data Mining and Process Mining . . . . .	113

# Chapter 1

## INTRODUCTION

### 1.1 Scope of the thesis

The thesis focuses on behavioral analysis in the whole of retail scenarios. Today, the retail world is one of the areas of application of the Computer Vision where many possible implementations can be developed. In fact, Computer Vision allows to carry out important developments and analyzes in the world of supermarkets. Today, knowing the consumer is a plus to orient marketing choices and to greatly improve the appeal of their points of sale, which become spaces that are greatly optimized for the 1to1 sale. In this context, the present PhD thesis is divided into two directions aimed at carrying out the customer's 'behavior analysis' within the point of sales. More specifically, to perform behavioral analysis it is necessary to work on two goals: analyzing the interaction of the customer using digital signage devices and tracking the customer inside the point of sale. This context itself lends to an application of the Computer Vision, since it is possible to reach the goals of behavioral analysis of the consumer thanks to the analysis of the multimedia flows acquired through distributed cameras. The first goal is implemented through Computer Vision techniques as Gender/Age Estimation, Emotion Detection, Pose Estimation techniques to map the customer's propensity to purchase through NLP (Neuro-linguistic programming). To achieve this objective, it was necessary to apply algorithms to photos/videos acquired through digital signage devices (interactive totem). The second goal is implemented by using a customized approach of Video Synopsis techniques (video summarization) applied to shopping carts equipped with egocentric camera. Compared to other approaches (e.g., WIFI, BLE), Computer Vision allows to minimize the costs for monitoring the retail scenario, allows for

adaptations to dynamic changes of the store and allows for continuous monitoring of the consumers' experience. The approach described in the PhD thesis has been named Visual Market Basket Analysis. To achieve this target, the videos have been acquired by micro-cameras placed on the carts, in a store of an important supermarket chain in southern Italy, called Futura. The approaches presented in this thesis for the two above mentioned goals are innovative and patented. The 'behavior analysis' is integrated into the ambitious goal of creating a 'visual genome' of the point of sale.

### 1.1.1 Context

The last 20 years are characterized by a rapid evolution in the communication media and consequently a revolution in the Marketing and Communication strategy. The evolution of these recent years is not very far from the Industrial Revolution or the massive spread of televisions in households. Just 20 years ago, Google was not existing, just 10 years ago Facebook and Instagram were not there. The way we consume is rapidly changing: The Observatory of Mobile B2C Strategy from Polytechnic of Milano is highlighting how the Mobile advertising is increasing this year 53% vs 2014 in Italy [8]. Speed is the key word for Companies who want to efficiently communicate with consumers: they need to be able to change, learn and update knowledge in a fast-paced environment. Content & Innovation are the 2 new key pillars of the marketing mix for a successful communication. The real challenge for Companies and Brands in the digital world is to be able to create and maintain a connection with the target audience. They need to deliver a meaningful content that resonate in a turbulent and noisy framework. They need to be able to attract the attention in a world where the competition is not anymore limited to the competitors' products voice, but is enlarged to a set of completely different industries such as television shows, films, videos and even sports events. Digital communication has, in fact, an ambivalent profile, being able to increase visibility for small and local brands or companies in the same way of a big multinational, but also creating a tremendous amount of noise. The pure emotional involvement is not enough anymore if not anchored to the relevant set of values of the addressed target. From the classic consumer centric marketing approach of Marketing 2.0 based mainly

on gender/ age/occupation segmentation, marketing is rapidly evolving to a “values key set” segmentation approach in Marketing 3.0 [9].

### 1.1.2 Marketing 3.0

This approach [10] is more adaptable to the shift from the “one-to-one” communication to the “many-to-many” world of Social Networks, where everybody is able to influence, build, comment, share and shape a product or a service, through his direct action and contribution via the digital media. GFK has developed the ValueScope® model [11] measuring 54 Personal Values demonstrating that:

- Personal Values are at the core of human kind attitudes, motivations and behaviors;
- They encompass all dimensions of life and are directly linked to the choices we make;
- They provide a common ground for understanding differences between and within cultures and demographics;
- They are relatively stable over time and enable to spot gradual societal changes.

Reassembling the values in key “value types” they can be used in innovation processes to develop products that get to the heart of consumers’ needs by considering their aspirations and helping people to meet them. They can be used in communications to establish messages that will deeply resonate with consumers and provide stronger cut through. They can also provide a more insightful way to encourage consumers’ loyalty, as value set is normally more stable in terms of time.

This context is related to the new digital born generation, the so called Generation Y or Millennials, that are the key target for almost all the marketers around the world and the most skeptical to standard advertising and communication. They are considered as the trendsetters’ productive stream of the population. The ones able to rapidly spread words of mouth and declare the ultimate judgment on a product or service success or failure. This generation is constituted by people who were born from early 80’s until early ’00. They are characterized by a natural and extensive use of digital technology and a great familiarity with communication. Being exposed their entire life to TV commercials, magazines and slogans, Millennials are

extremely diffident to standard advertising and just do not trust this way of being approached anymore. They tend to refuse the “one-direction” product communication and their attention is quite low, unless entertained, stimulated and involved in a cocreated content. The attention to content becomes therefore key to establish a relationship with this target, with a process that primarily focus on the key value type identification and adapt the message and the style even dramatically to the media supporting it. Brands Companies are not only deputed to satisfy a need but are asked to actively take part in playing a role to define a value frame for a “better world”. Brand entertainment, communities, story-telling and edutainment are some of the most efficient way to do it. Some of the most famous examples of these trends are Companies as PG who started branding the commercials with the PG logo not anymore only as a quality statement, but as a true values proposition and a community umbrella. Brands as Dove with their “natural Beauty” campaign or Coca-Cola with the new released “Taste the feeling” campaign were able to perform a shift from a gender/need/taste pillars segmentation to a logo used as the carrier of the key brand value and promise. Looking forward it will be key also for the retail industry to start segmenting the audience in a more deep and complex way, delivering more complex content than just “You can find the product X, better displayed and at a better price”. Opportunities of buying products are already not limited to proximity anymore, trust engagement is just a “never-ending-feed” process for new generations and competition is not anymore limited to other retailers. Embracing analytical cultures and invisible analytics, to address this fast changing world can help also the retail industry to become expert at converting that data into business success. The result is improved customer engagement, insight that informs creativity and better ways to customize offers [12]. Invisible analytics (identified by GFK as one of the most prominent technical trends of 2016) could, for example, be the key ingredient in the effective use of augmented and virtual reality in retail environments.

### 1.1.3 DOOH

A possible example of implementation of the presented scenario is DOOH. Digital out of home (DOOH) symbolizes the dynamic media distributed across placed-based



Figure 1.1: Examples of DOOH devices (jumbotrons and interactive kiosk)

networks through digital signage devices [1] as addressable screens, kiosks, jukeboxes and jumbotrons, with the aim to engage customers and extend the effectiveness of marketing experience. An example of DOOH is provided by Futura Point kiosk, appeared in 2012, whose functionalities are patented [2]. Futura Point is a web-oriented kiosk, characterized by a touchscreen monitor, a barcode 1D/2D reader, a printer for coupons, a camera and a computer. DOOH applications are one of the most important topic for retail environment [6], due to the change of traditional marketing into 1to1 approach, providing a custom shopping experience for the user. DOOH has received considerable interest from retailers and governmental institutions because of the benefits obtained in better managing and respond to the preferences of the users [7]. The implementation of DOOH scenario requires more and more technology, then there is an integration of Computer Vision techniques [13], Artificial Intelligence modules and Internet of Things. In fact, Computer vision provides algorithms to automatically collect soft biometrics of people in front of a smart screen or in a delimited controlled area. It is of great interest for industry. These approaches are mainly based on face detection module and involve gender recognition [14], age estimation [25], emotion recognition, gesture and pose detection algorithms [15]. In fact, the input about gender is very important to customize the shopping experience because it is possible to define two defined clusters, one for male and one for female. Thanks to the Age Estimation, it is possible to define some clusters for different age



targets. Combining gender, age and ethnicity input, emotion detection and skeletal tracking analysis, it is possible to improve the dynamicity interactive of DOOH applications. Artificial Intelligence provides semantic networks for a support ontology based on knowledge-base. In an interconnected world of devices, also DOOH is an Internet of Things scenario, in fact, all the interactive interfaces are interconnected sharing information for the dynamic interaction with the user. All these technologies are targeted to the analysis of the behavioral state of the buyer. In our patented approach about behavioral analysis [16], we consider at first the 4 possible “response modes” defined by Miller Heiman in “The New Strategic Selling” [17] that are: Growth, Trouble, Even Keel, Overconfident. These Concepts, specifically developed for B2B Buyer-Vendor transaction cases suggest that: Buyers in Growth mode perceive that by buying a product or service, they will produce growth for their company (opened to vendor offerings). In Trouble mode, buyers perceive that they have to fill rapidly a business gap (e.g. it is losing customers, losing money, decreasing productivity) and understand that something must change (opened to vendor offerings). Buyers in Even Keel mode do not perceive a large enough gap between their position and their goal, then there is no urgency to change anything (closed to vendor offerings). Buyers in Overconfident mode believe that they are doing so well that the suggestion that your offering might improve their situation is practically an insult (closed to vendor offerings). In fact, they have a weak grasp of the reality and their mind is then closed to suggestions before to the natural fall to the trouble mode. These concepts have been developed, applied and conceptually adapted, also to the B2C case and specifically utilized to tailor properly the Communication to each specific Customer Response Mode detected or targeted, considering the related variability as a function of the situation and information perceived by the individuals. This approach requires to consider also psychological aspects, as the Color Psychology Analysis [18] and Neuro-linguistic programming (NLP) analysis [19]. Neuro-linguistic programming (NLP) is an approach developed by Richard Bandler and John Grinder (in the 1970s) to communication, personal development, and psychotherapy. NLP teaches the ability to calibrate or “read” people (sensory awareness). It means the ability to interpret changes in muscle tone, skin color and shininess, lower lip size and breathing rate and location. From these and other indications it is possible to determine what effect these changes can have on other



Figure 1.2: Response Modes

people.

This information serves as feedback as to whether the other person is or not in a specific, desired state of study or interaction. An important and often overlooked point is to know the singling out moment when the other person is in the state that you desire to detect. The development of this system is a multidisciplinary topic (computer vision, artificial intelligence, psychology and marketing) and it can improve current applications, providing new DOOH systems thanks to the use of an approach of complete ‘reading’ of people mind. The whole analysis could be validated further via a correlation between a set of actual behaviors and responses, determined in experimental and properly authorized sites, to achieve scientific levels of results acceptance, eventually differentiating the readings by considering specific applications. In the second chapter we will focus on the Behaviour Analysis starting from the basics until the last updates in the field, considering also two important commercial benchmarks.

#### 1.1.4 Customer Tracking

Understanding customer behavior within the store is a very useful analysis for marketing purposes of a point of sale. There are numerous approaches to monitoring the consumer through, for example, the use of people counting devices, smart carts equipped with RFID technology, smartphone applications interconnected to the point-of-sale WIFI or BLE system, distributed camera systems within the point of sale. The people counting systems, in fact, make it possible to monitor the flows of people at the point of sale, considering the directions of movement. They are

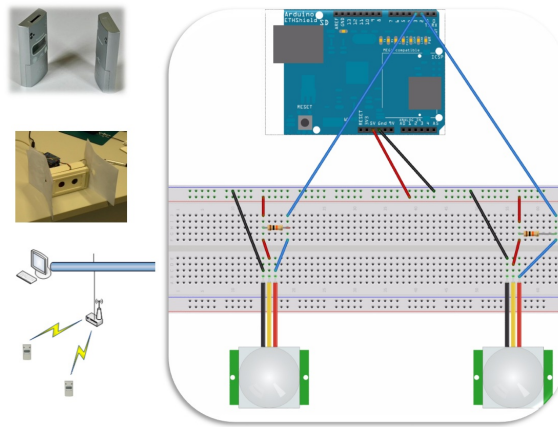


Figure 1.3: People counter

particularly useful for activating alerts, for example at the cash desk and at the departments served (e.g. butchers, delicatessen). However, this is a minimal information, useful for statistical purposes, but not very analytical, by virtue of the fact that the number of passages and the direction of travel is taken from the counting system in a certain position.

An evolution is represented by the tracking of the carts using RFID technology. This system provides for the installation in the point of sale of control points equipped with readers with RFID antenna (e.g. UHF.VHF) and on each cart a passive or active RFID. Passive RFID is cheap, it can be masked as a label, it does not require power as it is activated by the magnetic field induced by the reader. Active RFID, on the other hand, requires a greater investment as it has a battery and is larger than a label. Normally, for cost-effectiveness we rely on an important number of passive RFID carts. The passage of an RFID at a gate determines the detection of the passage. By interpolating the various passages of the carts between the various gates distributed, it is possible to reconstruct the customer's journey at the point of sale. It is thus possible to obtain the cold and hot areas of a point of sale based on the number of passes for the individual RFID gates.

Recently, with the advent of smartphones in everyday life, it has been possible to evaluate consumer tracking solutions through app interconnected with the WIFI intranet network of the point of sale. In this way it is possible to geolocalize the device and carry out promotional activities in the store. However, the user is required to use the application within the point of sale. A valid alternative to these tracking

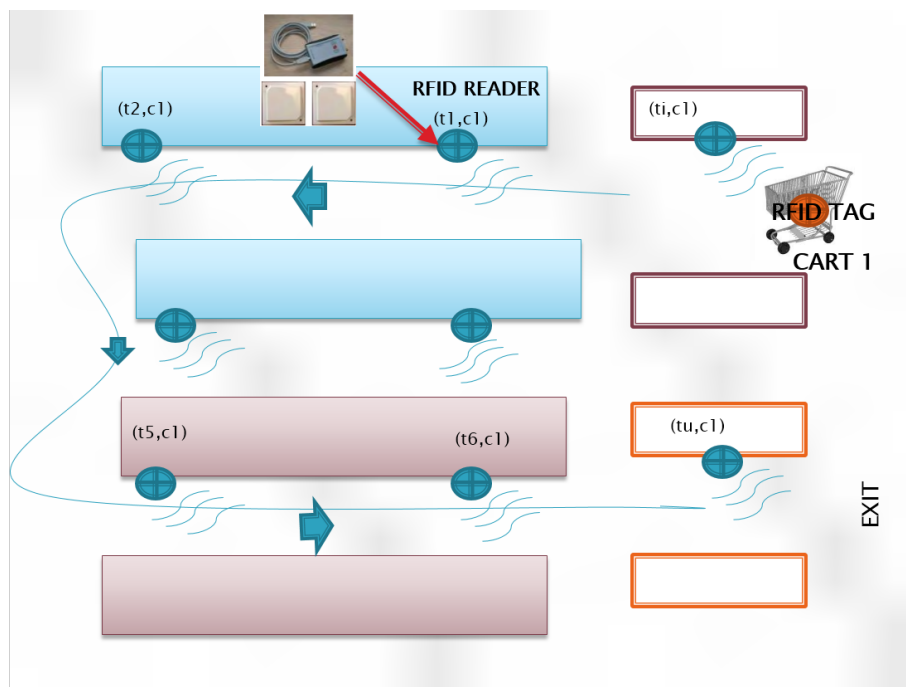


Figure 1.4: RFID Cart Tracking

systems is represented by the use of video surveillance systems, usually present at the point of sale, and which can be authorized by the territorial management of the work also for organizational and production improvement purposes. However, the number and positioning of the cameras is constrained on a regulatory basis in order to limit the invasiveness of these continuous-cycle active shooting systems. In addition, it is very difficult to place cameras to map the entire layout, adapting dynamically to the sales point of the store. Hence the opportunity to use self-centered, non-invasive vision systems on board the carts in order to monitor the shopping cart on time to reconstruct consumer behavior. This approach has been defined as Visual Market Basket Analysis, as it allows to carry out a timely monitoring of the shopping cart using the computer vision. This aspect will be dealt with in depth in the third chapter, considering a first traditional approach based on descriptors and machine learning and a second advanced approach based on deep learning. The described system has also been the subject of an invention patent.

An evolution combining also the traditional market basket analysis and considering by viewing also the picking and releasing activities by the users is presented as future development within the final chapter.



Figure 1.5: Egocentric vision on shopping carts

## 1.2 Papers and Patents

In this section we report the list of papers and patents produced during this PhD.

Conference Papers:

- V. Santarcangelo, G.M. Farinella, S. Battiato, ‘Gender Recognition: Methods, Dataset and Results’, IEEE International Conference on Multimedia Expo Workshops (ICMEW), 2015

*Results of this paper are presented in the sections 2.3, 2.4 of this thesis*

- V. Santarcangelo, G.M. Farinella, S. Battiato, ‘Egocentric Vision for Visual Market Basket Analysis’, ECCV Workshops, 518-531, 2016

*Results of this paper are presented in the section 3.3 of this thesis*

- M. I. Tariq, V. Santarcangelo, ‘Analysis of ISO 27001: 2013 Controls Effectiveness for Cloud Computing’, ICISSP, 201-208, 2016

*The approach about cloud computing cited in this paper is presented in the section 2.9 of this thesis*

- M. Giacalone, V. Santarcangelo, ‘Big Data Process Analysis: From Data Mining to Process Mining’, CLADAG, 2017

*Analysis dealt in this paper is presented in the section 4.4 of this thesis*

- V. Santarcangelo, ‘An innovative approach for the GDPR compliance in Big Data era’, SIS, 2018

*Approach shown in this paper is presented in the section 2.12 of this thesis*

- V. Santarcangelo, 'Tools for Risk Control of Data Management', ASA Pescara, 2018

*Analysis dealt in this paper is presented in the section 2.12 of this thesis*

Journal Papers:

- V. Santarcangelo, G.M. Farinella, A. Furnari, S. Battiato, 'Market Basket Analysis From Egocentric Videos', Pattern Recognition Letters, Volume 112, Pages 83-90, 2018

*Results of this paper are presented in the section 3.6 of this thesis*

Patents:

- A. Camporesi, V. Santarcangelo, 'Sistema per la misurazione della variazione dello stato comportamentale di un interlocutore', 17/11/2016, Patent: 0001425894

*This patent is described in the section 2.7 of this thesis*

- A. Camporesi, A. Meccariello, V. Santarcangelo, 'Metodo di identificazione avanzata basato sulla visione artificiale in un'area delimitata', 15/02/2017, Patent: 0001427186

*This patent is cited in the section 4.3 of this thesis*

- A. Camporesi, V. Santarcangelo, 'Metodo di localizzazione intelligente di oggetti basato sulla computer vision in un'area delimitata', 28/03/2017, Patent: 0001427883

- A. Camporesi, V. Santarcangelo, 'EMOTIONAL/BEHAVIOURAL/PSYCHOLOGICAL STATE ESTIMATION SYSTEM', Patent Pending: WO/2017/054871

*This patent is described in the section 2.7 of this thesis*

- V. Santarcangelo, G. M. Farinella, S. Battiato, A. Camporesi, 'Advanced Kinesthesia Analysis based on Artificial Vision and Audio Analysis for Process Control in a Delimited Area', International Patent, 18 January 2017, PCT/IT2017/000007

*This patent is described in the section 3.4 of this thesis*

Other publications:

- M. Giacalone, C. Cusatelli, V. Santarcangelo, 'Big Data Compliance for Innovative Clinical Models', *Big Data Research* 12, 35-40, 2018
- M. Giacalone, C. Cusatelli, A. Romano, A. Buondonno, V. Santarcangelo, 'Big Data and forensics: An innovative approach for a predictable jurisprudence', *Information Science*, 426, 160-170, 2018

## Chapter 2

# BEHAVIOR ANALYSIS IN RETAIL CONTEXT

### 2.1 DOOH and Retail

Digital out of home (DOOH) symbolizes the dynamic media distributed across placed-based networks through digital signage devices as addressable screens, kiosks, jukeboxes and jumbotrons, with the scope of engaging customers and extending the effectiveness of marketing experience [27]. DOOH application is one of the most important topic for retail environment, due to the change of traditional marketing into 1to1 approach, providing a custom shopping experience for the user. DOOH has received considerable interest from retailers and governmental institutions because of the benefits obtained in better managing and respond to the preferences of the users [7]. As a result, Computer Vision covers an important role for providing solutions for this scope [28]. Then, it is very important to exploit computer vision algorithms to automatically collect soft biometrics of people in front a smart screen. It is of great interest for industry [26]. These approaches are based on face detection module [1] and start with gender recognition algorithm. Infact, the input about gender is very important to customize the shopping experience because it is possible to define two defined clusters, one for male and one for female. An other important approach useful for DOOH is Age Estimation, infact, it is possible to define some clusters for different age targets. Combining gender, age and ethnicity input, emotion detection and skeletal tracking analysis, it is possible to improve DOOH applications. The two most important inputs of DOOH application are Gender and Age, then, in this chapter it is presented a full analysis about these two themes, with





Figure 2.1: Example of DOOH device (jumbotrons)

also an applied benchmark analysis and a patent about a behavioral analysis system.

## 2.2 Face Detection

To introduce the complexity about computer vision approaches starting from the identification of a face, it is important to evaluate how human beings perceive faces and their importance in social relationships. It is a fact that human beings can recognize the faces of family members from images at extreme resolution. The reason is not clear but it is a clear case of a cognitive process of information minimization. Border analysis is not enough to recognize, and face is processed somehow as a "whole" and not as composed by parts. In fact, from splitted parts of faces we can obtain new combined faces, that we recognize as other faces. It is very important to underline that eyebrows and impact of skin pigmentation are very important for the identification of faces and that faces can be recognized despite extreme distortions.

Faces seem to be encoded in memory in exaggerated caricature way: average face (averaged from a number of persons), some typical face, face created by taking big deviation from average. Such faces are recognized even better than typical ones. Newborn babies turn more attention to more face-like objects (upper row) than not face-like. From negative picture it is impossible to identify faces and face recognition is strongly compensated for the direction of illumination. Moreover, studies about response of neural cell of monkey in the face processing area of the



Figure 2.2: Example of extreme distortion of faces






	Faces	Cats	Schematic Faces	Objects
				
% MR Signal	1.6	1.6	0.9	0.6

Figure 2.3: Response of neural cells of human brain

brain show that response to something like face is much more stronger than for hand (millions and millions of cells are processing at the same time). Also in the measurement from human brain it is possible to see that signal from face-like picture is much stronger than from other objects. These considerations shown about faces indicate how sophisticated is information processing in biological systems. What is very amazing is getting correct results despite extreme distortions. For the most part, we do not know how this is done and we have difficulty in thinking how to develop algorithms which would have similar capabilities. The main problems for face detection are the high dimensionality, the heterogeneity of poses, lighting, occlusions, facial expressions (the face is not a rigid object), the presence of beard, glasses, mustache, and make-up.

The main approaches considered to implement the cognitive model of the face can be divided into:

- Knowledge-based (encode human knowledge of what constitutes a typical face by considering relationships between facial features - e.g., The central part of the face has constant luminosity. The difference in intensity between the central part and the upper part is significant; usually presents with two symmetrical eyes, a nose and a mouth)
- Feature-invariant (based on invariant structural features with several possible factors - e.g., edge, intensity, color and shape)
- Template matching (based on pre-calculated feature patterns with which to inspect an image by region)
- Appearance-based (templates or templates are created starting from a training set representing the variability of the subjects)

For face detection the well-known Viola and Jones approach is usually used (appearance based method).

## 2.3 Datasets

To properly approaching to Gender and Age recognition themes, in the various involved contexts, benchmark face datasets have been introduced. Large datasets are usually required to properly test and measure the performances. Face image datasets can be grouped in two main categories: constrained and unconstrained [25]. Constrained datasets are mainly composed of faces usually used for biometrical application purpose. A constrained dataset is characterized by images with controlled poses of the acquired subjects and pre-defined scene conditions. Examples of constrained datasets are AR [34], Lab2 [35], FEI Database [36], FERET [37], PAL [38], MORPH [39], better detailed in the following.

- **AR dataset**

It is characterized by over 3,000 color images of 116 people (63 men and 53 women) acquired at the Computer Vision Center of the U.A.B. under strictly controlled conditions with respect to the possible variabilities. The images are related to frontal view faces with different facial expressions, illumination conditions, and occlusions (e.g. sun glasses and scarf). This dataset is public

available at the following URI :

<http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html>.

- **Lab2 dataset**

This dataset is characterized by visible light images and near-infrared images of 50 subjects (12 females and 38 males). For each subject there are 20 visible light face images and the same number of near-infrared face images. Variability is related to facial expression and pose. Moreover images were acquired under four different source point illumination conditions: frontal illumination, left illumination, right illumination, both left and right illumination. The dataset can be found at the following URI:

<http://www.yongxu.org/databases.html>.

- **FEI face database**

It is a Brazilian face database that contains 2800 images (14 images for each of the 200 individuals, 100 males and 100 females) which have been acquired with homogeneous background. The acquired images are characterized of profile rotation up to about 180 degrees. The dataset can be obtained at <http://fei.edu.br/~cet/facedatabase.html>.

- **FERET database**

FERET contains 14051 grayscale images of human faces with different views (frontal, left and right profiles). It represents one of the most known and used dataset for face recognition purposes. Information on how to obtain the dataset are available at the following website : [http://www.itl.nist.gov/iad/humanid/feret/feret\\_master.html](http://www.itl.nist.gov/iad/humanid/feret/feret_master.html).

- **PAL dataset**

The PAL DB is characterized by 575 face images of adults ranging from 19 to 93 years of age (225 males and 350 female). The official website of this dataset is <https://pal.utdallas.edu/facedb/>.

Despite constrained dataset are frequently used in literature, in real application domain (such as DOOH) the images to be analyzed are taken in unconstrained settings. A lot of state of art approaches reach really good performances on

constrained datasets but could not have the same accuracy on image acquired on real life. So, in the recent years unconstrained datasets have been introduced. Unconstrained datasets are built considering images acquired by real life, with different poses and scene conditions. These datasets are principally built collecting images from public repository (e.g., Flickr) with the use of web image crawlers. The unconstrained face datasets currently used in literature are LFW [39] , Gallagher [48], Genki-4K [49] , Image of Groups [50], Kin-Face [51]. Gallagher and Image of Groups are unconstrained image of groups datasets, and are the most difficult one where testing these approaches. The development of dataset more and more useful for these application is a target to consider in the today research.

- **Labeled Faces in the Wild (LFW)**

It can be considered the most important unconstrained face dataset up today. It contains more than 13000 images of faces collected from the web (10256 male and 2977 female images). Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the dataset. The official link to obtain the LFW dataset is <http://vis-www.cs.umass.edu/lfw/>.

- **Gallagher dataset**

It is a collection of 931 real life digital images of people. The dataset can be downloaded from the author's website: <http://chenlab.ece.cornell.edu/people/Andy/GallagherDataset.html>.

- **GENKI-4K dataset**

It contains over 3,000 color face images labeled as either “smiling” or “non-smiling” by human coders (1539 females and 1,506 males extracted by Danisman et al [40].). The images contains faces spanning a wide range of illumination conditions, geographical locations, personal identity, and ethnicity. GENKI-4K can be downloaded at the link [http://mplab.ucsd.edu/wordpress/?page\\_id=398](http://mplab.ucsd.edu/wordpress/?page_id=398) .

- **Image of Groups**

This is a dataset of 5,080 images containing groups of people. It contains

28,231 faces labeled with age and gender. This is a useful dataset for the studies of groups of people in unconstrained settings. It can be downloaded at <http://chenlab.ece.cornell.edu/people/Andy/ImagesOfGroups.html>.

- **UB KinFace**

It comprises 600 images related to 400 people (most of them are real-world collections of public figures as celebrities and politicians from Internet). It can be organized into 200 groups (child, young parent and old parent images). The dataset is available online at : <http://www1.ece.neu.edu/~yunfu/research/Kinface/Kinface.htm>.

- **CASIA WebFace dataset**

The dataset [91] contains photos of actors and actresses born between 1940 and 2014 from the IMDb website. Images of the CASIA WebFace dataset include random variations of poses, illuminations, facial expressions and image resolutions. In total, there are 494,414 face images of 10,575 subjects. The dataset is available online at : <http://classif.ai/dataset/casia-webface/>.

Following table summarizes the main characteristics of the datasets and, whenever it is available, the number of male/female subjects.

Table 2.1: Face Datasets for Gender Recognition

NAME	TYPE	Images	Male Faces	Female Faces
AR	Constrained	3016	1638	1378
Lab2	Constrained	2000	1520	480
FEI	Constrained	2800	1400	1400
FERET	Constrained	14051	nd	nd
PAL	Constrained	575	225	350
Gallagher	Unconstrained	931	nd	nd
LFW	Unconstrained	13233	10256	2977
GENKI-4K	Unconstrained	3045	1506	1539
I.Groups	Unconstrained	5080	10303	9532
CASIA	Unconstrained	494414	nd	nd
UB KinFace	Unconstrained	600	440	160

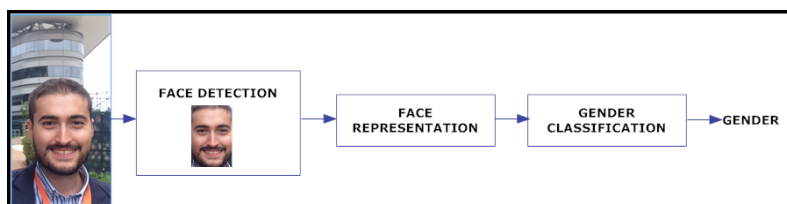


Figure 2.4: GENERAL PIPELINE OF A GENDER RECOGNIZER

## 2.4 Gender Recognition

Gender recognition approaches share a pipeline composed by three main modules (see Fig.2): face detection, face representation and gender classification. In this section we detail the components of a gender recognition method by considering the current state-of-the-art.

### 2.4.1 Face Detection

As first stage of a gender recognition engine, the faces of the people present into the acquired images have to be detected. To this purpose the well-known Viola and Jones object detection framework is usually used [33]. It is able of processing images in real time achieving high detection rates. The solution exploits the Integral Image representation so that Haar-like features can be computed at any scale or location in constant time. A learning algorithm based on AdaBoost [46] is used to select the most discriminative features bases for classification purposes. Combining different classifiers in a cascade the background regions of the image are discarded while faces are detected.

### 2.4.2 Face Representation

When faces are detected a discriminative representation have to be computed. Usually a pre-processing step to remote geometric variabilities (e.g., by aligning faces with respect to the eyes positions) as well as photometric variabilities (e.g., light conditions) is performed [1]. The obtained images are hence processed to extract features to be sual into the gender classification module. Different features methods have been used for gender recognition. Almost all approaches for gender recognition are based on LBP, HOG, SIFT features. These features can be combined also with

Color Histograms (CH) and Gabor features to improve the performances. It is useful to synthesize these approaches to better understand their application on gender recognition.

- **LBP descriptor**

It considers spatial comparison of local neighborhood of a pixel and creates labels which are then aggregated in histograms. LBP descriptors (and the related variants) are robust to illumination and to the rotation variations. LBP are useful to capture textures (e.g., beard).

- **SIFT descriptor**

The SIFT descriptor is invariant to scale, translation, rotation, partially invariant to illumination changes and robust to local geometric distortion.

- **HOG method**

It considers the image divided into a block structure cell-based. The block is characterized by the histogram of oriented gradients as a function of the edges. Histograms are subjected to normalization of contrast.

LBP, SIFT and HOG approaches have been also combined with the CH and Gabor features for gender recognition [20] [23]. The histogram of the colors (CH) extracts the characteristics on the distribution of the "colors" of the image, instead, Gabor filters are used to encode texture. Moreover, recent approaches uses CNN features (Convolutional Neural Network) reaching good results [44]. Once the features are extracted and the image is represented, a classifier is applied.

### 2.4.3 Gender Classification

The most used methods for gender classification are SVM (with RBF Kernel) and Adaboost combined with Linear SVM. A Support vector machine (SVM) [41] is a powerful classifier for two classes based problems: gender recognition problem (male vs female) is a good application scenario. The Adaptive boosting approach (Adaboost) [42] is an ensemble learning based method able to build a strong classifier from a combination of weak classifiers. In CNN (Convolutional Neural Network)



approaches it is possible to find also Softmax classifier. Softmax is a generalization of a Logistic Regression classifier to multiple classes, that yields the actual probability scores for each class label.

#### 2.4.4 State of the Art Results

In this paragraph we review the results obtained by the current state of the art approaches. For each approach we briefly describe the main ingredients used to build the gender recognizer, which dataset have been used in the experimental phase and well and the obtained results. The comparison among the different approaches is obtained by considering the results claimed by the authors in their papers. In this way approaches which used the same dataset in the experiments are straightforward compared. The comparison is reported in Table 2.

- **Danisman et al. [40]**

This approach shows as the use of a pre-processing step can improve the performances of gender recognition on unconstrained datasets. The method proposed by the authors is composed by face detector, a face alignment step to remove geometric variability, and a histogram equalization in which the face images are normalized with respect to a specific probability density function obtained considering the average face of the training dataset to remove illumination variability. Classification is obtained exploiting the SVM classifier with RBF kernel. This method has been tested on unconstrained datasets. The accuracy obtained on the LFW, Genki-4k and Groups datasets are respectively 91.87%, 91.07%, 88.16%.

- **Ersi et al. [20]**

The gender classification approach is based on the combination of LBP, SIFT and CH descriptors. Classification is obtained by exploiting a SVM with RBF kernel. The accuracy obtained on the Gallagher database (unconstrained images from the web) is 91.6%.

- **Liu et al. in [21]**

The authors introduce a new feature called Self-Similarity of Gradients (GSS)

which captures pairwise statistics of localized histogram of gradient distributions. For classification purpose both Adaboost and SVM are compared independently or in cascade (boosting for feature selection and SVM for classification). The performances reached on LFW face dataset by considering only the GSS descriptor are 88.96%, whereas the combination of HOG31, LBP and GSS features achieved an accuracy of 95.76%.

- **Borgi et al. in [22]**

This work proposes a new approach for gender classification called multi-regularized learning (MRL). It considers as first step a dimensional reduction of the faces feature space. Then the proposed multi-regularization feature learning approach is applied for classification purpose. The method obtained 92.83% of accuracy on the AR dataset, whereas an accuracy of 94% is reached on the FEI dataset.

Table 2.2: Performances

	[40]	[20]	[21]	[22]	[23]	[43]	[44]	[45]	[93]	[44]	[94]
Groups	88.16									91.34	90.14
Genki-4K	91.07										
LFW	91.87		95.80		98.00		96.86		97.31	98.90	98.00
Gallagher		91.60				88.60					
AR				92.83							
FEI				94.00							
FERET					98.78						
KinFace					96.50						
Feret+Morph								89.70			

- **Ren et al. in [23]**

This work considers a combination of the SIFT, HOG and Gabor filters as final descriptor for gender recognition. The classification is obtained through RealAdaboost with the use of a penalty term that considers the complexity of the feature combination. The combinations of the feature spaces with the penalty term reduces the computational complexity. The approach has been tested on both constrained and unconstrained datasets obtaining the following results: 98.78% on FERET, 96.50% on KINFACE, 98.01% on LFW.

- **Eidinger et al. [43]**

The authors presented a pipeline based on with four steps: detection, alignment, representation and classification. The detection is obtained by the use of Viola and Jones face detector, the alignment method is done considering the position of 68 specific facial features, the final classification is obtained by the use of a Dropout-SVM on LBP and FPLBP representation. The results of the method on Gallagher Dataset is of 88.6%.

- **Jia et al. [44]**

This contribution address the important challeng of training the gender classifier by considering a big dataset represented in a high dimensional feature space (four million images and 60 thousand features). The proposed approach use an ensemble of linear classifiers, and achieves an accuracy of 96.86% on the most challenging public database, Labelled Faces in the Wild (LFW).

- **Carcagni et al. [45]**

The paper presents a comparison among LBP, HOG and SWLD (Spatial Weber Local Descriptor) descriptors for gender recognition on constrained dataset. Specifically, the authors have used a fusion of FERET and MORPH datasets for testing purposes. The best accuracy is obtained by using HOG descriptors coupled with SVM and exploiting RBF Kernel (89,70%).

- **Jia et al. [92]**

This work presents an interesting comparison between CNNs with approaches with other types of vision features from different facial regions, introducing a novel approach. Specifically, the authors have used CNN features with Softmax classifier on the Labeled Faces in the Wild (LFW) dataset achieving an accuracy of 98.90%, and on the Images of Groups (GROUPS) dataset, achieving an accuracy of 91.34% for cross-database gender classification.

- **Antipova et al. [93]**

In this paper authors presents an approach of gender recognition based on CNN features with Softmax classifier. The accuracy obtained on the Labeled Faces in the Wild (LFW) dataset is 97,31%.

- **Santana et al. [94]**

This approach considers a combination of LBP, HOG and CNN features for gender recognition. They used SVM as classifier and achieved the highest accuracy on the LFW and GROUPS in a cross-database setting, 98% and 90.14% respectively.

To complete the analysis, it is important to consider also the impact of the image resolution regarding the performances obtained by a gender classifier. Andreu et al. [47] performed tests on different dataset (FERET, PAL, AR) considering them at different resolution. The study shows that a size between 22x18 and 90x72 pixels is recommended for the problem of gender recognition. The authors pointed out that a size of 45x36 pixels provides enough information to infer the gender recognition from images.

The results reported above show that the best performances are obtained on constrained datasets (98.78%) [23]. Considering unconstrained datasets, the feature descriptors better performing are CNN [92] (98.80%) and the combination of HOG, SIFT and Gabor descriptors [23] (98.01%). It is important to notice that some datasets (e.g., the one with images of groups) are more complex than others in gender recognition, due to the fact of the presence of groups of people, children, different ethnic groups and a large age gap.

## 2.5 Age Estimation

Although humans are able to perform age estimation from human face images with a certain accuracy, it is a pretty complex task to be performed automatically. Initially, the first approaches of automatic age estimation were based on anthropometric models (e.g. craniofacial growth, skin aging, facial lines and wrinkle estimator). An evolution was represented by Active Appearance Model (AAM) considering both shape and texture [32]. The evolution of these approach has been similar to Gender recognition, then Age estimation is based on a pipeline composed by three main modules: face detection, face representation, gender classification. Almost all approaches for age estimation are based on LBP, HOG, SIFT and CLBP (Completed Local Binary Pattern) features with the combination of Gabor Filter. Methods for the classification are mainly SVM classifier with RBF Kernel. In Age Estimation

approaches, datasets are divided mainly in 5 Age Groups : 16-24, 25-39, 40-59 and 60+. However, the presence in the dataset of a little number of pictures for the class 60+ make system not very robust for detection of seniors. The results of the state of the art show that the best performances on constrained dataset are obtained considering CLBP approach (73%) and as expected, considering the confusion matrix for the age classes, most of the confusion occurs between adjacent classes [1]. Combining LBP+SIFT+HOG [20] it is possible to reach 63% of accuracy in unconstrained image of groups datasets, that represent an example of application in the real context. CNN approaches [95] reach lower results (59.90%).

## 2.6 Research Proposal

Research proposal for DOOH is about the analysis of digital signage applications and the study of methods about gender, age, ethnicity, emotion, clothing attributes and skeletal/gesture recognition [29] to develop a system for the behavioral state analysis of a buyer. The behavior analysis of a buyer can be defined analyzing the 4 possible “response modes” defined by Miller Heiman in “The New Strategic Selling” [28] approach that are: Growth, Trouble, Even Keel, Overconfident. These Concepts, specifically developed for B2B Buyer-Vendor transaction cases suggest that: Buyers in Growth mode perceive that by buying a product or service, they will produce growth for their company (opened to vendor offerings). In Trouble mode, buyers perceive that they have to fill rapidly a business gap (e.g. it is losing customers, losing money, decreasing productivity) and understand that something must change (opened to vendor offerings). Buyers in Even Keel mode do not perceive a large enough gap between their position and their goal, then there is no urgency to change anything (closed to vendor offerings). Buyers in Overconfident mode believe that they are doing so well that the suggestion that your offering might improve their situation is practically an insult (closed to vendor offerings). In fact they have a weak grasp of the reality and their mind is then closed to suggestions before to the natural fall to the trouble mode. These concepts have been developed, applied and conceptually adapted, also to the B2C case and specifically utilized to tailor properly the Communication to each specific Customer Response Mode detected or targeted, considering the related variability as a function of the situation and

information perceived by the individuals. The core of this research in DOOH is to project and develop a system for the “response mode” recognition, through a combination of inputs as that of Gender, Age, Ethnicity, Emotion and Skeletal’s systems combined to feedback from device interaction (e.g. text input, page choice, opinion mining, social network interaction, pir sensor feedbacks) thanks also to the use of semantic networks. This research field requires to consider also psychological aspects, as the Color Psychology Analysis [30] and Neuro-linguistic programming (NLP) analysis [31]. Neuro-linguistic programming (NLP) is an approach developed by Richard Bandler and John Grinder (in the 1970s) to communication, personal development, and psychotherapy. NLP teaches the ability to calibrate or ‘read’ people (sensory awareness). It means the ability to interpret changes in muscle tone, skin colour and shininess, lower lip size and breathing rate and location. From these and other indications it is possible to determine what effect these changes can have on other people. This information serves as feedback as to whether the other person is or not in a specific, desired state of study or interaction. An important and often overlooked point is to know the singling out moment when the other person is in the state that you desire to detect. The development of this system is a multidisciplinary topic (computer vision, artificial intelligence, psychology and marketing) and it can improve current applications, providing new DOOH systems thanks to the use of an approach of complete ‘reading’ of people mind. The whole analysis could be validated further via a correlation between a set of actual behaviors and responses, determined in experimental and properly authorized sites, to achieve scientific levels of results acceptance, eventually differentiating the readings by considering specific applications.

## 2.7 Behavioral Analysis System

To develop this approach, a patent about a system for measuring the variation of the behavioral state of an interlocutor has been written and granted by Centro Studi. Now it is reported the description of the invention. The recognition of emotions through computer vision applied to the face of an interlocutor is one of the main areas of current research, however, the result of this analysis risks being in itself not

very reliable as it is mainly linked to the characteristics of a person's face in a determined moment, without considering environmental boundary factors and the body language in toto (NLP). According to what is known, both in the culture of NLP (neuro-linguistic programming) and sales strategies there are response modalities ("response modes") or behavioral states. In the literature, the "response modes" presented by Robert B. Miller in the text "The NEW Strategic Selling" [52] are:

- Growth (Development): in which the interlocutor "is always ready to say YES" to get something better that introduces a "discrepancy" to fill with a consequent high probability of sales / interaction success;
- Trouble (Crisis): in which there is a "defeat" to be solved immediately, with a consequent high probability of sales / interaction success;
- Even Keel (Satisfied): in which there is no "discrepancy" to fill, resulting in a low probability of sales / interaction success;
- Overconfident (Too Optimistic): in which the user is convinced that he has more than what he actually has, with the consequent probability of a successful sale / interaction.

Knowing how to determine the "behavioral state" of an interlocutor in a given moment, and being able to measure the variation of the behavioral state over time, is a fundamental tool to better direct communication in any field and to predict the relative probability of success of the interaction. In particular, the patent finds application in any scenario in which there is interaction with the user and the need to obtain feedback on his behavioral state. In order to implement this system it is necessary to consider the potential of neuro-linguistic programming, which allows to model the behavior of people in order to obtain useful input in communication. In this context, the technical task underlying the present invention is to propose a system for determining the behavioral state and measuring its variation over time. Moreover, the present invention extends Miller's "response modes" by introducing two possible combinations for each "mode", RATIONAL (with a precise perception by the interlocutor), NOT RATIONAL (with the absence of a precise perception by the 'party). Consequently, the 4 "modes" CRISIS (C), DEVELOPMENT (S), TOO OPTIMIST (TO), APPROVED (A) specialize through the two RATIONAL (1) and NON RATIONAL (0) states in 8 "modes": RATIONAL CRISIS (C1), NON-RATIONAL CRISIS (C0), RATIONAL DEVELOPMENT (S1),

NON-RATIONAL DEVELOPMENT (S0), TOO VERY RATIONAL OPTIMISM (TO1), TOO MANY NON-RATIONAL OPTIMIST (TO0), RATIONAL APPARATUS (A1), NON-RATIONAL APPROVED (A0).

In particular, it is an object of the present invention to provide an automatic system characterized by a camera 11, a camera with sensors for the "skeletal / gesture recognition" 12, color sensor 13, microphone 14, proximity sensor (PIR or ultrasonic) 15, pressure sensor for floor 16, possible sensor magnetometer SQUID 17, input device (keyboard, touchscreen monitor, joystick, mouse) 18, output device (display, monitor) 19, and from information from social networks, loyalty loops, speaker barriers, user flow monitoring systems (e.g. systems with RFID technology) 21, indoor environmental information (brightness, humidity, temperature) 7 and outdoor (weather) 22. The specified technical task and the specified purpose are substantially achieved by a system for measuring the variation of the behavioral state of an interlocutor comprising the technical characteristics exhibited in one or more of the appended claims. Further characteristics and advantages of the present invention will become clearer from the indicative, and therefore not limiting, description of a preferred but not exclusive embodiment of a system for measuring the variation of the behavioral state of an interlocutor, as illustrated in the accompanying figure 2.5, which is a schematic representation of a system in accordance with the present invention. The system for measuring the variation of the behavioral state of an interlocutor in accordance with the present invention is applicable in any case within a closed or closed delimited area. As an example, this delimited area can be that of a shopping center, a local council, a service station, any DOOH scenario (digital out of home). In fact, the system lends itself, for example, to be integrated into interactive multimedia totems or gaming devices (such as slot machines). The system provides, initially, to associate with the genus sizes 23, age 24, facial expression 25, position of the ocular pupils 26 (extracted from the face test 1), shape of body 2, detected colors 3, tone of heading 27, type of breath 28 (extracted from the acquired audio signal 4), distance 5, position of the feet 29, speed of interaction with the input devices 6 detected in the area delimited "A" a corresponding value. In a database 10, the measured values are saved. This information is accompanied by any choices made using input devices 30, information extracted from social networks (using opinion mining algorithms), loyalty circuits, user flow monitoring systems or



cash barriers 21, from weather information 22 and indoor environmental conditions 7 (including any information on the level of ELF waves obtained magnetometers). The group of sensors 8 is mounted on board an embedded device or a personal computer 9 equipped with a tcp / ip network interface. The values obtained are then interpreted according to the reference values present in the database 10, which represent the basis of knowledge for interpreting the recorded data. This knowledge base contains the scheme of interpretation of NLP, logic related to the psychology of colors, logics for the joint interpretation of data (semantic tree structures) and mathematical models for self-learning and not. For processing unit 20 we can mean a common computer or a server containing the database 10. The communication of the data collected to the unit 9 occurs via the tcp / ip network protocol. Advantageously, the device 9 can also be a remote server. Alternatively, the embedded device 9 can be queried via a tcp / ip (wifi / ethernet) network interface to obtain a real-time feedback via a web interface html based on a standard logic of real-time data interpretation. This device 9 can therefore be queried by devices such as PCs, notebooks, PDAs, smartphones, tablets with web browsers and tcp / ip connectivity. The device 9 can in turn be integrated into the central unit of an interactive totem. In addition to monitoring the measured values in real-time, it is therefore possible to check their time evolution. The system returns with respect to a time frame established by the user the variation of the behavioral state, measured in the various time instants by means of the algorithm referred to below. Considered "N1" the first number describing the synthesis value obtained from the information (gender, age, facial expression, position of the pupils, colors) obtained by means of the acquisition with camera and color sensor, "N2" the synthesis value obtained from the information of the sensor camera to the skeleton recognition (template), "N3" the synthesis value obtained from the voice / breath information, "N4" the synthesis value obtained from the interaction detected by input devices, "N5" the value of synthesis obtained from the distance of the interlocutor, "N6" the synthesis value obtained from the additional information (social networks, loyalty circuits), "N7" the synthesis value obtained from the historical information, "N8" the synthesis value obtained from the information relating to the indoor and outdoor environment, "N9" the synthesis value obtained from the additional additional information

obtained from the sensors (e.g. electric fields) at a determined moment "t0" of detection the numerical value that describes the behavioral state "I (t0)" is a number obtained after a normalization of the data, which include values of importance of the individual components, ie:  $V1 = 1 - (N1 / N1max)$ ;  $V2 = 1 - (N2 / N2max)$ ;  $V3 = 1 - (N3 / N3max)$  ...  $Vn = 1 - (Nn / Nnmax)$  with  $N1max$ ,  $N2max$ ,  $N3max$ , ...  $Nnmax$  established on the basis of theoretical knowledge. Defined  $\gamma_1, \gamma_2, \gamma_3 \dots \gamma_n$  (weights of importance of the single components,  $0 \leq \gamma_i \leq 1$ ) the value that identifies a behavioral state "IC" for  $n = 3$  is established by the formula:

$$IC = \frac{((\gamma_1) \times V1 + (\gamma_2) \times V2 + (\gamma_3) \times V3)}{(\gamma_1)^2 + (\gamma_2)^2 + (\gamma_3)^2}$$

This value in relation to a numerical reference threshold value (depending on the knowledge base and mathematical models) determines its behavioral state. In addition to giving feedback on the behavioral state, the system 2.6 is able to suggest a type of action to be taken with the interlocutor and the relative probability of success.

## 2.8 Benchmark Analysis

In order to understand how the major world players are dealing with this issue, the Google Cloud library and an embedded device provided by Centro Studi have been taken into consideration. Although the costs are accessible and the high-performance system, the risk of using Google Cloud services is linked to a possible violation of the GDPR, as Google Cloud acts as the controller, transferring the entire management responsibility to the customer (Data Controller). In tests carried out in retail contexts, also thanks to the help of semantics, the Google library does not perform correctly on individual products. In context analysis, on the other hand, it is more reliable. Excellent performances have been found thanks to the use of the OCR module. Regarding the recognition of the face and emotions, the system is particularly efficient.

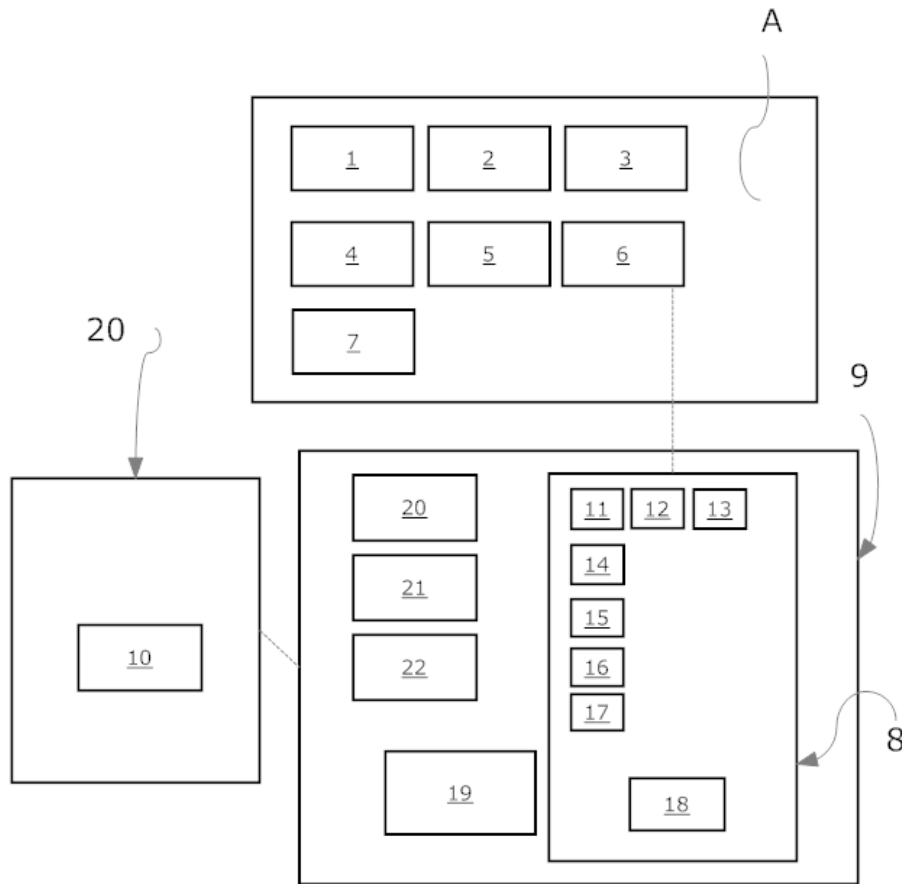


Figure 2.5: Blocks scheme of patented system (A environment, processing unit 20, server 9)

## 2.9 Google Cloud Platform

To introduce the Google Cloud approach, it is important to introduce the comparison about Cloud approaches [97]. In fact, we have to consider the differences between Paas, Saas and Iaas cloud services. Software as a Service, also known as cloud application services, represent the most commonly utilized option for businesses in the cloud market. SaaS utilizes the internet to deliver applications to its users, which are managed by a third-party vendor. A majority of SaaS applications are run directly through the web browser, and do not require any downloads or

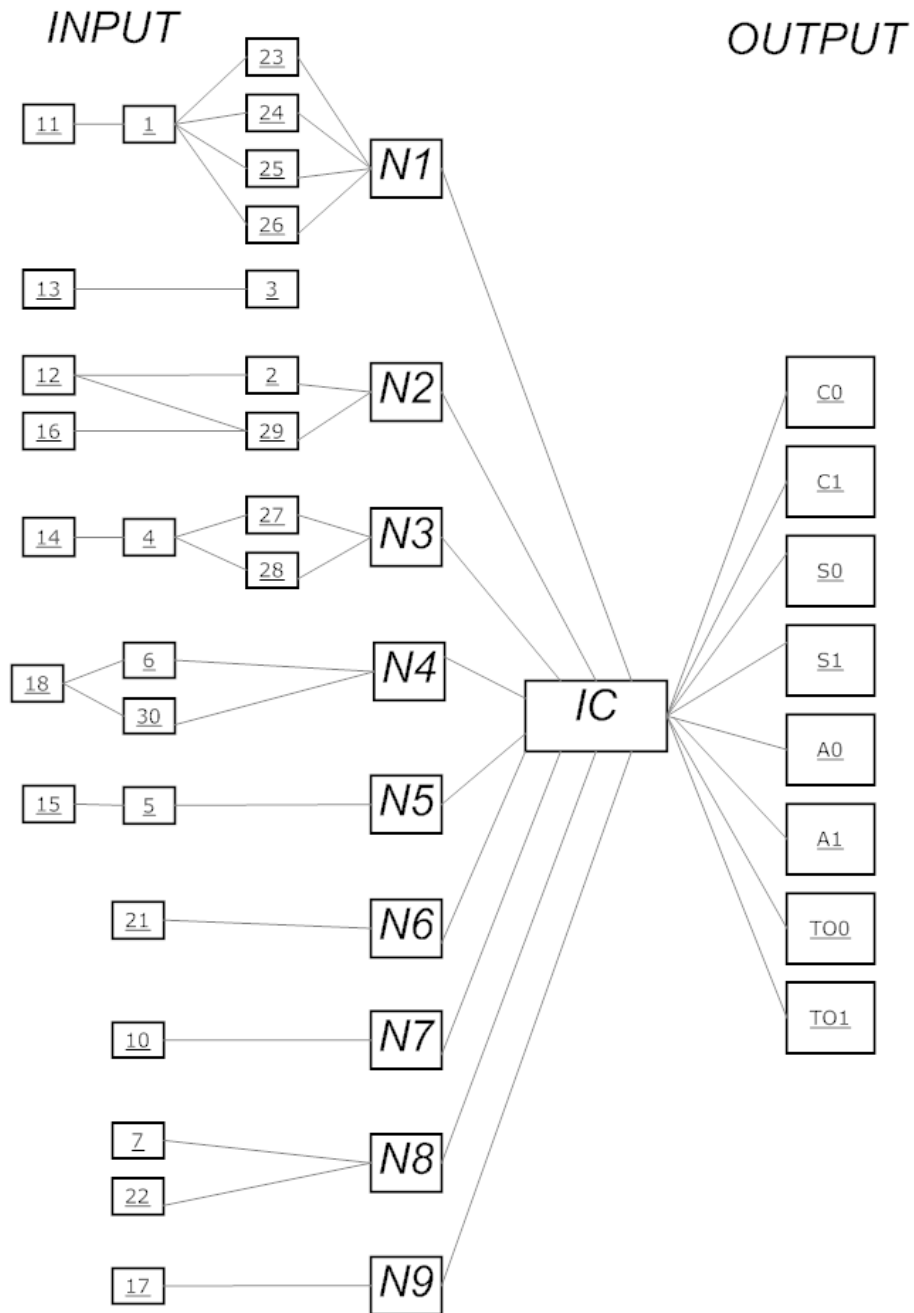


Figure 2.6: Detailed scheme of patented system (inputs read, IC behavioral state, system outputs)

installations on the client side. Platform as a Services (PaaS) provide cloud components to certain software while being used mainly for applications. PaaS provides

a framework for developers that they can build upon and use to create customized applications. All servers, storage, and networking can be managed by the enterprise or a third-party provider while the developers can maintain management of the applications. Infrastructure as a Services (IaaS) are made of highly scalable and automated compute resources. IaaS is fully self-service for accessing and monitoring things like compute, networking, storage, and other services, and it allows businesses to purchase resources on-demand and as-needed instead of having to buy hardware outright. Google, along with Microsoft, was one of the first big companies to offer PaaS services for the consumer market. An example of PaaS service is Google Cloud Platform.

### 2.9.1 Google Cloud Platform: Image Analysis

Google Cloud Vision API allows developers to understand the content of an image by encapsulating powerful machine learning models in an easy to use REST API. It is possible to classify images into thousands of categories (e.g., "sailboat", "lion", "Eiffel Tower") [96], detect individual objects and faces within the images, and find and read printed words within the images by OCR. It is possible to generate metadata for an image catalog, moderate offensive content, or enable new marketing scenarios through image sentiment analysis. With this platform it is very easy to detect broad sets of objects in images, from flowers, animals, or transportation to thousands of other object categories commonly found within images. Vision API enables also to detect different types of inappropriate content such as violent content. Vision API uses the power of Google Image Search to match results with celebrities, logos, or news events. It considers Visually Similar Search to find similar images on the web. Optical Character Recognition (OCR) enables to detect text within images, along with automatic language identification. Then, with this platform it is possible to detect sets of categories within an image, ranging from modes of transportation to animals. Label detection is definitely the most interesting annotation: this feature adds semantics to any image or video stream by providing a set of relevant labels (i.e. keywords) for each uploaded image. Labels are selected among thousands of object categories and mapped to the official Google Knowledge Graph. This allows image classification and enhanced semantic analysis, understanding, and reasoning. Technically, the actual detection is performed on the image as a whole, although

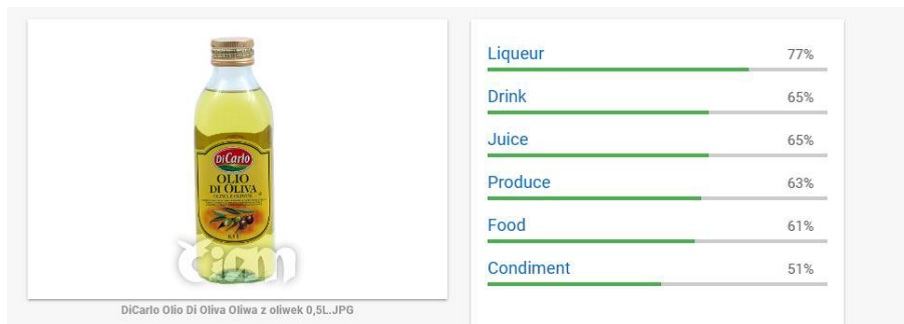


Figure 2.7: Label Detection of an olive oil bottle

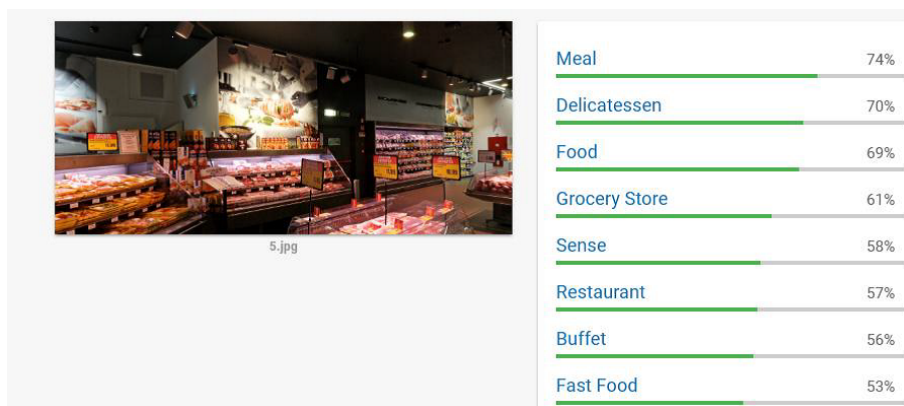


Figure 2.8: Label Detection of a meat department

an object extraction phase may be executed in advance on the client in order to extract a set of labels for each single object. In this case, each object should be uploaded as an independent image. However, this may lead to lower-quality results if the resolution is not high enough, or if the object context is more relevant than the object itself — for the application’s purpose. In the following images, we have provided some examples about label detection. Testing this platform on an olive oil bottle, it is possible to notice some misclassification: liqueur/drink/juice are provided as best results, confirming the very difficult topic of the product retail classification. Good results are provided in the analysis of retail scenes (meat and vegetable department).

We have also tested an interesting feature of this platform concerning the detection of popular product logos within an image. In fact, it is possible to see in the following image how the logo has been detected.

This can be done also by the use of OCR, which always gives good results.

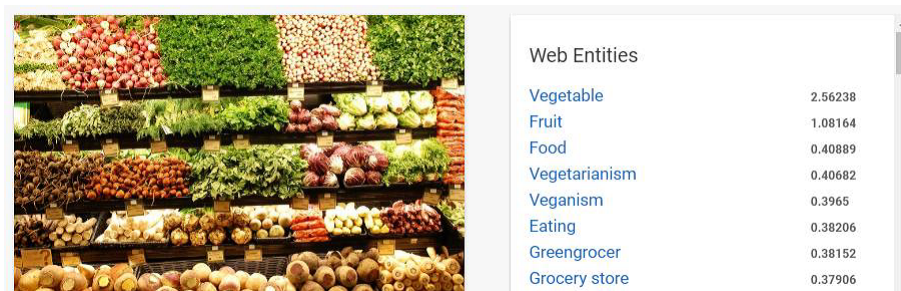


Figure 2.9: Label Detection of vegetable department



Figure 2.10: Logo detection in a supermarket

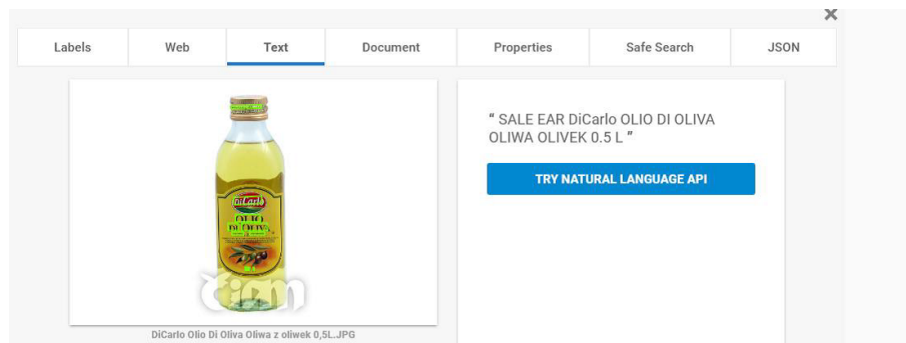


Figure 2.11: OCR of an Olive Oil Bottle

The Vision API, in fact, runs OCR - similar to the model which is used in Google Translate - to extract the following text from the image. In addition to a bounding box for the entire text, the platform also get a bounding box for the position of each word in the image, to analyze it further or translate it.

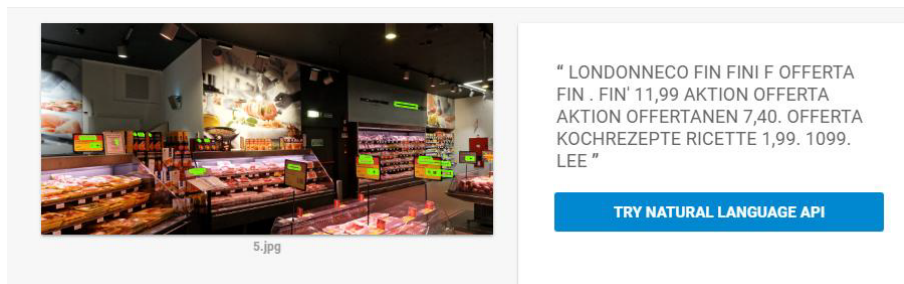


Figure 2.12: OCR of a meat department

## 2.9.2 Google Cloud Face and Emotion Detection

The platform is very useful also to detect multiple faces within an image, along with the associated key facial attributes like emotional state or wearing headwear. Face detection aims at localizing human faces inside an image. It's a well-known problem that can be categorized as a special case of a general object-class detection problem. It is important to define that:

- It is NOT the same as Face Recognition, although the detection/localization task can be thought of as one of the first steps in the process of recognizing someone's face. This typically involves many more techniques, such as facial landmarks extraction, 3D analysis, skin texture analysis, and others;
- It usually targets human faces only.

If you ask the Google Vision API to annotate your images with the FACEDETECTION feature, you will obtain the following:

- The **face position** (i.e. bounding boxes);
- The **landmarks positions** (i.e. eyes, eyebrows, pupils, nose, mouth, lips, ears, chin, etc.), which include more than 30 points;
- The main **face orientation** (i.e. roll, pan, and tilt angles);
- **Emotional likelihoods** (i.e. joy, sorrow, anger, surprise, etc), plus some additional information (under exposition likelihood, blur likelihood, headwear likelihood, etc.).



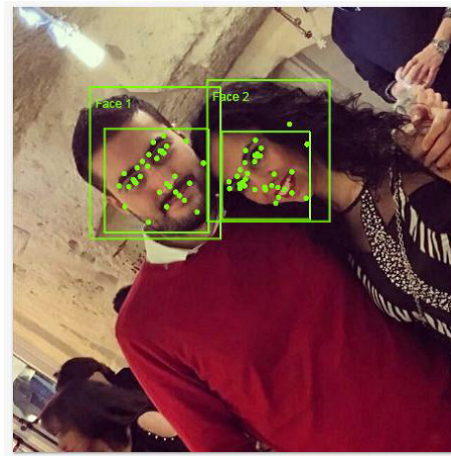


Figure 2.13: Face Detection Test

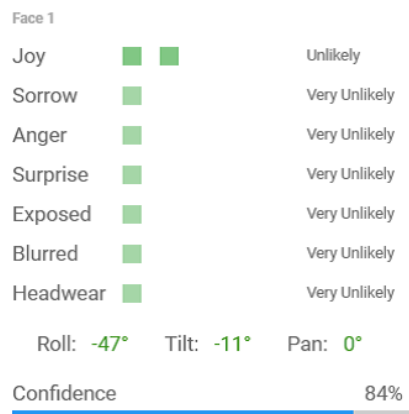


Figure 2.14: Face 1 results

We have carried on our study giving some examples about face detection in an image concerning a couple. The algorithm detects the two faces and for each face the relative emotions, with a score from 1 to 5 and a confidence factor. Comparing OpenCV with Google Cloud Vision [98] on the Cohn-Kanade AU-Coded Facial Expression Database with Fisherface technique, the OpenCV implementation got the best performance.

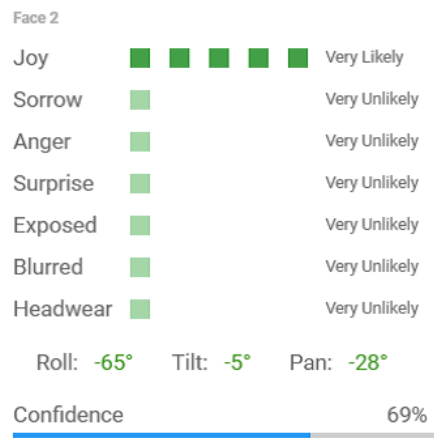


Figure 2.15: Face 2 results

Furthermore, the platform detects also general attributes of the image, such as dominant colors and appropriate crop hints.



Figure 2.16: Attribute Face Detection Test

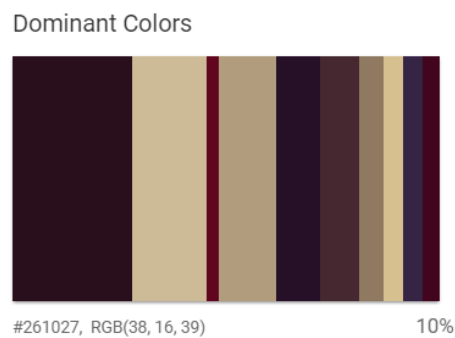


Figure 2.17: Dominant Colors in Face Detection Test

It is very important to underline that behind the development of Google Cloud Vision there is Fei-Fei Li, director of the Stanford Artificial Intelligence Lab (SAIL) and the Stanford Vision Lab. Among her best-known work is the ImageNet project, which has revolutionized the field of large-scale visual recognition. Fei-Fei is the recipient of the 2014 IBM Faculty Fellow Award and the 2012 Yahoo Labs FREP Award.

## 2.10 Embedded Device Tests

A further benchmark analysis was provided by the use of an embedded device provided by Cento Studi S.r.l. This device is equipped with processing unit (based on deep learning model) and acquisition unit and can be interfaced via USB port. Unfortunately, for reasons of intellectual property Cento Studi S.r.l. has not provided any further technical specifications.

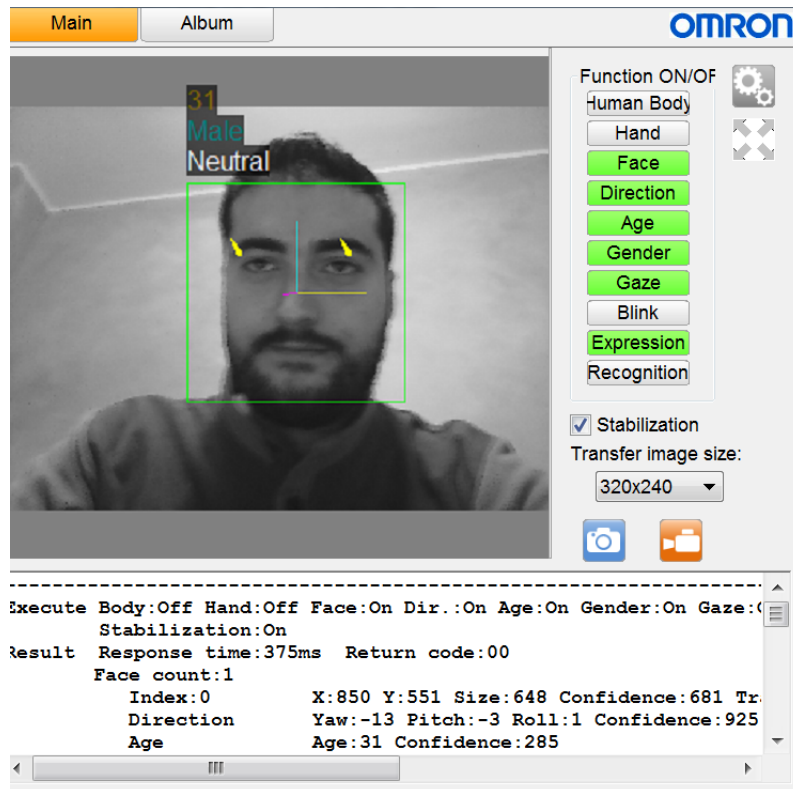


Figure 2.18: Device used during tests.

The device can be used for the identification of the human body (skeletal analysis), tracking of the hand, identification of the face and its recognition (from the previously loaded training database). In particular, it makes the following useful processing for a behavioral analysis:

- FACE DIRECTION ANALYSIS;
- DETERMINATION OF THE AGE;
- SANCTIFICATION OF SEX;

- IDENTIFICATION OF EMOTIONS;
- EYE TRACKING.

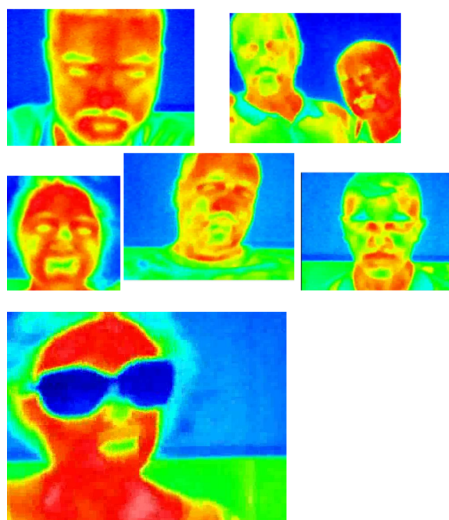


The following is an example of the log returned by the application

```

Execute Body:Off Hand:Off Face:On Dir.:On Age:On Gender:On
Gaze:On Blink:Off Expr.:On Recog.: Off
Stabilization:On
Result Response time:375ms Return code:00 Face count:1
Index:0 X:850 Y:551 Size:648 Confidence:681
TrackingID:0
Direction Yaw:-13 Pitch:-3 Roll:1 Confidence:925
Age Age:31 Confidence:285
Gender Gender:Male Confidence:1000
Gaze Yaw:-5 Pitch:7
Expression Expression:Neutral

```

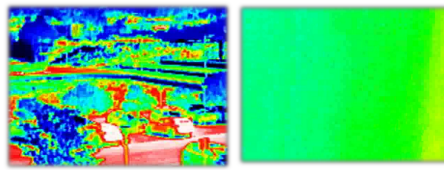


Score : Neutral:98 , Happiness:0 , Surprise:1 , Anger:0 ,  
Sadness:1 Degree:0

The various elaborations take place around the second half, with almost 100% recognition of sex. Moreover, from the tests carried out a recognition of a higher age was obtained also from the literature CLBP.

## 2.11 Thermal Cameras

Recently, the analysis of images acquired by thermal cameras has been introduced within the computer vision scenario. An infrared camera is a non-contact device that detects infrared (heat) energy emitted by an object and converts it into an electronic signal that is subsequently processed to produce a thermal image on a display. The heat detected by a thermal camera can be quantified extremely precisely. Each object emits an infrared radiation, which passes through the thermal lenses of the camera and is processed and returned by the detector as a thermal image. An experiment in the DOOH field was to acquire and perform some experiments on the behavioral analysis performed by analyzing thermal images. For this experiment, an AXIS thermal camera was used. Unlike a conventional camera, a transparent material has a shielding function (the temperature of the first detected material is detected). A glass therefore has a barrier function. The following is an example of detection made with the presence of a window panel, and after opening the window.



The presence of glass makes it impossible to recognize the temperature of the various objects in the scene. Tests were carried out on the GENDER and AGE estimation using images acquired with the camera. The results are much lower than those obtained by analyzing RGB images. However, in the literature it is possible to find experiments with the use of thermal image on FACE DETECTION [100] and EXPRESSION RECOGNITION (it is facilitated by the recognition of temperature) and application cases in the field of STRESS DETECTION, BREATH ANALYSIS and NEUROLOGY [99]. As application developments in DOOH scenario, it is possible to consider the use of thermal images of the consumer to evaluate the "QUALITY OF LIFE" of the interlocutor, and applications related to the GDO as a suggestion of products for customer care (e.g. creams, tanning, hair products) depending on the temperature level detected (e.g. hypothermia detects a significant level of cellulite, it is possible to display a map of the hair) to define the "QUALITY OF SKIN" and "QUALITY OF HAIR".

## 2.12 GDPR and DOOH

As a consequence, the need arises to regulate the use of Big Data with the help of European legislation: the EU 679/2016 General Data Protection Regulation (GDPR) was born from this need, and the aim of this work is to provide an overview of the new legislation and to introduce a new index to measure GDPR compliance (Corrales M. et al, 2017). The GDPR, approved by the European Parliament in April 2016, entered into force on May 25, 2018. The goal is to harmonize the laws on the confidentiality of information and privacy of all European countries and keep safe the sensitive user data processed by companies, and to limit uses according to the principles of [56]):

- lawfulness, correctness and transparency: data must be processed in such ways;

- limitation of purposes: they must be determined, explicit and legitimate, then clearly identified;
- data minimization: data must be adequate, relevant and limited;
- accuracy: the data must be updated;
- restriction of storage: data must be kept for a limited period of time to achieve the purposes;
- integrity and confidentiality: adequate security of personal data must be guaranteed.

The GDPR, replacing the regulations of the individual European countries that differ from one another, represents an important step forward in terms of standardizing European policies and data protection at the continental level [57]. What changes is the extension of the jurisdiction to all companies that process personal data of subjects residing in the European Union, regardless of the geographical location of the company or the place where the data are managed and processed. Non-European companies that process data of European citizens will also have to appoint an EU representative [58]. It is essential that European companies identify immediately how to adapt to the new legislation, thus avoiding being unprepared to face what is considered the most significant change in the history of data protection over the last 20 years. It is necessary that companies immediately review their internal processes, placing user privacy as a primary element to guarantee priority and precedence. It is also necessary for companies to strengthen internal corporate communication through specific training programs so that anyone in a position that implies access to personal data of users correctly knows the extent to which they can carry out their profession. The concept of "privacy by design", a fundamental point on which the GDPR is concentrated [59] establishes that the data protection measures must be planned with the relative supporting IT applications starting from the planning of the business processes. This implies that only the data that are really indispensable for the performance of one's professional duties are processed and that access to information is limited only to those who have to carry out the processing. Another important point of the legislation concerns the "Breach Notification": data breach notifications are mandatory where the violation

can put at risk the rights and freedoms of individuals. The notification must be made within 72 hours from the time the violation is verified and the customers are required to be informed "without undue delay". The changes that the GDPR will bring are not only linked to the relationship between companies and users, but also concern the internal structure of the company: the new legislation will give greater prominence to the IT team and the company CIOs, making their tasks, nevertheless many managers still consider the GDPR as a waste of money and time, not understanding the importance of data protection today [64]. With the GDPR, the figure of the Data Protection Officer (DPO) is established within the company with the task of monitoring the internal processes of the structure and acting as a consultant: the controllers of the monitoring and data processing activities are still required to notify their activities to local Data Protection Advisors (DPAs) which, for example within multinationals, can be a real bureaucratic nightmare, since each Member State has different notification requirements [60]. With the introduction of the DPO, appointed on the basis of professional quality, expert in the field of law and data protection practices and equipped with the appropriate resources, the control of internal data management processes will be simplified. The new legislation pays particular attention, in addition to what has already been said, to the requests for consent made to the subjects [61]: the GDPR wants the requests to be submitted to the user in an "intelligible and easily accessible" manner, so that it is immediately clear what is the purpose of data processing. The companies will also have to guarantee users the right to delete personal data (Right to be forgotten), the possibility to request information about their treatment and to obtain a free copy in electronic format. The new regulation will be the cause of severe sanctions for companies that do not respect it, with fines of up to 4% of the total annual turnover or 20 million, whichever is the greater of the two. But the consequences will not be only economic: failure to comply with the new rules will also have repercussions on the reputation and image of the company, which will not be considered as attentive to the privacy of users and their sensitive data. The GDPR has shed light on the issues of Data Protection [62], a theme that, also due to the latest cyber attacks, requires ever more attention. It is well known that the threats against IT security and data protection are not going to decrease: just think of the recent attack of the WannaCry ransomware that hit more than 150 countries between Europe and



Asia causing serious damage all over the world. Such a serious attack makes us understand the skills of today's hackers, always in search of flaws and inadequacies in IT systems, which must also be protected with the help of specialists in the sector [63]. By taking advantage of effective security solutions, companies can protect themselves completely, thus guaranteeing their users that their data is always safe and that there is no risk of it being lost. The use of computer vision applied to DOOH must necessarily be subject to compliance with current regulatory requirements. The face test must therefore be managed in accordance with the new data protection regulation. In this regard it is necessary, for real-time analysis without any registration of the face, to provide appropriate information to the processing of data to the user concerned. This provided you do not save this information associated with personal data or for profiling. In fact, from the analysis of the face it is possible through the behavior analysis to determine a behavioral state that represents in all respects a particular information, subject to the related DPIA (Data Protection Impact Assessment) by the data controller. In the event of an anonymous real-time acknowledgment, the system, upon presentation of an information notice to the processing of data, may use this information to perform certain operations (e.g. statistics, provision of a coupon, interaction with the user). In the event that the information must be acquired together with the identification data, this aspect necessarily requires an explicit and optional consent to the user. The user will then authorize the acquisition of this feature for purposes defined by the data controller, in the manner and in the times set out in the information sheet together with the consent. Please note that for profiling activities, consent must be obtained every year and for marketing activities every 2 years. If such acquisition activity involves features such as age and sex, the DPIA will represent a low risk, different evaluation concerns as regards ethnicity, recognition of emotions and behavioral states. These aspects, in fact, represent particular data, the use of which must be justified. Performing a profiling of a user with regard to such data represents in all respects an important injury to the person, and could be used for discrimination even through automated choice. Consequently, since there is no prior request to the Guarantor, it is obvious that a high-risk DPIA exposes the data controller very seriously. This aspect naturally concerns both in-house applications and through outsourced systems. In fact, the outsourcer in all respects represents the person in charge of data

processing, i.e. the supervisor of operational and technical measures, guaranteeing the RID requirements (confidentiality, integrity and availability) of the service, but does not determine the purposes, which are expressly Head to the Data Controller. Even the aid of a Data Protection Officer (DPO) would not be a sufficient guarantee on the considerable risk associated with these treatments. Consequently, behavior analysis in the context of GDPR should be limited to real-time uses. Even the justification for the use of sex and age memorization would be questionable, as these statements can be provided by the user in a more lawful and less invasive way towards the data subject. This scenario therefore opens up countless debates, and requires the competent authorities to give clear indications regarding the DPIA, as done for video surveillance systems and GPS systems for remote control, which are in turn regulated and authorized by the DTL (direction territorial work) according to the workers' statute. In fact, in case of use of computer vision devices in environments where there is the presence of employees, a prior authorization from the competent bodies (DTL or RSA) is always mandatory. Engaging in compliance with the GDPR is an important collective achievement. With regard to the permitted behavior analysis activities, it is therefore necessary that no data recording is made (visual assays) and no correlation is made between the processed test / information and the customer data. For this reason it is necessary to transform the acquired information in a way that is not attributable to an identifiable person, and is provided for inclusion in the treatment register with the related DPIA. Register of the Data Controller and related DPIA compliant to GDPR for DOOH realtime application in a retail context.

#### REGISTER OF THE HOLDER OF THE TREATMENT

NAME OF TREATMENT: Behavior Analysis activity for real-time marketing purposes without profiling

PRODUCT DESCRIPTION / PROCESS / SERVICE: DOOH system (Interactive Kiosk) present at the point of sale. The system interacts with the user thanks to the feedback obtained from the analysis of visual assays acquired by means of a camera. The system does not record personal information and does not require to enter data related to the loyalty card thus avoiding any profiling activity.

DESCRIPTION OF DATA PROCESSING: The data processing will concern the acquired real-time visual assay (without conservation) and elaborated by the algorithm present within the system. Only a descriptor extracted from the image and not invertible can be saved for statistical and research purposes (it will never be possible, starting from the descriptor, to reconstruct the wise of the acquired face and then identify the subject).

PURPOSE: Increase of business productivity thanks to customer behavioral analysis

CATEGORIES OF INTERESTED: Customers

CATEGORIES OF PERSONAL DATA: Visual essay

TECHNICAL MEASURES ADOPTED: The visual essay of the face is acquired and processed without any preservation of the image. This process therefore allows us to preserve the identity of the subject

INFORMATION: Yes, information provided to the customer before interaction

INFORMED CONSENT: NO (Legitimate interest of the data controller). No recording of the visual test is performed.

DURATION OF TREATMENT: the essay is acquired and elaborated in a real-time way. Treatment is immediate. Data (descriptors) kept for production and research purposes are not subject to duration of treatment.

#### DPIA (DATA PROTECTION IMPACT ASSESTMENT)

MANAGEMENT MODE: The system is installed inside the computers located in the point of sale interactive kiosks. Real-time processing takes place within individual computers. No retention of the face test is performed.

STORAGE TIME AND ACCESS MODE: The acquired real-time data are saved as a descriptor. However, it would be possible to reconstruct the visual test from the information present in the RAM memory of the PC. For this reason access to the station is reserved and profiled. This guarantees the confidentiality, integrity and availability of the data.

TECHNICAL DETAILS: The system is developed in PHP language and uses an SQL database. The system is active on Apache webserver present on Linux Operating System.

## Chapter 3

# VISUAL MARKET BASKET ANALYSIS

### 3.1 Market Basket Analysis

Market Basket Analysis [105] is a technique aimed at uncovering associations and connections between specific objects. In this kind of analysis, you look for combinations of products that frequently co-occur in transactions. For example people who buy flour and sugar might be looking for eggs as well (all these products are needed in order to bake a cake). Business men can thus take their decisions relying on the outcomes of this analysis. Scope of this chapter is to introduce Market Basket Analysis and its evolution through Computer vision, called Visual Market Basket Analysis.

#### 3.1.1 Terminology

To introduce Market Basket Analysis it is very important to introduce reference terminology. Items are the objects among which we are identifying associations (we will use the letter “i” and a subscript in order to identify an item) Item set (or itemset) is a group of items; we will refer to itemsets containing n elements as n-itemsets.

$$I = \{i_1, i_2, \dots, i_n\}$$

**Transactions** are instances of groups of items co-occurring together.

$$I = \{i_i, i_j, \dots, i_k\}$$

**Rules** are statements that infers an item from an item set (the item inferred does not belong to the item set). We will refer to the set on the left of the arrow as the Left Hand Side (**LHS**) of the rule, while the item on the right will be called Right Hand Side (**RHS**) of the rule

$$I = \{i_1, i_2, \dots, i_k\}$$

**Support** (of an item or item set) is the fraction of transactions in a dataset that contain an item or an item set; when referred to a rule we calculate the support of the union of the sets participating in the rule **Confidence** (of a rule) is the likelihood that a transaction including a certain item set also contains the item implied by the rule confidence

$$Confidence(I_m \Rightarrow I_n) = support(I_m \cup I_n) / support(I_m)$$

**Lift** (of a rule) is the ratio of the support of the item set (which the rule refers to) co-occurring with the items implied by that rule, divided by probability that these items co-occur if they are mutually independent

$$lift(I_m \Rightarrow I_n) = support(I_m \cup I_n) / (support(I_m) \times Support(I_n))$$

### 3.1.2 How it works

The data we need to analyse are usually stored in a binary matrix in which each line displays a transaction and each column represent an item. There are many different algorithms that can be used to generate the rules, but most of them share the concept that the support of a rule is at most equal to the support of the set generating it; in this way it is possible to study only the rules that are considered valuable. It is not uncommon to make use of commercial software that aims to visually represent the outcomes of the market basket analysis, in order to make it easier to understand. In general, it is nice to identify rules that have a high support, as these will be relevant to a large number of transactions. If lift is greater than 1, it suggests that the presence of the LHS of a rule has increased the probability that the RHS will occur on this transaction. If the lift is less than one means that the presence of the LHS of a rule make it less likely to find the item on the RHS on the

same transaction. If the lift equals 1, it means that the data we are analysing don't give useful in order to predict the presence of the RHS.

## 3.2 MBA technical approach

The core of market basket analysis is constituted by the Frequent Itemsets Mining (FIM) algorithms used to create the association rules. We say that an itemset is frequent if it has a support that is not below a given minimum threshold.

### 3.2.1 Apriori

The apriori algorithm [102] is based on an horizontal breadth-first search algorithm. The apriori principle can reduce the number of itemsets we need to examine. Put simply, it states that if an itemset is infrequent, then all its subsets must also be infrequent. The algorithm follows these steps:

1. Start with itemsets containing just a single item
2. Determine the support for itemsets. Keep the itemsets that meet your minimum support threshold, and remove itemsets that do not
3. Using the itemsets you have kept from the previous step, generate all the possible itemset configurations
4. Repeat the steps 2 and 3 until there are no more new itemsets.

It has to be noticed that the support threshold that you pick up in the second step could be based on formal analysis or past experience. If you discover that sales of items beyond a certain proportion tend to have significant impact on your profits, you might consider using that proportion as your support threshold. Once high-support itemsets have been identified, using the apriori principle to identify item associations with high confidence or lift would be less computationally expensive, since confidence and lift values are calculated using support values. Even though the apriori algorithm reduces the number of candidate itemsets to consider, this number could still be huge when dealing with large datasets or setting a low threshold. However, an alternative solution would be to reduce the number of comparisons by

using advanced data structures, such as hash tables, to sort candidate itemsets more efficiently. When working on large datasets, we might need to reduce the support threshold in order to detect certain associations, though doing this can lead to a higher number of deceitful associations detected.

### 3.2.2 Eclat

The Eclat algorithm [103] improves upon the Apriori approach by using a depth-first search to avoid keeping many itemsets in memory. The Eclat algorithm relies on what is called a vertical database representation; this kind of representation indicates the list of transactions where each item appears. For an itemset  $i$ , the list of transactions containing it is called its TID-list (Transaction IDentifiers) and it is denoted as  $\text{tid}(X)$ . This vertical representation can be obtained by scanning the original database only once and it is very useful in itemsets mining because of the following properties:

1. For any itemsets  $X$  and  $Y$ ,  $\text{tid}(X \cup Y) = \text{tid}(X) \cap \text{tid}(Y)$
2. For any itemset  $X$ ,  $\text{support}(X) = |\text{tid}(X)|$

Using these two properties, vertical algorithms such as Eclat can explore the search space by scanning the database only once to create the initial TID-lists. The algorithm follows the below steps (assuming we have already scanned the database the first time, let  $I$  is above the chosen threshold):

1. Take an itemset  $X \in I$
2. Perform a search to find frequent itemsets extending  $X$  with one item and adds them to a set  $E$  of frequent extensions of  $X$
3. Execute recursively the previous steps on the elements included in  $E$
4. Perform these steps for each  $X \in I$

### 3.2.3 Pattern-growth

To face the main limitation of algorithms such as Apriori and Eclat, a major advance in the field has been the development of pattern-growth algorithms such as FP-Growth, H-Mine and LCM. We can assume without loss of generality, that there



exists a total order on the set of all items, such as the lexicographical order. The main idea of pattern-growth algorithms [104] is to scan a database in order to find itemsets, and thus avoid generating candidates that do not appear in the database. Furthermore, to reduce the cost of scanning the database, pattern-growth algorithms have introduced the concept of projected database to reduce the size of databases as an algorithm explore larger itemsets with its depth-first search. The projected database of an item  $i$  is defined as the set of transactions where  $i$  appears, but where the item  $i$  and items preceding it have been removed. A pattern-growth algorithm explores the search space using a depth-first search by recursively appending items, according to the said order, to frequent itemsets, to obtain larger frequent itemsets.

In order to simplify the explanation of this kind of algorithms, we will use the notation here explained:

- $D$ : a transaction database
- $t$ : more responsibility
- $X$ : more satisfaction

A pattern-growth algorithm goes on following these steps:

1. Take an itemset  $X = \emptyset$
2. Scan the database  $D$  to find the set  $Z_D$  of all frequent items in  $D$
3. Choose a  $[z \in Z_D]$  according to the existing order, start from the first
4.  $X^1 = X \cup \{z\}$  is a frequent itemset
5.  $D^1 = projection(D, X^1)$
6. Execute from step 2 replacing  $D^1$  to  $D$  and  $X^1$  to  $X$  until no more frequent itemsets are found
7. Execute from step 3 for each  $z \in Z_D$

A major advantage of pattern-growth algorithms is that they only explore the frequent itemsets in the search space, thus avoiding considering many itemsets not appearing in the database or infrequent itemsets. Besides, the concept of projected database is also useful to reduce the cost of database scans. A common question about the concept of projected database is: is it costly to create all these copies of the original database? The answer is no if an optimization called pseudo-projection is used, which consists of implementing a projected database as a set of pointers on the original database, to avoid creating a copy of the original database. Note that many other optimizations can also be integrated in pattern-growth algorithms. For example, **LCM** also integrates a mechanism to merge identical transactions in projected databases to further reduce their size, and an efficient array-based support counting technique called occurrence-delivery. The **FP-growth** and **H-mine** algorithms respectively introduce a prefix-tree structure and a hyperlink structure for representing projected-databases to also reduce memory usage.

### 3.2.4 Observation

When using Market Basket Analysis to make any kind of business decisions, we need to keep in mind that the output of the analysis reflects how frequently items co-occur in transactions; we must be aware that this is a function of both the real correlation between items and the way in which those items are displayed to customers. This means that items presented next to one another are likely to result related at the end of the analysis, but the only correlation between them might be the placement inside the store (or website). The point of using Market Basket Analysis is seeking to understand the purchase behaviour of customers. This information can then be used for purposes of cross-selling (selling an additional product or service to an existing customer) and up-selling (inducing customers to purchase more expensive items), in addition to influencing sales promotions, loyalty programs, store design and discount plans.

### 3.3 Egocentric vision for Visual Market Basket Analysis

The evolution of Market Basket Analysis using Computer Vision may be defined Visual Market Basket Analysis [5]. In this paragraph is explained all the steps followed to perceive the result. Egocentric vision is a new emerging area in Computer Vision [52,53]. By exploiting wearable devices it is possible to collect hours of videos that can be processed to obtain a log of the monitored scenarios. Different papers on egocentric vision applications have been published in the recent literature. The main tasks addressed in this area are related to scene recognition [55], motion understanding [65], objects and actions recognition [54,76–78], 3D reconstruction [74,75] and summarization [70,73]. Among the others, context aware computing is an important research area for egocentric (first-person) vision domain [55,79,80]. Temporal segmentation of Egocentric Vision is also fundamental to understand the behavior of the users wearing a camera [25,65]. Recently, the retail scenario has become of particular interest for applications related to the geo-localization of the user's positions and the reconstruction of the spaces [67]. In the retail context, one of the possible developments of interest concerns the monitoring of the paths of customers, thereby enabling to carry out an analysis of their behaviors. Nowadays customers monitoring it is partially employed by using loyalty cards, counting devices connected with Bluetooth and WiFi systems, employing RFID tags [68], as well as fixed cameras (e.g., video surveillance). Differently than classic approaches %, drawing on the approach of the monitoring of carts based on RFID tags [68], and considering the potentials and spread of egocentric cameras, in this work we consider to turn an ordinary cart in a 'narrative shopping cart' by equipping it with a camera. The acquired egocentric videos are processed with algorithms able to turn the visual paths in customers' behaviour. By doing so it is possible to acquire all over the route travelled by carts and (hence by the customers), from the cart picking to its release. Visual and audio data can be collected and processed to monitor pauses, understanding the areas of personal interest, estimating the path speed, estimates the most busy areas of the retail by clustering routes, registering the reactions opposite to the audio announcements in the store, as well as infer the inefficiencies (e.g., slowness at cash desk). We call this kind of behavioral monitoring in a retail



Figure 3.1: Information useful for VMBA.

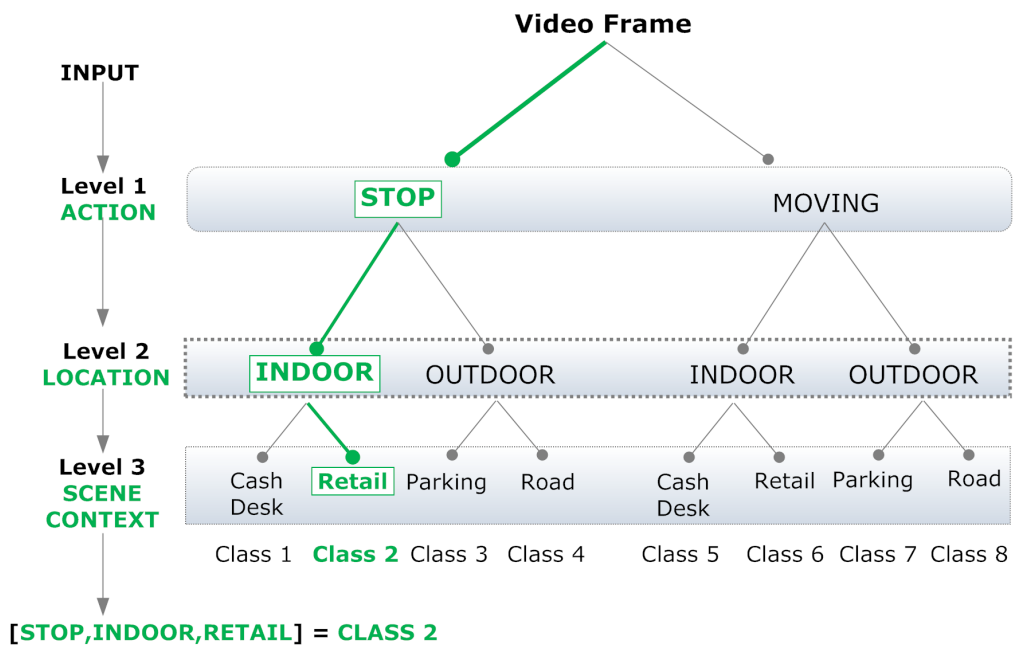


Figure 3.2: Considered VMBA behavioral classes organized in a hierarchy.

‘Visual Market Basket Analysis’ (VMBA) since it can be useful to enrich the classic ‘Market Basket Analysis’ methods [66] used to infer the habits of customers.

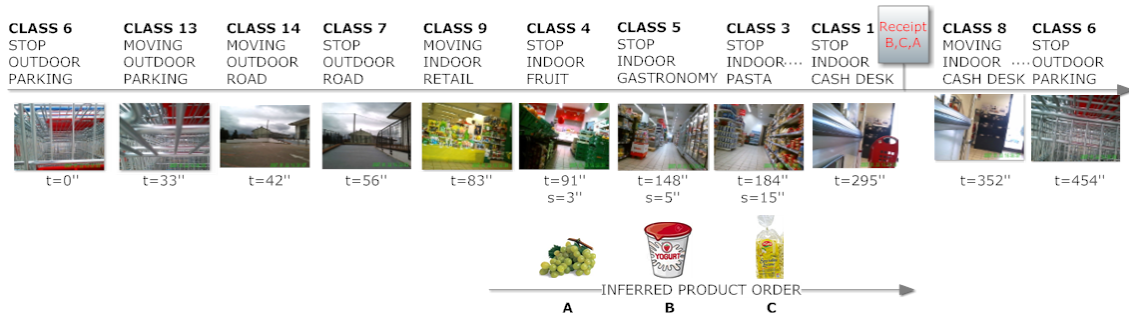


Figure 3.3: VMBA timeline temporally segmented considering the 8 classes.  $t$  denotes the time, whereas  $s$  denotes the stopping time.

### 3.3.1 Methods for VMBA

At first we have introduced the problem of VMBA considering three different high-level information related to the customers which is carrying the narrative cart (Fig. 3.1): location (i.e., indoor vs outdoor), action (i.e., stop vs moving), and scene context (i.e., cash desk, retail, parking, road). These high-level information can be organized in a hierarchy to produce 8 different behaviors useful in the retail domain to log the storyline of the shopping of the customers that can be eventually associated to others information (e.g., receipts) for retail management purposes. The 8 classes are shown as path, from the root to the leaves, of the tree in Fig. 3.2. Given a frame of the video acquired with the narrative cart camera, at each instant we wish to know a triplet corresponding to a path in the tree (e.g., [STOP, INDOOR, RETAIL] in Fig. 3.2). By classifying each frame of the acquired egocentric videos with the proposed 8 classes (i.e., the 8 possible triplet of the hierarchy in Fig. 3.2), it will be simple to perform an analysis of what are the custom behaviors, and also understand if there are problem to be managed in the store. An example of a narrative cart egocentric video together with a temporal segmentation with respect to the 8 defined classes is shown in Fig. 3.31. For example from the segmented narrative cart video it will be simple to understand how long are the stops to the cash desk by considering the frames classified with the triplets [STOP, INDOOR, CASH DESK] and [MOVING, INDOOR, CASH DESK]. This can be useful to eventually plan the opening of more cash desks to provide a better service to the customers. By analyzing the inferred triplets of a narrative cart video, it will be simple to understand if there are carts outside the cart parking spaces in order to take actions (e.g., if

there is a long sequence of the triplet [STOP, OUTDOOR, ROAD] which does not change for long time). A lot of other considerations for a better management of the retail can be done by considering the narrative cart egocentric videos when those have been temporal segmented by classifying each frame with the 8 possible triples (i.e., behavioral classes). By combining the receipt with the temporal segmented video and algorithms for visual re-localization [81] it will be simple to establish the order in which the products have been taken, hence increasing the information that are usually exploited by the classic "Market Basket Analysis" algorithms [66] and opening new research perspectives (Fig. 3.31). To set the first VMBA challenge we propose a new dataset of 15 sequences (VBMA) obtained by collecting and labeling real video sequences acquired in a retail. The proposed dataset is available for the research community upon request to the authors. We benchmark the dataset by considering a Direct Acyclic Graph SVM approach [72] coupled with classic descriptors for the representation of visual content (GIST [71]), motion (Optical Flow [69]), and audio (MFCC [88]). Experiments show that a classification accuracy of more than 93% can be obtained on the proposed VMBA dataset when the 8 behavioral classes are considered.

### 3.3.2 Actions

The first level analyzes the customer behavior from the point of view of the motion of the narrative cart by considering two possible states: stop and moving. In order to understand such states from the egocentric video, in our benchmark we tested the MFCC audio features [83] and the optical flow features computed with the classic block matching approach [84] (Fig. 3.4). For the optical flow we have just considered the frames divided in 9 blocks, so for each frame we have got a 9-dimensional features vector. The audio processing produced a feature vector of 62 components. We have decided to exploit audio because there is %by looking at the audio waveform (Fig. 3.5) we have noticed that there is a visual correlation between the audio waveform with the narrative cart motion and locations (Fig. 3.5). The exploitation of the optical flow feature is straightforward since the problem to be solved. In our experiments we have tested the two considered features separately and jointly.

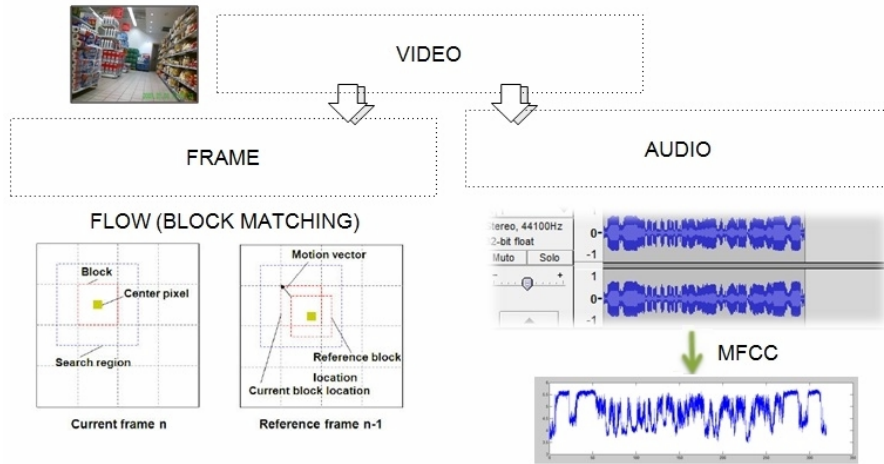


Figure 3.4: Features used at the first level.

### 3.3.3 Location

The second level of the tree in Fig. 3.2 has the scope to identify the high level location where the user is acting: indoor vs outdoor. As for the first level we have considered MFCC features after visual inspection of waveform (Fig. 3.5). Indeed the waveform is more pronounced in the outdoor environment than in the indoor location. To benchmark the VMBA problem addressed in this thesis to discriminate indoor vs outdoor locations we have also tested the GIST visual descriptor [89], which is able to encode the scene context with a feature vector composed by 512 components (Fig. 3.6). In our experiments we tested the indoor vs outdoor discrimination by considering audio and visual features independently and combined.

### 3.3.4 Scene Context

The third level of the hierarchy in Fig. 3.2 is related to the analysis of the scene context considering four different classes: cash desk, retail, parking and road. As described before, the first two contexts are related to the indoor environment, whereas the other two describe in more details the outdoor location. For this level of description we have used the GIST descriptor [89] again since its property in capturing the shape of the scene for context discrimination.

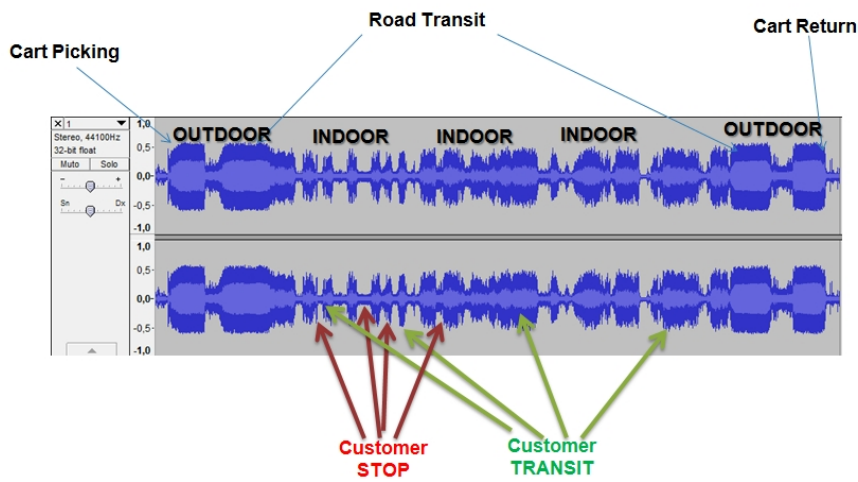


Figure 3.5: Audio waveform and behaviors.

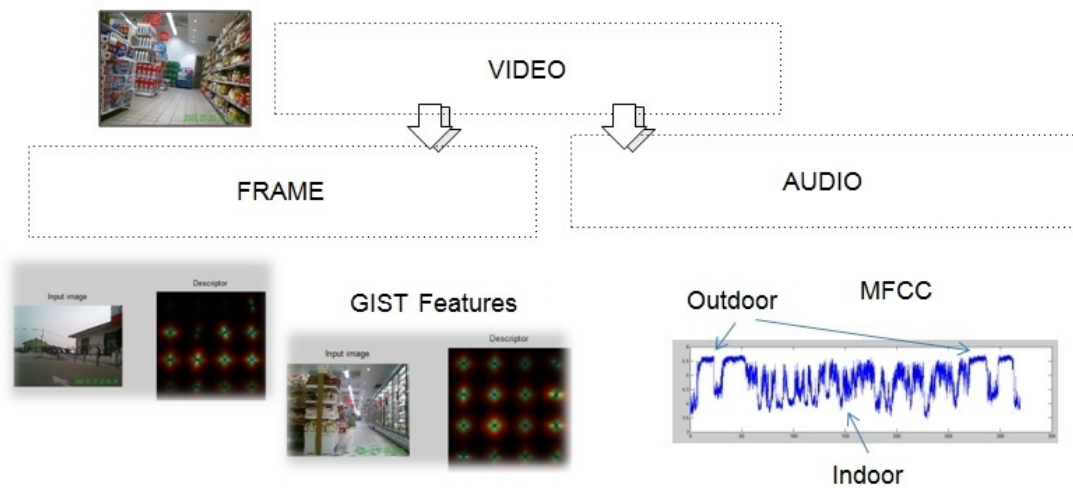


Figure 3.6: Features used at the Second level.

### 3.3.5 Classifications

After representing a frame of the egocentric video as described in previous sections, a classifiers have to be employed to infer one of the 8 considered classes (i.e., one of the 8 possible triplet corresponding to a path of the tree in Fig. 3.2).





Figure 3.7: Narrative Carts.



Figure 3.8: Some visual examples of frames related to the egocentric videos of the VMBA15 dataset. The eight scenes represent the eight possible classes with order from top to bottom, left to right. Notice that some classes are characterized by similar visual content but different actions, such as in the case of the image at the first row of the second column (CLASS 2) and the third image in the second column (CLASS 6). The images at the first row are related to CLASS 1 (left) and CLASS 2 (right), and share the same location (INDOOR) but show different scene context (RETAIL vs CASH DESK).

In this work we benchmarked three different classification modalities:

- combination of the results obtained by three different SVM classifiers in correspondence of the three different levels of the hierarchy;
- a single SVM trained on the 8 possible classes;
- a Direct Acyclic Graph SVM learning architecture (DAGSVM) [72] which reflects the hierarchy in Fig. 3.2 on each node.

Experiments reported in Section IV demonstrate that good classification accuracy can be obtained considering the hierarchical classification with DAGSVM.

### 3.3.6 VMBA15 Dataset

To set the first VMBA challenge and perform the benchmark on the considered problem we acquired a dataset composed by 15 different egocentric videos with narrative carts in a retail of the Southern of Italy during real shopping. To this aim we have mounted a narrative cam veho muvi pro [82] into the front of a classic shopping cart as depicted in Fig. 3.7. Each narrative cart video has a duration between 3 to 20 minutes and resolution of 640x480 pixels. Audio has been also recorded since it can be useful to discriminate indoor vs outdoor environment. From each narrative cart video we have sampled and manually labeled frames and audio at 1 fps considering the 8 possible paths of the tree shown in Fig. 3.2. The total number of sample is 7976 (see Table I for more retails about the dataset). Some examples of frames extracted from the VMBA15 dataset are shown in Fig. 3.32. The labeled data is available upon request to the authors.

Table 3.1: Number of samples per class for each egocentric video.

VIDEO	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$	$C_8$	TOTAL
1	0	13	2	0	8	193	17	89	322
2	0	17	4	0	9	266	10	84	390
3	0	19	4	0	12	226	10	96	367
4	0	20	3	0	10	277	13	106	429
5	0	20	4	0	10	213	9	107	363
6	69	10	28	2	59	134	35	91	428
7	0	3	13	0	7	102	16	119	260
8	6	36	7	0	8	233	8	75	373
9	142	186	8	0	18	550	9	85	998
10	0	5	3	0	7	106	13	75	209
11	42	90	31	0	10	406	16	89	684
12	0	36	22	0	26	436	23	104	647
13	56	80	57	4	7	130	28	133	495
14	50	396	7	0	3	485	11	46	998
15	81	528	0	27	3	310	3	46	998



Figure 3.9: On the left a typical scene of the narrative cart when in the parking space. On the right an example of a frame acquired by the narrative cart in retail. The distribution of vertical and horizontal edges could generate confusion in the classification.

### 3.3.7 Experimental settings and results

We have performed experiments by randomly splitting that dataset in three parts composed by five egocentric video each. The experiments have been repeated three times considering 10 videos for the training and 5 video for the tests. The final results are obtained by averaging among the three runs. As first we have compared the different features employed at the different levels of the hierarchy independently exploiting a SVM classifier with RBF kernel. This was useful to understand which are the best features (or combination of them) to be employed at each level for the final classification of each frame with respect to the 8 classes. In Table II are reported the results of the stop vs moving classification (i.e. First Level). Both audio and visual feature achieve good performance, however, visual feature (the flow) outperforms this audio features with a margin of about 5%. Interestingly the combination of audio and visual features improve the results and obtains an accuracy of 94.50% in discriminating stop vs moving actions. The obtained results pointed out that the combination of MFCC and flow features has to be used at the first level.

Also in the case of the discrimination of the main location where the narrative cart is moving (or stopping), the visual feature outperform audio feature with a good margin obtaining 95.79% of accuracy (see Table III). Differently than in the first level, the combination of audio and visual features do not improve the indoor vs outdoor classification. Hence for the second level we decided to employ the GIST descriptor alone.

For the third level of the hierarchy we have obtained an accuracy of 92.42% with the



Figure 3.10: Some examples of frames with occlusions (at the cash desk).

GIST descriptor. Note that in this case a multi-class SVM with RBF kernel has been trained to discriminate this four possible Scene Contexts without considering the prior indoor vs outdoor. The results respect to the four scene contexts are reported in Table IV. The main confusion is related to the class parking and retail (first column in Table IV). This is probably due to the encoding of the scene information by the GIST descriptor. Indeed, when the narrative cart is in the parking space, the scene is mainly composed by vertical and horizontal edges that can be confused with the vertical and horizontal edges of some scenes in the retail (see Fig. 3.9).

Table 3.2: STOP VS MOVING classification

	FLOW	MFCC	COMBINED
Accuracy%	92.50	87.04	94.50
TP RATE%	73.03	61.54	84.76
TN RATE%	99.18	95.21	97.65
FP RATE%	0.82	4.79	2.35
FN RATE%	26.97	38.46	15.24

Table 3.3: INDOOR VS OUTDOOR classification

	GIST	MFCC	COMBINED
Accuracy%	95.79	88.00	91.77
TP RATE%	89.3	49.51	67.49
TN RATE%	97.8	97.66	97.1
FP RATE%	2.20	2.34	2.90
FN RATE%	10.7	50.49	32.51

Table 3.4: Scene Context classification

		PREDICTED		
	PARKING	ROAD	RETAIL	CASH DESK
PARKING	54.25%	17.01%	25.09%	3.64%
ROAD	0.55%	88.94%	9.5%	1.01%
RETAIL	0.17%	1.09%	98.46%	0.28%
CASH DESK	0.13%	7.47%	17.21%	75.19%

As demonstrated by the results reported later, this problem is mitigated when the classification is performed by the DAGSVM approach since it introduces a prior on the main location (indoor vs outdoor). One more problem in the classification is due to strong occlusions as the one in the examples reported in Fig. 3.10.

The aforementioned experiments pointed out that the best features to be employed in the hierarchy are the combination of MFCC and FLOW for the first level, whereas the GIST descriptor for the second and third level. Since the main goal is the classification with respect to the 8 possible triplets generated by the hierarchy in Fig. 3.2, after selecting the features for the three levels independently we have compared the three classification modalities described in Section II.D . For the combination of three different classifiers (one for each level) we have considered the concatenation of the labels given by three different SVM (with RBF kernel) when trained independently on the best selected features of the three levels. For the multi-class SVM with 8 classes we have trained a SVM with RBF kernel on the concatenation of MFCC, GIST and FLOW features.



Figure 3.11: Examples of frames correctly classified by the proposed DAGSVM approach. These frames are misclassified by the other two compared approaches. The frame on the left is related to the parking space of the carts, but is recognized as retail by both the combined approach and Multi-Class SVM. The frame on the right is related to outdoor, but is recognized as indoor by both the combined approach and Multi-Class SVM.

Table 3.5: Results of the classification considering the 8 classes

	Combination	Multi-Class SVM	DAGSVM
Accuracy%	87.36	69.54	93.47

Finally, we have trained a DAGSVM [72] reflecting the hierarchy in Fig. 3.2. Each node of the DAG is composed by a SVM with RBF kernel in which the best features to solve the problem at each node are exploited. The results of the three different approaches are reported in Table V. The final results are in favor of the DAGSVM approach which obtain an accuracy of 93.47%. It is worth to note that a combination of the MFCC features with the FLOW and GIST descriptors does not allow a multi-class SVM to reach good accuracy (69.54%). Finally, the results of the combination of the three different classifiers stated at the second place in the classification a ranking (87.36%). Visual examples for the assessment of the output given by the proposed DAGSVM-based approach are available in Fig. 3.11 and at the following URL: <http://iplab.dmi.unict.it/icpr2016> .

### 3.4 Patented System

The present system has been described as an invention patent relates to a method for the identification and advanced kinesthetic analysis based on artificial vision and audio information for the control of the process in a delimited area. In particular,

the present invention finds application in shopping centers or shops in general, self-shopping barriers, logistics spaces such as port areas or storage warehouses. The use of artificial vision in retail environments is always of greatest interest given the many applications ranging from security to anti-shoplifting, to the counting of trolleys and people, up to customer re-identification. The methods and systems mainly used are those based on RGB cameras and in recent years also thermal. However, the potential of these systems is not fully exploited today, as the enormous visual and audio information that can be collected by heterogeneous distributed acquisition systems can give a true summary of the sales point trend over time. In fact, technologies such as RFID, Bluetooth Low Energy, surveillance cameras for monitoring objects (trolleys) and for interacting with customers at the point of sale are often used. However, this technology requires multi-level design and installation, and it is difficult to apply because of the dynamism of the sales spaces. At the level of anteriority, we underline the use of "cart tracking" through mainly RFID technology or point of sale surveillance cameras such as the patents CN104637198A and US8325982B1. Knowing how to monitor the behavior of users at the point of sale and evaluating the process at the point of sale in terms of global analysis represents a factor of considerable advantage for a sales space. In particular, the present invention wants to analyse any scenario in which it is necessary to monitor the flow of users, their behavior and the interaction with the process or the processes concerning the scenario analyzed. In supermarkets, a current problem is represented by the evaluation of the result of marketing and communication activities (e.g. promotional result) and the identification of process problems at the point of sale. In this context, the technical task underlying the present invention is to propose a method and relative system for process control by computer vision and audio information. In particular, it is an object of the present invention to provide an automatic method characterized by the acquisition of images and video by means of cameras mounted on board carriages (1) which we can call "narrative cart"; capture images and videos using mobile robots (2) distributed at the point of sale; acquire images and videos through the use of static or mobile interactive kiosks (3) distributed at the point of sale; capture images and videos using mobile devices (4). From the video streams it is possible to extract the relative audio flow allowing to have a "kinesthetic" information characterized by the audio and video components. To this information is



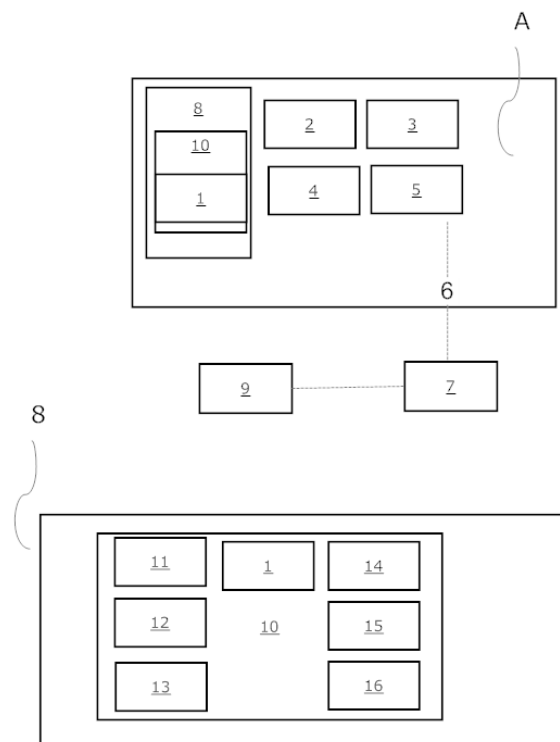


Figure 3.12: Block scheme of Narrative Cart System (shopping cart 8 with cameras 1 in A environment)

added the detection of information through appropriate sensors (5) within a network of multimedia sensors (6) in the Internet of Things. These information flows are then processed by computer (7), returning the numeric, position and trajectories carried out over time by individual customers (narrative cart) (8). This information is useful for extracting an analytical map of the sales point (9). Through this analytical map it is possible to evaluate the customer's behavior, thanks to the information of the breaks in the sales point areas, the interactions carried out in the areas (with objects / shelves / exhibitors / staff), the information obtained from the interactive kiosks, and from the barrier (including information related to loyalty cards).

From the activity in the literature called "market basket analysis", adding visual information, we move to a "visual market basket analysis" or "cynesthetic market basket analysis" crossing the visual and auditory information with those coming from the barrier. In this way it is also possible to calculate an index relating to the effectiveness of communication and marketing activities at the point of sale,



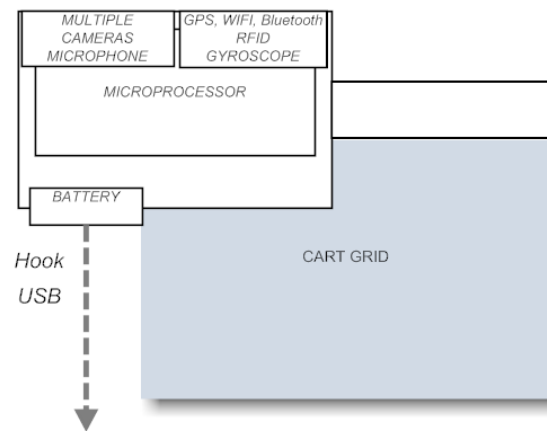


Figure 3.13: Narrative cart setup scheme (embedded device with cameras mounted on cart grid)

thus correcting previous inputs. Furthermore, kinesthetic information based on audio and video is useful for identifying process inefficiencies (e.g. noise pollution, presence of barriers in transit, dirt, queues in the wards, queues at the cash desks). In acquiring visual information it is possible to reconstruct the 3D of the scene using techniques such as SFM. The huge amount of information acquired over time by distributed devices allows you to make a summary of the status of the sales point over time, allowing you to derive the visual genome (synthesis / story) of a point of sale. Moreover, the "narrative cart" allows to provide a real-time tool to support the disabled thanks to the previously introduced tracking technology. The specified technical task and the specified purpose are substantially achieved by a method and system for the identification and advanced kinesthetic analysis based on the artificial vision in a delimited area comprising the technical characteristics exhibited in one or more of the appended claims. The "narrative cart" is a cart characterized by an embedded system as shown in unit 2 composed of a computer equipped with wifi / rfid / bluetooth card, cameras (front / side), microphone, battery, power input capable of acquiring and sending multimedia streams and elaborate information on

the location and identification of the position and objects detected (inside the cart and in the surrounding environment). Further characteristics and advantages of the present invention will become clearer from the indicative, and therefore non-limiting, description of a preferred but not exclusive embodiment of a method and system for advanced kinesthetic identification and analysis based on artificial vision and audio information for control of the process in a delimited area, as illustrated in the accompanying Figure 1 which is a schematic representation of a method in accordance with the present invention. The method of identification and advanced analysis based on the artificial vision in a delimited area according to the present invention is applicable in any case within a delimited area (A) closed or open. It is evident that the system is more efficient in well-defined areas (e.g. point of sale). The method provides, preliminarily, to acquire a continuous video stream of a scene through cameras of the "narrative cart", to identify the position in time of the cart, and the interactions carried out in its surroundings, to correlate this information with that acquired by the cameras of the "Mobile robots" distributed at the point of sale (using face / people detection algorithms, product recognition), with that acquired from interactive kiosks, extracting the map of the point of sale of the scene (A), determining an index on the effectiveness of communication-marketing of the sales point according to the information extracted from the distributed devices and the checkout barrier, monitor and generate alerts for anomalies detected at the point of sale. This analysis can be viewed in real time by the employee of the cash barrier (e.g. on monitor / tablet), and produce alerts and snapshots of the state of the point of sale on site, produce analysis of "visual market basket analysis" for users of the site. The method and system based on "narrative cart" also allows for navigation within the point of sale for the availability of specific products to be used through applications for smartphones / tablets, interactive kiosks or mobile robots for assistance. For the disabled, it is also possible to exploit this technology to allow assisted navigation at the point of sale. This approach can be summarized by a representation of 3 Person Vision Equipment (First, Second, Third), a module for Kinesthesia Analysis and an Environment Sensory Network. First person vision is characterized by the use of egocentric camera, Second person vision is done by robots and kiosks that interacts with the subject, Third person vision is obtained by other video acquisition (for example TVCC system of the scene).

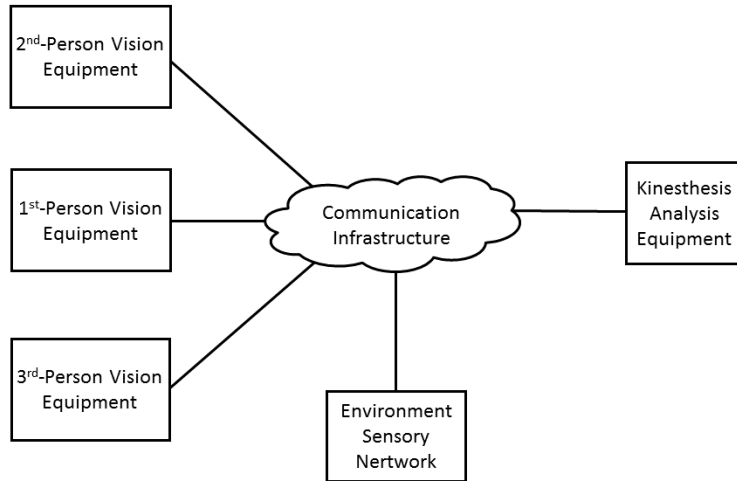


Figure 3.14: Vision Equipment

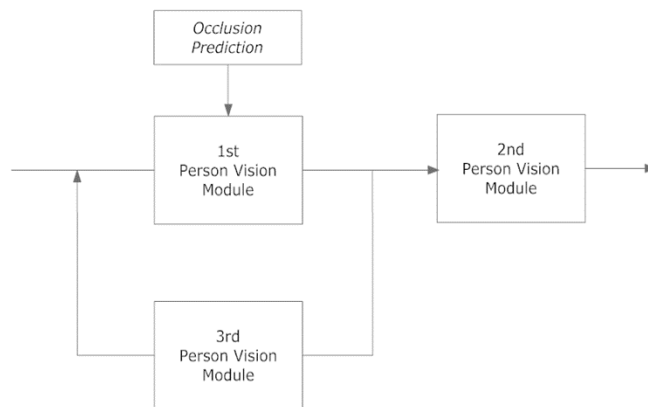


Figure 3.15: Person Vision Block System



Figure 3.16: Difference between Machine Learning and Deep Learning.

### 3.5 Deep Learning Overview

It is very important, to show the evolution of our system, to provide an overview of deep learning, of differences with respect to machine learning, with particular focus on CNN (convolutional neuronal networks) [108]. The explanation will be done with the use of practical examples of implementation and application to images of retail scenarios. To start it is very important to show what is machine learning and what is deep learning.

Deep learning is a specific form of machine learning. The machine learning process begins after the extraction of the features (according to a particular descriptor), making the classification through the realization of a model. In deep learning there is not a preventive phase of extraction of features, but the process starts from the images, making the automatic extraction of the features and the relative classification. An example of Deep Learning is provided by convolutional neuronal networks. In deep networks from an image the features are extracted using layers to arrive at the final classification. Then, a deep network is a neural network characterized

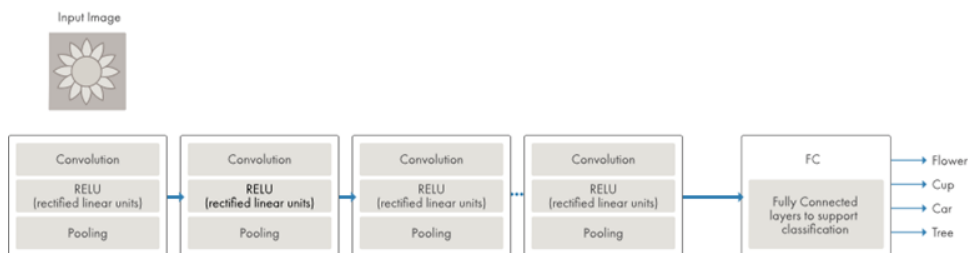


Figure 3.17: Input Image

by different layers. For each image given as input, in different resolutions, filters are applied (layers) and the output of each convex image is used as input for the next layer. Convolutional layers can initially extract very simple features, such as

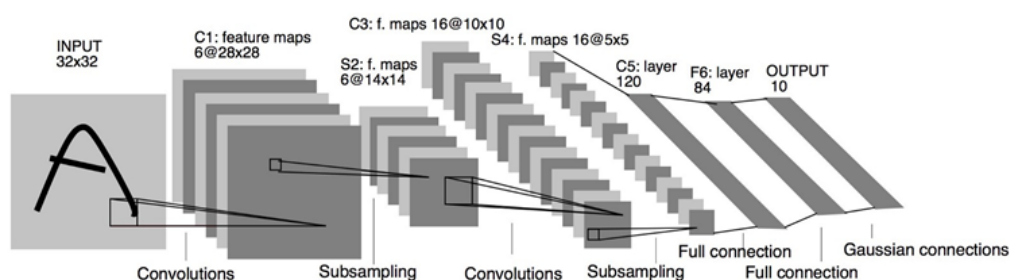


Figure 3.18: CNN model by Yann Lecun.

brightness and edges, to take on increasingly complex shapes up to the recognition of faces, objects and scenes. The final layer (fully connected) with the softmax (statistical module) returns the final classification.

CNN networks are inspired by Yann Lecun (1994) model (fig. 3.18). The model used the sequence of 3 layers: convolution, pooling, non-linearity (these are the basic sequences of Deep Learning layers). This model used a multi-layer neural network (MLP) as the final classifier.

### 3.5.1 CNN structure

It is important to introduce the structure of CNN. A CNN network is characterized by a sequence of convolutional layers that are used to extract features. We can consider some main types of CNN layers:

- RELU Layers increase the non-linearity property of the network activation function;
- Pooling layers reduce the number of parameters and control overfitting;
- FullyConnected Layers transform volume to vector (they group the information), they are not followed by convolutional layers.

Convolutional filters (or layers) allow the extraction of features by conveying the input image with filters that allow to recognize lines, angles, outlines, objects, animals, scenes (fig. 3.19) The last layers are useful for normalization (e.g. LRN for normalization) or output layer where from the fully connected layer. using a probability function, the classification is returned.

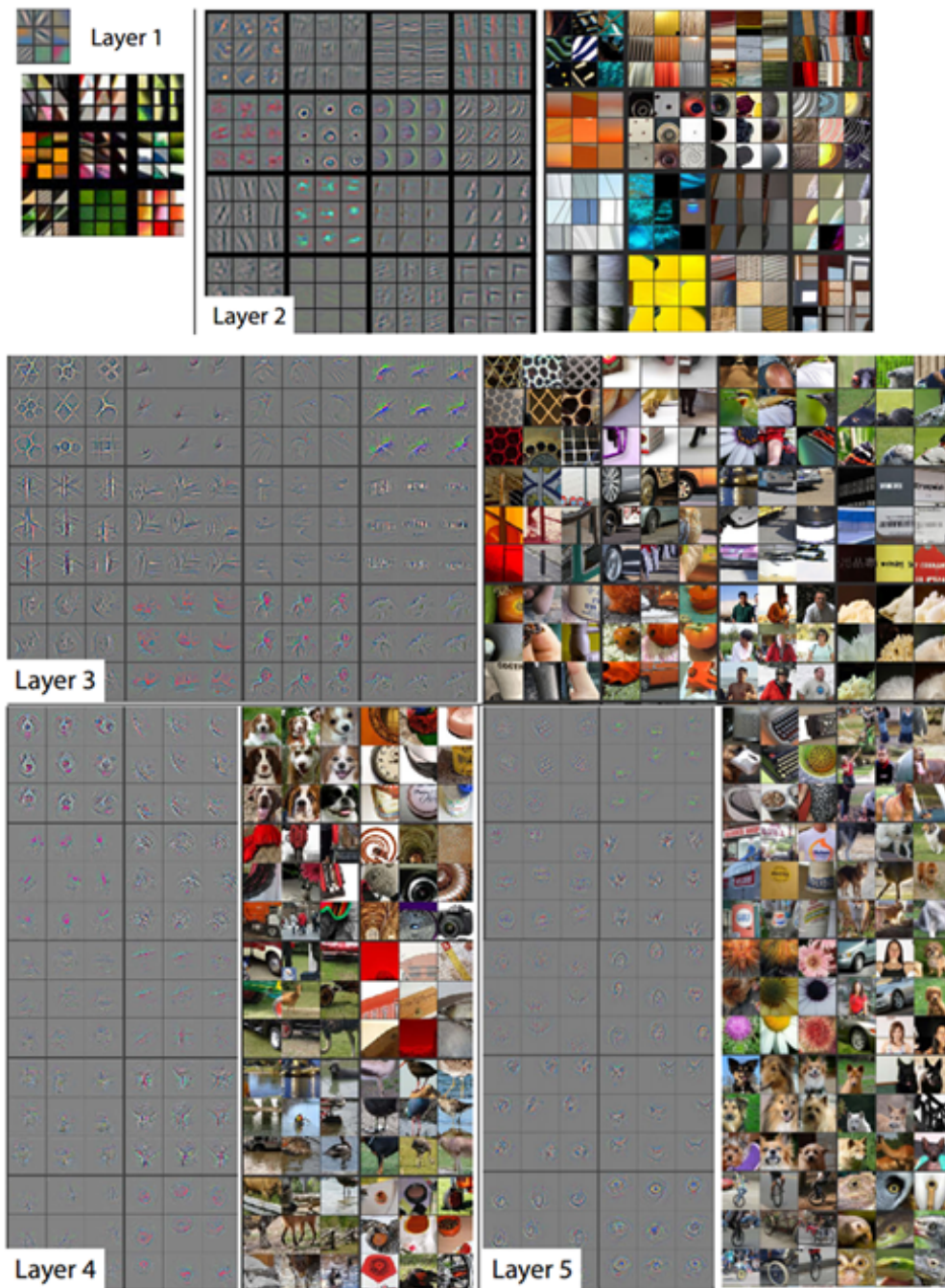


Figure 3.19: Convolutional filters.

### 3.5.2 CNN training

The training of a CNN requires millions of images and long processing times, consequently the three most followed approaches alternative to CNN training which are:

- Use of an **already trained CNN** (if the classes are useful for the purpose)
- Transfer Learning** (tuning a pre-pulled model);
- Use of **models** (e.g. AlexNet) to extract the features of the images to be used in cascade with machine learning (SVM)

Each CNN model has been trained on a specific "big-dataset" of images (e.g. Imagenet) and has its own labeling for classification (e.g. Imagenet labeling). Furthermore, CNN neuronal networks can be feed-forward (networks stratified by layers) where the nodes of each layer are connected to those of the next layer (e.g. AlexNet). However, CNN neuronal networks can be recurrent (RNN) or feedback where the connections between neurons form a direct cycle, i.e. having a direction. In this case, it creates a real internal state of the network, which allows it to have a dynamic behavior over time. The RNN can use their internal memory to process various types of input (e.g. ResNet).

The most important model for CNN networks are:

- VGG-16
- GoogLeNet
- Microsoft ResNet (152 layers)
- AlexNet (25 layers)

The image 3.20 shows a comparison between different CNN models, considering operations and accuracy.

To understand the CNN networks it is fundamental to consider two fundamental parameters, the accuracy and the computational complexity (number of operations required). It is very important to underline that accuracy does not depend only on the network but also on the amount of data available for training. Usually networks are compared on a standard dataset called ImageNet. ImageNet project is an ongoing effort and currently has over 14 million images of 21841 different categories. Since 2010, ImageNet has been running an annual competition in visual

recognition where participants are provided with 1.2 million images belonging to 1000 different classes from Imagenet data-set. So, each network architecture reports accuracy using these 1.2 million images of 1000 classes. The second parameter is the computation. In fact, most CCN have huge memory and computation requirements, especially during training. The size of the final trained model is very important for mobile application. A more computationally intensive network produces more accuracy, then, there is always a trade-off between accuracy and computation. In addition, there are many other factors as ease of training and the ability of a network to generalize.

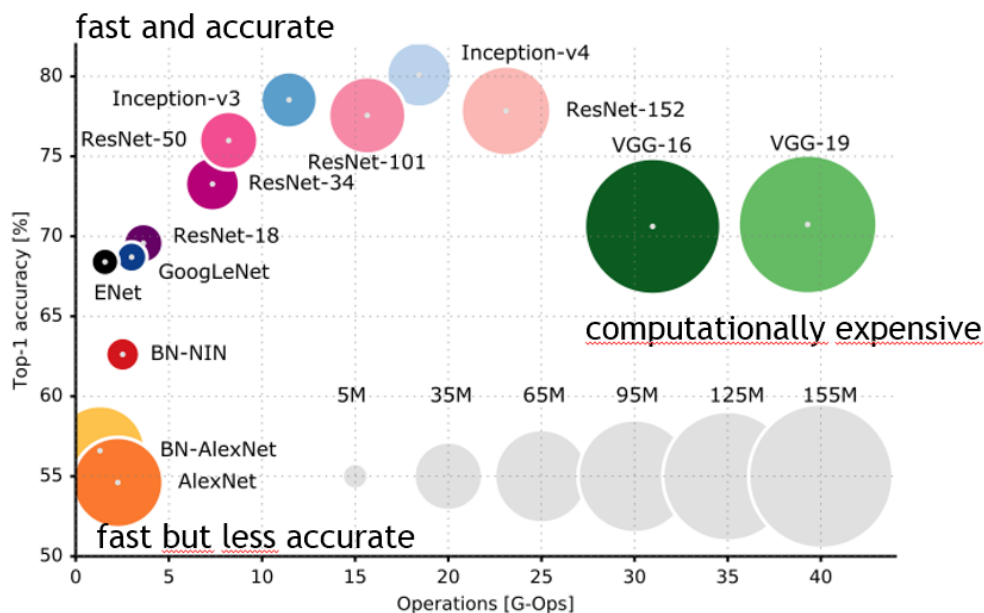


Figure 3.20: Comparison between different CNN.

A very important network to understand CNN is Alexnet. Alexnet is a CNN characterized by 25 levels that is able to recognize 1000 categories.



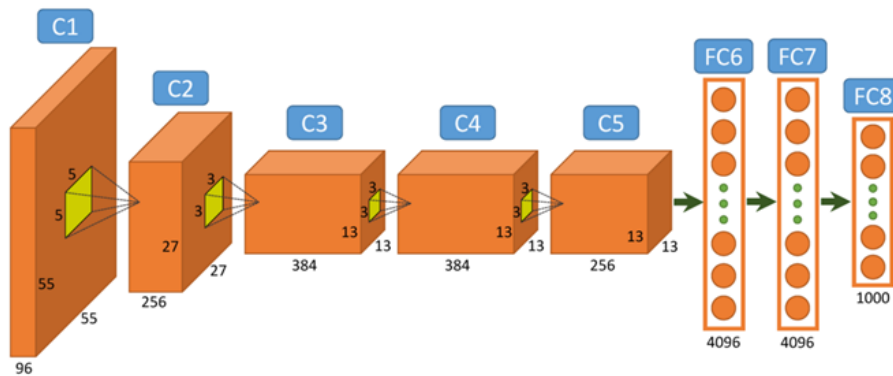


Figure 3.21: Alexnet structure

The structure of the Alexnet is composed of 25 levels, 1 input, 1 output, 5 convolution layers with RELU and Max Pooling, 3 FC layers, 1 SoftMax (tab. 3.7). The following table describes all the structure of Alexnet.

Table 3.6: 25x1 Layer array with layers:

1	'data'	Image Input	227x227x3 images with 'zerocenter' normalization
2	'conv1'	Convolution	96 11x11x3 convolutions with stride [4 4] and padding [0 0]
3	'relu1'	ReLU	ReLU
4	'norm1'	Cross Channel Normalization	cross channel normalization with 5 channels per element
5	'pool1'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0]
6	'conv2'	Convolution	256 5x5x48 convolutions with stride [1 1] and padding [2 2]
7	'relu2'	ReLU	ReLU
8	'norm2'	Cross Channel Normalization	cross channel normalization with 5 channels per element
9	'pool2'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0]
10	'conv3'	Convolution	384 3x3x256 convolutions with stride [1 1] and padding [1 1]
11	'relu3'	ReLU	ReLU
12	'conv4'	Convolution	384 3x3x192 convolutions with stride [1 1] and padding [1 1]
13	'relu4'	ReLU	ReLU
14	'conv5'	Convolution	256 3x3x192 convolutions with stride [1 1] and padding [1 1]
15	'relu5'	ReLU	ReLU
16	'pool5'	Max Pooling	3x3 max pooling with stride [2 2] and padding [0 0]
17	'fc6'	Fully Connected	4096 fully connected layer
18	'relu6'	ReLU	ReLU
19	'drop6'	Dropout	50% <i>dropout</i>
20	'fc7'	Fully Connected	4096 fully connected layer
21	'relu7'	ReLU	ReLU
22	'drop7'	Dropout	50% <i>dropout</i>
23	'fc8'	Fully Connected	1000 fully connected layer
24	'prob'	Softmax	softmax
25	'output'	Classification Output	crossentropyex with 'tench', 'goldfish', and 998 other classes

VGG-16 architecture is from VGG group from Oxford. It makes the improvement over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3X3 kernel-sized filters in cascade. With a given receptive field (the effective area size of input image on which output depends), multiple stacked smaller size kernel is better than the one with a larger size kernel because multiple non-linear layers increases the depth of the network which enables it to learn more complex features, and that too at a lower cost. While VGG achieves a phenomenal accuracy on ImageNet dataset, its deployment on even the most modest sized GPUs is a problem because of huge computational requirements, both in terms of memory and time. It becomes inefficient due to large width of

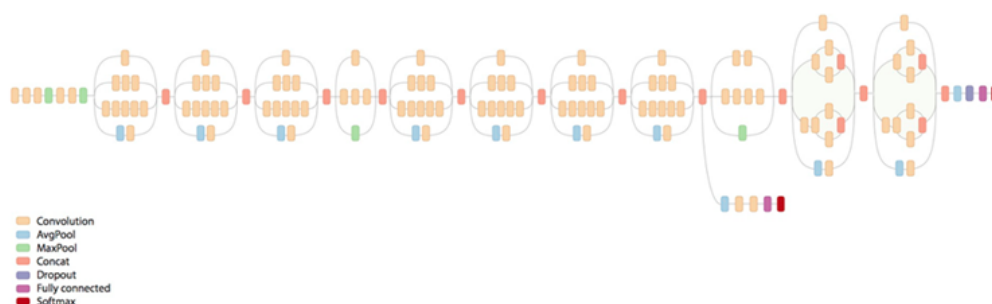


Figure 3.22: GoogLeNet structure

convolutional layers.

GoogLeNet (fig. 3.22) has a significantly more complex structure than AlexNet (fig. 3.21). The GoogLeNet builds on the idea that most of the activations in a deep network are either unnecessary (value of zero) or redundant because of correlations between them. Therefore the most efficient architecture of a deep network will have a sparse connection between the activations, which implies that all 512 output channels will not have a connection with all the 512 input channels. GoogLeNet devised a module called inception module that approximates a sparse CNN with a normal dense construction. Since only a small number of neurons are effective as mentioned earlier, the width/number of the convolutional filters of a particular kernel size is kept small. Also, it uses convolutions of different sizes to capture details at varied scales (5X5, 3X3, 1X1). Another change that GoogLeNet made, was to replace the fully-connected layers at the end with a simple global average pooling which averages out the channel values across the 2D feature map, after the last convolutional layer. This drastically reduces the total number of parameters. This can be understood from AlexNet, where FC layers contain approx. 90% of parameters. Use of a large network width and depth allows GoogLeNet to remove the FC layers without affecting the accuracy. It achieves 93.3% top-5 accuracy on ImageNet and is much faster than VGG.

ResNet is characterized by residual learning. Like GoogLeNet, it uses a global average pooling followed by the classification layer. It achieves better accuracy than VGGNet and GoogLeNet while being computationally more efficient than VGGNet. ResNet-152 achieves 95.51 top-5 accuracies. The architecture is similar to the VGGNet consisting mostly of 3X3 filters.

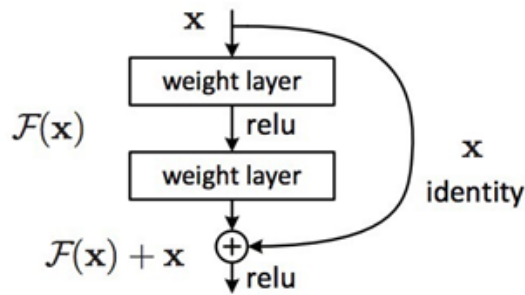


Figure 3.23: Residual block of ResNet

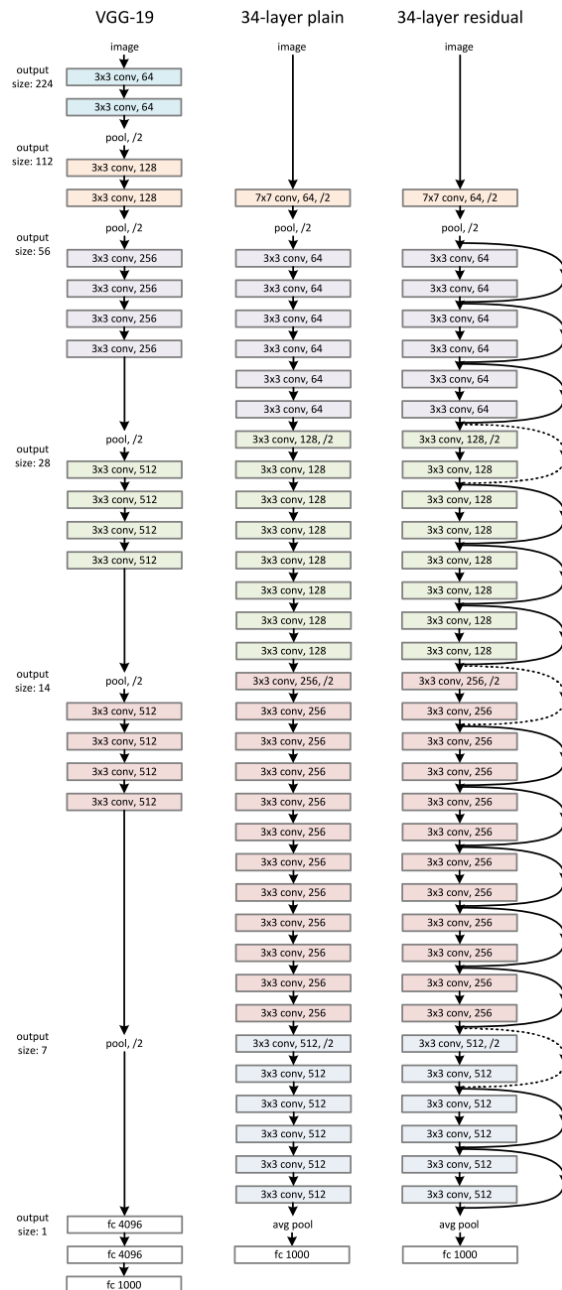


Figure 3.24: ResNet structure

### 3.5.3 Deep Learning Frameworks

The main frameworks for deep learning [107] are (ordered from the most complex to the least complex) :

1. TORCH
2. TENSORFLOW
3. CAFFE'
4. MATLAB

Torch is a computational framework with an API written in Lua that supports machine-learning algorithms. Some version of it is used by large tech companies such as Facebook and Twitter, which devote in-house teams to customizing their deep learning platforms. Lua is a multi-paradigm scripting language that was developed in Brazil in the early 1990s. Torch strengths are:

- Lots of modular pieces that are easy to combine
- Easy to write your own layer types and run on GPU
- Lots of pretrained models

Torch weakness are:

- You usually write your own training code (Less plug and play)
- No commercial support
- Spotty documentation

Google created TensorFlow a deep-learning framework written with a Python API over a C/C++ engine. Google's acknowledged goal with Tensorflow seems to be recruiting and making code of their researchers shareable, standardizing the approach to deep learning of software engineers, and creating an additional draw to Google Cloud services, on which TensorFlow is optimized. TensorFlow is not commercially supported. TensorFlow generates a computational graph and performs automatic differentiation. TensorFlow strengths are:

- Python language
- Computational graph abstraction
- Faster compile times
- TensorBoard for visualization
- Data and model parallelism

TensorFlow weakness are:

- Slower than other frameworks
- Not many pretrained models
- Computational graph is pure Python, therefore slow
- No commercial support

Caffe is a well-known and widely used machine-vision library. Caffe is a deep learning framework written in C++ developed by Berkeley AI Research (BAIR). Caffe is not intended for other deep-learning applications such as text, sound or time series data. Like other frameworks mentioned here, Caffe has chosen Python for its API. Caffe strengths are:

- Good for feedforward networks and image processing
- Good for finetuning existing networks
- Train models without writing any code
- Python interface is pretty useful

TensorFlow weakness are:

- Need to write C++ / CUDA for new GPU layers
- Not good for recurrent networks
- No commercial support

- Python interface is pretty useful

MATLAB makes deep learning easy, thanks to tools and functions for managing large data sets. It also offers specialized toolboxes for working with machine learning, neural networks, computer vision, and automated driving. With just a few lines of code, MATLAB lets you do deep learning, creating and visualizing models, and deploying models to servers and embedded devices. Matlab strenghts are:

- Perfect for testing and prototype
- Easy development
- Commercial support

Matlab weakness are:

- Very slow computation

### 3.5.4 Code and test

In this section it is described the test done using Matlab code (with pre-trained Alexnet on Imagenet) executed on VMBA15 dataset. In fact, one of the easy way to test deep model is to use directly a pre-trained network on a dataset.

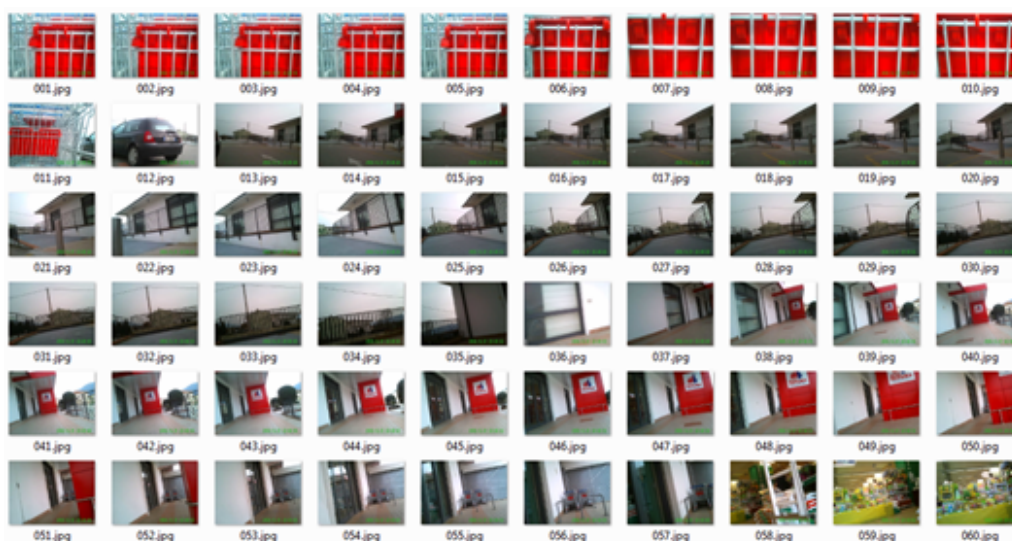
The Matlab code is very easy, and it is here reported:

```
% Access the trained model
net = alexnet
% See details of the architecture
net.Layers
% Read the image to classify
I=imread('foto_test.jpg');
label=classify(net, I)
I=imresize(I,[227,227]);
% Show the image and the classification results
figure
imshow(I)
text(10,20,char(label),'Color','white')
```

In this case we can see some misclassified frames due to the fact that Alexnet model used was trained on Imagenet and not on a retail image dataset. In fact, the shopping cart is classified as "shopping cart", the refrigerated counter as "tobacco shop", the shelf with the "grocery store" bread and the confectionery shelf for "confectionery" (fig. 3.25). We provide the sequences obtained from the run of CNN Alexnet network on the VMBA15 Dataset.



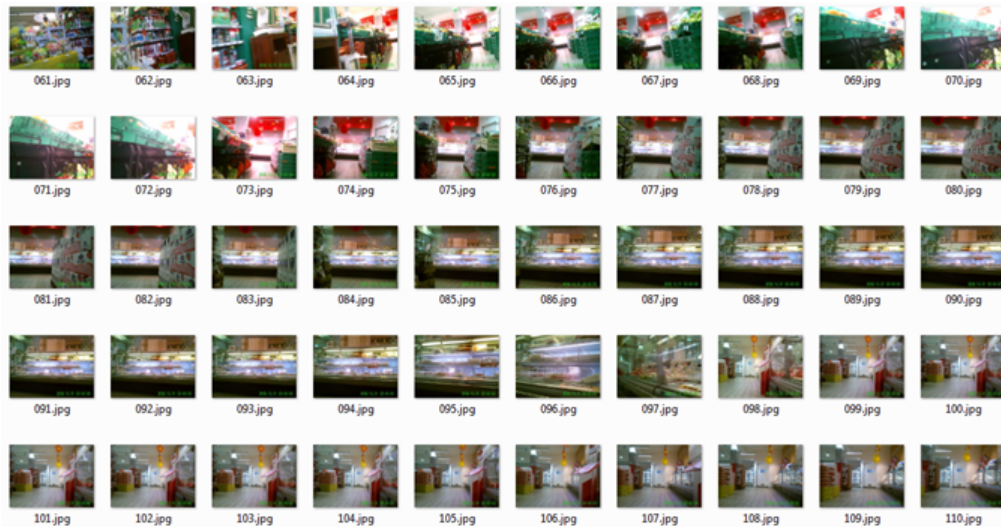
Figure 3.25: Store photos



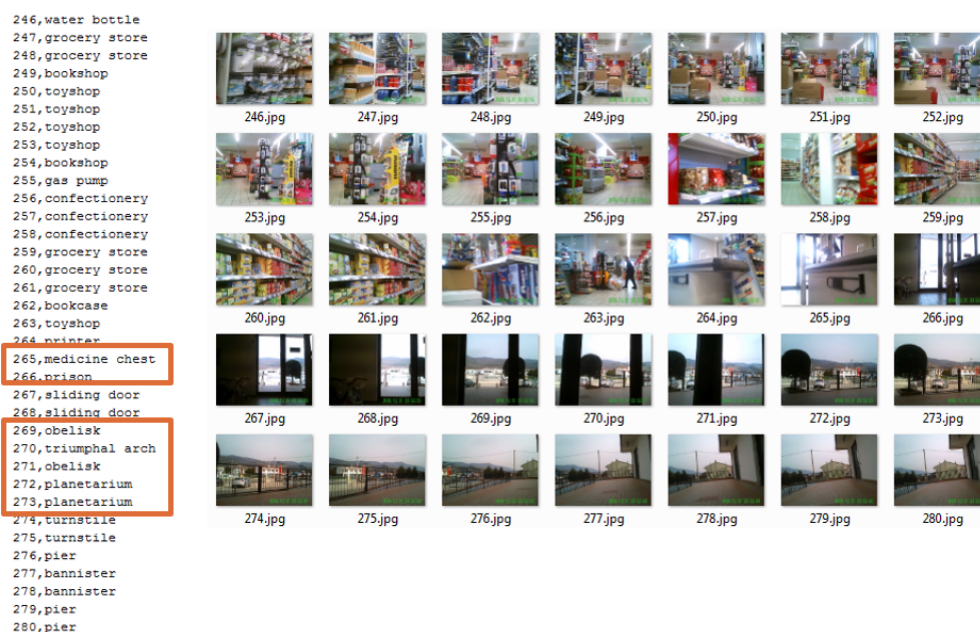
1,shopping cart - 2,shopping cart - 3,shopping cart - 4,shopping cart - 5,shopping cart - 6,paddlewheel - 7,pick 8,pick - 9,pick - 10,parallel bars - 11,shopping cart - 12,minivan - 13,patio - 14,patio - 15,pier - 16,pier 17,pier - 18,pier- 19,pier - 20,bannister - 21,bannister - 22,bannister - 23,bannister - 24,bannister - 25,bannister - 26,pier - 27,pier - 28,pier - 29,pier - 30,pier - 31,pier - 32,bannister - 33,bannister - 34,bannister - 35,bannister - 36>window screen - 37,sliding door - 38,sliding door - 39,mailbox - 40,cash machine - 41,jinrikisha - 42,moving van - 43,jinrikisha



- 44,moving van - 45,home theater - 46,cash machine - 47,cash machine - 48,cash machine - 49,binder - 50,bannister - 51,bannister - 52,sliding door - 53,pay-phone - 54,sliding door - 55,shopping cart - 56,shopping cart - 57,gas pump - 58,grocery store - 59,toyshop - 60,toyshop



61,toyshop - 62,bookshop - 63,bookshop - 64,shoe shop - 65,stage - 66,grocery store - 67,garbage truck - 68,grocery store - 69,garbage truck - 70,steam locomotive - 71,steam locomotive - 72,harvester - 73,feather boa - 74,butcher shop - 75,shopping cart - 76,confectionery - 77,paddlewheel - 78,confectionery - 79,confectionery - 80,confectionery - 81,confectionery - 82,vending machine - 83,dock -84,bakery - 85,bakery - 86,vending machine - 87,bakery - 88,bakery - 89,bakery - 90,bakery - 91,bakery - 92,bakery - 93,bakery - 94,bakery - 95,bakery - 96,bullet train - 97,grocery store - 98,refrigerator -99,refrigerator - 100,refrigerator - 101,refrigerator - 102,refrigerator - 103,refrigerator - 104,refrigerator - 105,refrigerator - 106,barbershop - 107,bookshop - 108,bookshop - 109,bookshop - 110,bookshop



We can see that the deep model provide a story about the narrative cart video, where there are some important misclassified output as obelisk, planetarium, prison, triumphal arch, that are not presented in the retail context. However, it is possible to notice that with a mapping of the classes, it is possible to use the TRANSFER LEARNING for the INDOOR / OUTDOOR classification (recoding all the Imagenet labels in indoor/outdoor classes), with good responses.

### 3.5.5 R-CNN

It is very important to notice that in retail context (that are high complexity scene) it could be useful to use R-CNN approach [109]. R-CNN is region with CNN features approach, and it applies CNN to parts of an image in order to recognize various objects in a scene.

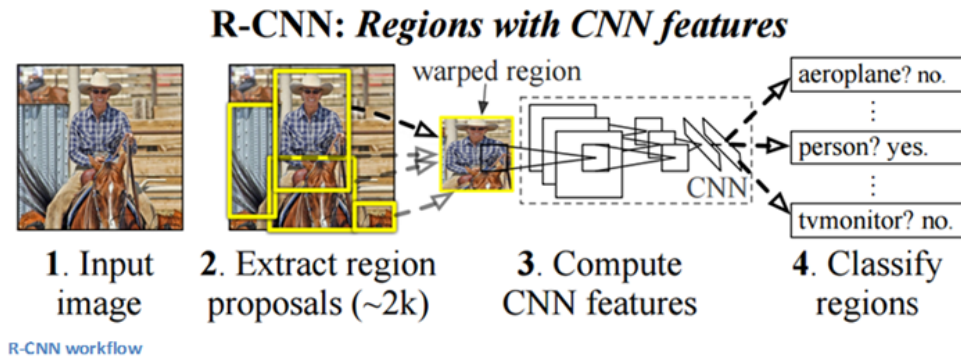


Figure 3.26: R-CNN Workflow

We can see misclassified items using CNN Alexnet (pretrained on Imagenet), on the previous scenario, also considering the region based approach in Matlab. In fact, carriage grille is classified as "street sign", the refrigerated counter as "vending machine" and the confectionary shelf for "tobacco shop" (fig. 3.27). Some error are obtained also in other images with products (fig. 3.28), where pasta composition is detected as "carpenter's kit". These results underline that it is very important to use a CNN network trained on retail images.



Figure 3.27: Store photos



Figure 3.28: Products on the shelves

The following image shows first convolutional layer weights of Alexnet CNN model.

N.B. The convolution layers are 2,6,10,12,14



Figure 3.29: First Convolutional Layer weights

### 3.5.6 Retail CNN Approach

A possible approach to obtain better results at the application level is to combine Deep Learning (CNN features) with Machine Learning (e.g., SVM). In this approach it is possible to use the Deep Learning for the extraction of features (e.g. from FC7 layer) and the Machine Learning is used for the classification useful for its application purpose (in this way it avoids being limited to the database of the used training labels only).

To perform our INDOOR / OUTDOOR classification instead of using a literature descriptor (e.g. GIST) it is possible to use for example the FC7 features extracted for each image and then classify them with an SVM. Following this method, Deep Learning (CNN) becomes a Feature Extraction tool for Machine Learning (SVM).

Table 3.7: 322x4096 single

0	1	2	3	4	5	6	7	8	9	10
1	-9.4664	-3.3576	2.2642	-8.4319	-3.7602	-4.2530	-6.8796	-3.6031	-6.0106	1.1480
2	-9.6076	-3.3343	2.1569	-8.4030	-3.8072	-4.2159	-6.8346	-3.6153	-5.9767	1.2833
3	-9.5561	-3.3343	2.3014	-8.5379	-3.7608	-4.2630	-6.8018	-3.6436	-6.1133	1.0852
4	-9.4814	-3.3631	2.1973	-8.2865	-3.7620	-4.1980	-6.1819	-3.2450	-6.1629	1.2508
5	-10.8098	-3.4016	1.7872	-8.4559	-4.1161	-4.0851	-6.0978	-3.8547	-6.0726	1.6742
6	-12.7350	-6.5178	2.6649	-11.1932	-6.5308	-5.2891	-9.8291	-9.0657	-7.8156	2.2535
7	-8.1952	-4.0098	3.9222	-9.1783	-4.5058	-4.6257	-7.3864	-4.7640	-4.9595	3.8088
8	-10.5645	-6.3131	6.0953	-12.7227	-4.1293	-6.4612	-10.3004	-5.8750	-6.4634	3.6549
9	-10.8741	-4.9534	7.0251	-12.2903	-4.1753	-4.8785	-9.6151	-6.4749	-6.4881	3.0362
10	-14.2501	-7.7967	5.2502	-13.6217	-5.5534	-4.4549	-11.3036	-7.7832	-5.6913	3.0841
11	-10.0489	-2.7909	-2.8347	-7.6427	-2.4630	0.2548	-3.8220	0.1725	-6.8434	-1.6078
12	1.5456	-0.1375	-5.8012	2.3662	-7.6919	-2.6768	-2.8995	-2.4686	-6.4381	-0.3892
13	-2.8086	-4.2513	-4.6263	2.2179	-2.2435	-2.6794	-4.1647	-2.5432	-0.7657	1.3046

In this regard, we report the matlab code sequence to perform the extraction of the features from the FC7 of the Alexnet-driven CCN on Imagenet and its classification with an SVM. In this way the deep features are classified by machine learning (SVM) approach.

```

DEEP FEATURE EXTRACTION
img = imread(foto);
net = alexnet;
net.Layers;
I= imresize(img, [227 227]);
labeltotali = classify(net, I)
fprintf(foggetti, '\%s,\%i,\%s\n', foto, i, labeltotali);
layer = 'fc7';
trainingFeatures = activations(net, I, layer);
descriptor(i,:) = trainingFeatures;
SVM CLASSIFICATION
A LABEL TRAINING
B DATA MATRIX TRAINING
C LABEL VALIDATION

```



## D DATA MATRIX VALIDATION

```

model = svmtrain(A, B, '-c 2 -t 1 -d 3 -g 0.5 -e 0.05 ');
model = svmtrain(A, B, '-c 1 -g 2 ');
y=svmpredict(C,D,model)

```

To summarize this synthesis about deep learning, we can say that:

- Deep Learning is a revolution of Machine Learning;
- Deep Learning can replace Machine Learning in some scenarios (e.g. object recognition, scene detection) but performs on the recognition of up to 1000 categories;
- The union of Deep Learning (CNN features) with Machine Learning (SVM, Neural Networks) is a very interesting solution to perform specific tasks (e.g. recognition of specific areas of the sales point, indoor / outdoor, picking / releasing, gender recognition , emotion analysis);
- Deep Learning applied to Thermal images? It could be a new experimentation frontier.

### 3.6 VMBA with 14 classes

Then, an evolution of the previous approach on VMBA has been to extend the classes from 8 to 14 with the introduction of the concept of Deep Features [4].

We introduce the problem of Visual Market Basket Analysis (VMBA) considering three different high-level behavioural categories related to the customers carrying the shopping cart : action (i.e., *stop vs moving*), location (i.e., *indoor vs outdoor*), and scene context (i.e., *cash desk, retail, pasta, fruit, gastronomy, parking, road*). Each of the three considered high-level categories provides a meaningful source of information for the VMBA problem. For instance, knowing that a cart is stopping, rather than moving in a particular area of the store, can provide insights on whether the customer is experiencing difficulties in locating the desired products. Likewise, being able to understand when the cart is inside or outside the store and, ultimately, in which part of the store it is located, allows to obtain real-time information on the distribution of customers in the store.

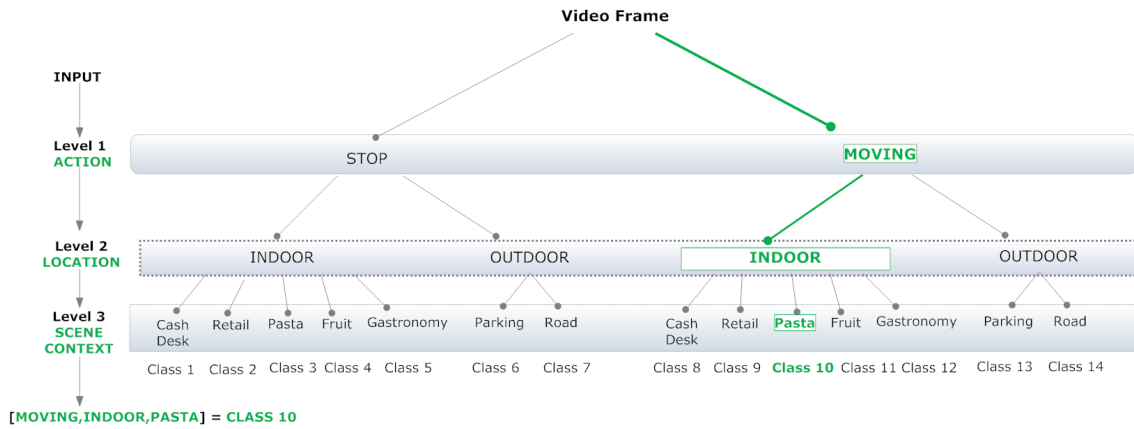


Figure 3.30: Visual Market Basket Analysis (VMBA) behavioural classes organized in a hierarchy. Best seen in digital version.

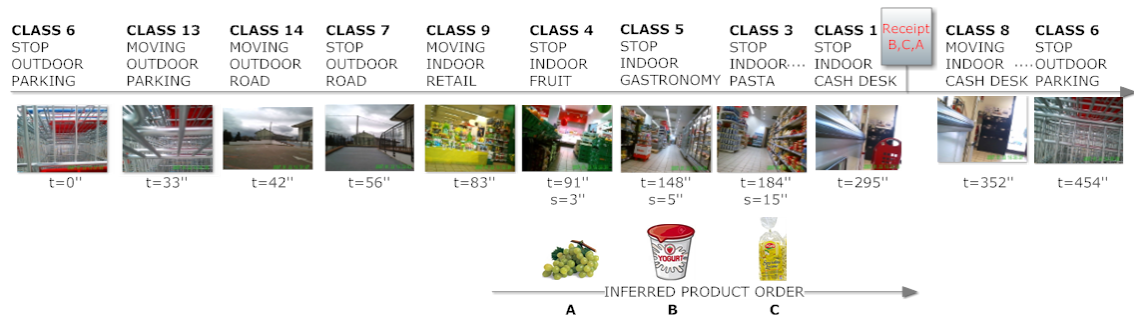


Figure 3.31: VMBA timeline temporally segmented considering the 14 classes.  $t$  denotes the time, whereas  $s$  denotes the stopping time.

We propose to organize these categories hierarchically, as shown in the tree depicted in Figure 3.30. Each of the paths from the root to a leaf, identifies one of 14 different behavioural classes. Given a frame of the video acquired by the camera mounted on the cart, we aim to infer a triplet identifying the path of the tree corresponding to the observed behaviour (e.g., [MOVING, INDOOR, PASTA] in Figure 3.30). The classification of each frame with respect to the proposed 14 classes allows to analyse the customer’s behaviour and can be also useful to understand if there are problems on strategic issues to be fixed by the management.

Figure 3.31 shows some examples of egocentric images acquired with a shopping cart together with the temporal segmentation with respect to the 14 considered classes. From the segmented egocentric video it is possible to understand how much time customers spend at the cash desk by considering all the frames classified



with the triplets [STOP, INDOOR, CASH DESK] and [MOVING, INDOOR, CASH DESK]. This may be useful, for example, to plan the opening of more cash desks in order to provide a better service to the customers. By analysing the inferred triplets, it is also possible to understand if there are carts outside the cart parking spaces in order to take appropriate actions (e.g., if a given cart is associated to the triplet [STOP, OUTDOOR, ROAD] for long time). Combining the customers' receipts with information arising from temporally segmented videos and algorithms for Visual Market Basket Analysis' re-localization [81] could help infer the order according to which products have been taken during the shopping (Figure 3.31), hence increasing the amount of information usually exploited by the classic Market Basket Analysis' algorithms [66]. This opens also new research perspectives in the context of egocentric vision.

### 3.6.1 Proposed Method

Our aim is to temporally segment the acquired egocentric videos into chapters. To this aim, each frame should be automatically labelled according to one of the considered 14 behavioural classes. In this context, a chapter is a set of consecutive frames which present the same behaviour, e.g., a sequence of frames with the same label [MOVING, INDOOR, PASTA]. We propose to explicitly consider the hierarchy illustrated in Figure 3.30 to classify each frame. The first level of the hierarchy is related to the action of moving or stopping, which reflect the basic actions of a customer in the retail store. The second layer of the hierarchy identifies the high level locations where the customer is moving, i.e., indoor or outdoor. The third level considers the scene-context in which the user is located during the shopping. The final classification with respect to the 14 classes can hence be obtained by considering the three classification problems and predicting a triplet [*Action*, *Location*, *Scene Context*]. To perform each of the three level classification tasks, two main components are needed: a suitable representation and a classification algorithm. In the following, we describe the representations used for the three different levels of the hierarchy in Figure 3.30, as well as the classification approaches exploited in the proposed study.

The first layer of the hierarchy analyses the user behaviour from the point of view of the motion of the cart. To this aim we have considered the approach shown

in previous section.

The second layer of the hierarchy aims to identify the general location of the user: *indoor* vs *outdoor*. At this level we consider both visual and audio features. Concerning visual features, we have considered , in addition to the popular GIST descriptor proposed in [89], Deep Features [87]. The GIST is able to encode the visual scene with a feature vector of 512 components. The Deep Features are obtained by the FC7 layer of Alexnet CNN architecture trained on ImageNet [88], a feature vector of 4096 components. We also consider the MFCC feature representation after visual inspection of the audio waveform. Indeed, the waveform is more pronounced in the outdoor environment than in the indoor location. In our experiments we consider audio and visual features both independently and in combination.

The third layer of the hierarchy is related to the analysis of the context in which the shopping cart is located. Among the considered seven classes, i.e., *cash desk*, *retail*, *pasta*, *gastronomy*, *fruit*, *parking and road*, the first five are related to the indoor environment, whereas the other two are related to the outdoor location. Also for this level of description we have considered the GIST features, given their property of capturing the “shape of the scene” for context discrimination [89]. As for the previous layer, we have also tested deep features [87].

After representing a frame of the egocentric video as described in previous sections, a classifier has to be used in order to infer one of the 14 behavioural classes. To benchmark the VMBA problem and assess the impact of the proposed hierarchical organization shown in Figure 3.30, the following different classification modalities have been considered:

- combination of the results obtained by three different SVM classifiers trained to perform classification with respect to each of the levels of the tree shown in Figure 3.30;
- A single SVM classifier trained to discriminate among the 14 considered classes;
- a Direct Acyclic Graph SVM learning architecture (DAGSVM) [platt] which reflects the hierarchy in Figure 3.30 on each node.

We acquired a dataset of 15 different egocentric videos during real shopping sessions in a retail store. The dataset is referred to as VMBA15. All videos have been



Figure 3.32: Some frames extracted from the egocentric videos of the proposed VMBA15 dataset. Each frame is related to one of the 14 considered classes. Images are presented in the same order as Figure 3.30 from left to right and from top to bottom. Notice that some classes are characterized by similar visual content but different actions (frames in the top row are related to the “stop” categories, whereas frames in the second row are related to the “moving” categories).

Table 3.8: Number of samples per context label for each egocentric video.

VID	PARK	ROAD	CASH DESK	RETAIL	GASTRONOMY	FRUIT	PASTA	TOTAL
1	23	89	8	117	43	13	29	322
2	10	84	9	209	16	32	30	390
3	10	96	12	163	36	27	23	367
4	13	106	10	156	20	35	3	343
5	9	107	10	159	24	32	4	345
6	39	93	128	123	4	6	35	428
7	19	119	7	81	3	4	27	241
8	20	75	14	210	14	7	33	373
9	22	85	160	646	5	7	73	998
10	13	75	7	48	48	14	5	210
11	41	89	52	355	28	19	100	684
12	27	104	26	376	38	37	39	647
13	51	137	63	129	34	12	79	495
14	25	46	53	832	4	6	32	998
15	6	73	84	761	3	61	10	998

acquired using a narrative cam<sup>1</sup> mounted in the front part of the shopping cart. Figure 3.32 shows some samples extracted from the dataset. The duration of each video is comprised between 3 and 20 minutes and has a resolution of 640x480 pixels. Audio has been also recorded since it can be useful to discriminate indoor vs outdoor environments. Each video has been manually labelled at 1 fps according to the 14 different behavioural classes arising from the possible paths root-leave of the hierarchy shown in Figure 3.30. This implies labelling each frame according to action (i.e., “stop” or “moving”), location (i.e., “indoor” or “outdoor”) and scene context (i.e., “cash desk”, “retail”, “pasta”, “fruit”, “gastronomy”, “parking” and “road”). The dataset contains a total of 7839 labelled samples. Table 3.8 reports the number of samples for each of the 8 scenes contained in the egocentric videos, while Table 3.9 reports the number of samples for each of the 14 considered class. Each class  $C_i$  is related to a path in the tree of Figure 3.30. Specifically: C1 [STOP INDOOR Cash Desk], C2 [STOP INDOOR Retail], C3 [STOP INDOOR Cash Desk], C4 [STOP INDOOR Fruit], C5 [STOP INDOOR Gastronomy], C6 [STOP OUTDOOR Parking], C7 [STOP OUTDOOR Road], C8 [MOVING INDOOR Cash Desk], C9 [MOVING INDOOR Retail], C10 [MOVING INDOOR Cash Desk], C11 [MOVING INDOOR Fruit], C12 [MOVING INDOOR Gastronomy], C13 [MOVING OUTDOOR Parking], C14 [MOVING OUTDOOR Road]. The labelled dataset is available for research purposes at <http://iplab.dmi.unict.it/vmba15>.

### 3.6.2 Experimental Settings and Results

The experiments have been performed splitting the dataset randomly in three parts, each composed of five egocentric videos. All experiments are repeated three times using two parts (10 videos) for training and one part (5 videos) for testing. The reported results are obtained by averaging over the three runs. We begin our analysis comparing the investigated representations when used to solve independently one of the three considered classification tasks. This test has been performed by exploiting a SVM classifier with an RBF kernel for each level separately. This experiment is useful to determine the best representation (or a combination of them) to be employed at each level of the hierarchy.

---

<sup>1</sup>[www.vehomuvi.com](http://www.vehomuvi.com)

Table 3.9: Number of samples per class for each video. Each class  $C_i$  is related to a path in the tree of Figure 3.30. See text for the details.

VID	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
1	0	3	2	6	18	6	0	8	114	27	7	25	17	89
2	0	15	4	2	0	0	0	9	194	26	30	16	10	84
3	0	4	4	3	12	0	0	12	159	19	24	24	10	96
4	0	15	3	0	5	0	0	10	141	0	35	15	13	106
5	0	12	4	2	4	0	0	10	147	0	30	20	9	107
6	69	20	28	3	2	4	2	59	103	7	3	2	35	91
7	0	11	13	0	0	3	0	7	70	14	4	3	16	119
8	6	20	7	2	4	12	0	8	190	26	5	10	8	75
9	142	150	8	2	2	13	0	18	496	65	5	3	9	85
10	0	5	3	2	2	0	0	7	43	2	12	46	13	75
11	42	79	31	7	2	25	0	10	276	69	12	26	16	89
12	0	24	22	4	3	4	0	26	352	17	33	35	23	104
13	56	41	57	6	10	23	4	7	88	22	6	24	28	133
14	50	380	7	2	2	14	0	3	452	25	4	2	11	46
15	81	482	0	18	1	3	27	3	279	10	43	2	3	46

Table 3.10 reports the results of the *stop vs moving* classification (i.e., first level). Both audio and visual features obtain good performance. However, the optical flow features outperform audio features with a margin of about 5%. The combination of audio and optical flow features allows to increase accuracy to the value of 94.50%. The obtained results point out that the combination of audio and visual features are the best suited for the first level.

Table 3.10: Results for stop vs moving classification. Per-row best results are reported in bold numbers.

	FLOW	MFCC	COMBINED
Accuracy%	92.50	87.04	<b>94.50</b>
TP RATE%	73.03	61.54	<b>84.76</b>
TN RATE%	<b>99.18</b>	95.21	97.65
FP RATE%	<b>0.82</b>	4.79	2.35
FN RATE%	26.97	38.46	<b>15.24</b>

Table 3.11 reports the results of the *indoor vs outdoor* classification (second

layer). In this case, deep features outperform both GIST and audio features with a good margin obtaining an accuracy of 97.26%. Hence, we employ the Deep Features descriptor alone in the second level.

Table 3.11: Results for indoor vs outdoor classification. Per-row best results are reported in bold numbers.

	GIST	MFCC	DEEP
Accuracy%	95.79	88.00	<b>97.26</b>
TP RATE%	89.3	49.51	<b>94.34</b>
TN RATE%	97.8	97.66	<b>98.33</b>
FP RATE%	<b>2.20</b>	2.34	1.66
FN RATE%	10.7	50.49	<b>5.64</b>

For the *scene-context* classification (third level), we obtain an accuracy of 85.05% using the GIST descriptor and 90.34% with the use of Deep Features descriptor. Note that, in this case, a multi-class SVM with RBF kernel has been trained to discriminate between the seven possible scene contexts without using priors given by the previous level in the hierarchy (i.e., indoor vs outdoor). We also report the confusion matrices with respect to the considered seven scene contexts in Table 3.12 and Table 3.13. When using GIST descriptors, the main classification errors are related to the confusion between the “parking”, “road” and “retail” classes (first row of Table 3.12). The confusion between parking and retail classes is probably due to the encoding of the scene information by the GIST descriptor. Indeed, when the cart is in the parking space, the scene is mainly composed by vertical and horizontal edges which can be confused with the vertical and horizontal edges of some scenes in the retail (see Figure 3.9). As it is shown in Table 3.13, As it will be discussed in the experiments, explicitly enforcing a hierarchy using a DAGSVM classifier helps reducing ambiguities by enforcing a prior on the main location (*indoor vs outdoor*). As it is shown in Table 3.13, results improve using Deep Features descriptor. For instance, the classification errors for the “parking” and “road” classes are reduced. Misclassification between the “retail” classes are mainly due to the visual similarity of different retail sectors in the frames. Other classification problems

are due to strong occlusions caused by people in the scene as shown in the examples in Figure 3.10.

Table 3.12: Confusion matrix for scene context classification with the GIST descriptor.

			PREDICTED				
	PARKING	ROAD	CASH DESK	RETAIL	GASTRONOMY	FRUIT	PASTA
PARKING	<b>54.25</b>	22.88	1.96	20.26	0.00	0.00	0.65
ROAD	7.25	<b>84.46</b>	6.57	0.99	0.54	0.00	0.19
CASH DESK	1.23	13.35	<b>78.67</b>	6.75	0.00	0.00	0.00
RETAIL	1.82	0.02	1.19	<b>92.46</b>	1.72	0.80	1.99
GASTRONOMY	0.00	3.89	0.00	27.16	<b>68.95</b>	0.00	0.00
FRUIT	0.00	0.00	0.00	37.10	0.00	<b>62.90</b>	0.00
PASTA	0.56	0.56	0.00	28.65	0.00	0.00	<b>70.23</b>

Table 3.13: Confusion matrix for scene context classification with Deep Features.

			PREDICTED				
	PARKING	ROAD	CASH DESK	RETAIL	GASTRONOMY	FRUIT	PASTA
PARKING	<b>85.51</b>	1.87	0.0	7.94	0.0	0.0	0.0
ROAD	0.38	<b>94.08</b>	2.86	1.34	0.0	0.0	0.0
CASH DESK	0.0	12.59	<b>68.53</b>	17.83	0.35	0.0	0.0
RETAIL	0.0	0.12	0.32	<b>98.2</b>	0.28	0.2	0.84
GASTRONOMY	0.0	2.6	0.0	38.31	<b>59.09</b>	0.0	0.0
FRUIT	0.0	0.67	0.0	35.57	0.00	<b>61.74</b>	2.01
PASTA	0.0	0.0	0.0	35.78	0.0	0.0	<b>64.22</b>

### 3.6.3 Overall Classification

The experiments presented in the previous sections pointed out that the best features to be employed in the first level of the hierarchy are the combination of MFCC and FLOW, whereas the DEEP Features descriptor can be employed in the second and third levels. Since the main goal is the classification with respect to the 14 possible triplets corresponding to the leaves of the tree in Figure 3.30, after selecting the features for the three levels independently, we have compared the three different classification modalities discussed in Section 3.6.1, namely: 1) combination of the results of three different SVM classifiers trained to address classification at each of the three considered levels; 2) a multiclass SVM classifier which discards the proposed

hierarchy and classifies samples into the 14 possible classes; a DAGSVM [platt] classifier which reflects the hierarchy proposed in Figure 3.30 to perform classification.

The results of the three different approaches are reported in Table 3.14. Best results are obtained by the DAGSVM approach, which obtains an accuracy of 87.71%. It is worth noting that the simple concatenation of MFCC features with the FLOW and DEEP descriptors does not allow the multi-class SVM to reach the best accuracy (67.50%). The proposed DAGSVM approach also outperforms the concatenation of three different classifiers trained separately at each level (80.10%). Some visual examples for the assessment of the output given by the proposed DAGSVM-based approach are available in Figure 3.11.

Table 3.14: Results of the three considered classification approaches.

	Combination	Multi-Class SVM	DAGSVM
Accuracy%	80.10	67.50	87.71

Table 3.15 reports the confusion matrix related to the performances of the DAGSVM. Class C1 [STOP, INDOOR, CASH DESK] obtains a True Positive Rate (TPR) of 63,4%, with the largest misclassification obtained on the C2 class [STOP, INDOOR, RETAIL]. The best performances are obtained with the C9 class [MOVING, INDOOR, RETAIL]. C3 class [STOP, INDOOR, PASTA] is misclassified with the generic C2 class that represents the generic retail environment. Similar performances are obtained for the classes C4 and C5 related to the fruit and gastronomy departments. For the class C6 [STOP, OUTDOOR, PARKING], the performance obtained is of 83,10%, due to the misclassification with retail scenarios (the cart grid is mistaken for windows and shelves). Class C7 [STOP, OUTDOOR, ROAD] is classified with 93.8% of accuracy. Considering the results, the proposed approach can be used for rough localization of customers in the sale point.



Table 3.15: Confusion Matrix of the DAGSVM approach.

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	<b>63.4</b>	14.5	0	0	0.21	0	9.8	4.9	3.4	0	0	0.1	0	3.7
C2	0.25	<b>93.6</b>	0.7	0.1	0.23	0	0.1	0.1	4.6	0.2	0.1	0.1	0	0.1
C3	0	34.5	<b>61.2</b>	0	0	0	0	0	1.3	3.1	0	0	0	0
C4	0	32.1	1.8	<b>62.9</b>	0	0	0.62	0	1.4	0.1	0.9	0	0	0.1
C5	0	34.2	0	0	<b>59.7</b>	0	2.1	0	2.1	0	0	1.2	0	0.7
C6	0	8.5	0	0	0	<b>83.1</b>	2.4	0	2.8	0.4	0	0	2.5	0.2
C7	1.7	1.1	0	0	0	0.4	<b>93.8</b>	0.3	0.1	0	0	0	0	2.6
C8	7.6	2.1	0	0	0.1	0	1.9	<b>61.2</b>	14.6	0	0	0.3	0	12.2
C9	0.1	3.1	0.1	0.1	0	0	0	0.2	<b>95.4</b>	0.6	0.1	0.3	0	0.1
C10	0	6.1	3.2	0	0	0	0	0	30.2	<b>60.3</b>	0	0	0	0
C11	0	3.2	0.3	1.7	0	0	0.1	0	32.7	1.9	<b>59.6</b>	0	0	0.4
C12	0	0.9	0	0	1.3	0	0.5	0	37.5	0	0	<b>58.1</b>	0	2.1
C13	0	2.3	0	0	0	3.6	0.3	0	9.2	0	0	0	<b>82.5</b>	1.9
C14	0.5	0.8	0	0	0	0	2.3	2.1	1.9	0	1.3	0	0.3	<b>90.8</b>

To demonstrate the system, in Table 3.16, we report some sample predictions at different times in a retail store. The system analyses 15 input videos and classifies them at different instants to infer their behaviour. Specifically, each row reports the number of predicted charts belonging to a given category. Ground truth predictions are reported in parenthesis. A video demo of the proposed method is also available at our web page: <http://iplab.dmi.unict.it/vmba15>.

### 3.7 Conclusion

This chapter has introduced the problem of Visual Market Basket Analysis (VMBA). To set the first VMBA challenge, a new egocentric video dataset (VBMA15) has been acquired in a retail store with cameras mounted on shopping carts. The VBMA15 dataset has been labelled considering 14 different classes arising from a hierarchical organization of *Actions*, *Location* and *Scene Contexts*. A first benchmark has been performed considering classic representations and classification modalities. Experiments pointed out that audio, motion and global visual features are all useful in the VMBA application domain when coupled with a Direct Acyclic Graph based SVM leaning architecture.

Future works will investigate the design of a framework based on deep learning trainable in an end-to-end fashion to address Market Basket Analysis from Egocentric Videos processing and fusing information coming from the three levels. Moreover, the analysis will be extended to data collected in more retail stores.

Table 3.16: Number of Shopping Charts Predicted for Each Class at a Given Time. Ground Truth Predictions are Reported in Parenthesis. Reported times are in MM:SS format.

Time	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
00:00	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (14)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
03:10	0 (0)	2 (2)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6 (5)	5 (6)	0 (0)	0 (0)	0 (0)	1 (1)
06:30	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (3)	0 (0)	0 (0)	0 (0)	1 (2)	2 (2)
09:50	0 (1)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	3 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
13:10	0 (1)	3 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
16:30	1 (2)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)

## Chapter 4

# FUTURE DEVELOPMENT

### 4.1 Introduction

This chapter describes some experimental tests carried out in relation to the issues such as DOOH and Visual Market Basket Analysis addressed in the two previous chapters. In particular, with regard to DOOH a starter experiment is described regarding the recognition of Gender and Age on a dataset acquired in the infrared band for dermatological purposes. In the context of Visual Market Basket Analysis, on the other hand, a correlation approach is proposed between the video sequences acquired and the list of products tracked in the receipt (associated with a loyalty card). Moreover, through the video acquisition inside the shopping cart it is possible to analyze the picking and releasing activities of the products, being able to carry out a precise description of all the purchasing behavior carried out by the user, therefore considering in a marketing key which products they are winning compared to others, and determining in perspective functional alerts for a better management of the layout of the point of sale.

### 4.2 DOOH on infrared spectrum

Among the various experimentation activities conducted there was also that relating to testing the algorithms of Face extraction, Gender recognition and Age estimation relative to images acquired in the infrared spectrum. These acquisitions are experimented in the dermatological field for the diagnosis and control of skin diseases, however, the recognition of sex has led to performances inferior to those obtained

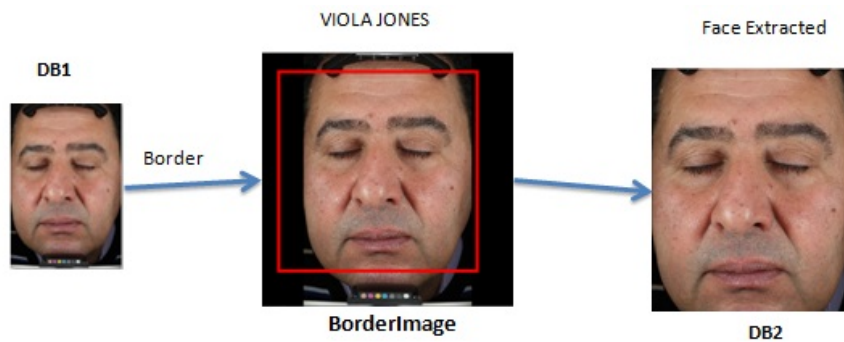


Figure 4.1: Extraction of the face.

by analysis of the RGB band. The activity on the dermatology dataset concerned the following activities:

1. Extraction of the face from the individual datasets (eliminating the boundary information)
2. Test of the Gender Recognition on the normal band
3. Test of the Gender Recognition on the fourth band (near infrared)
4. Age Estimation test on the normal band
5. Age Estimation test on the fourth band (near infrared)

### 4.2.1 Face extraction

In order to achieve the recognition of sex and age on the image dataset considered, it was necessary to carry out the extraction of the faces, eliminating all the environmental information related to the device used (base of shooting of the device), in order not to introduce external factors of disturb in the recognition of sex / age. The first activity carried out was therefore the application of the Face detection algorithm: giving an image input containing the face with other factors of disturb, we move to a new one concerning only the face. The Face detection algorithm appropriately calibrated for the considered dataset worked 100% on the entire data set considered. It was therefore possible to obtain a new dataset with only the faces.

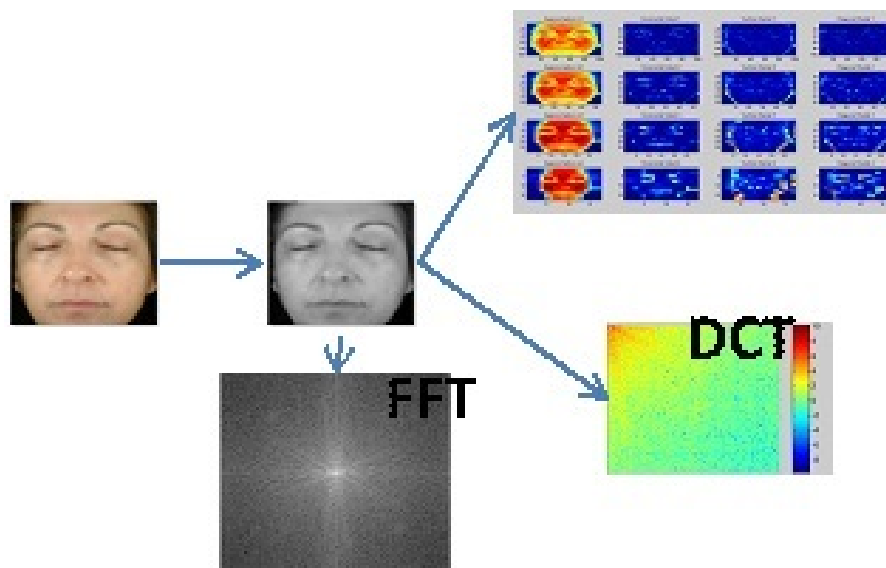


Figure 4.2: Features and Classifiers

## 4.2.2 Gender Recognition

In order to achieve the recognition of sex and age on the image dataset considered, it was necessary to carry out the extraction of the faces, eliminating all the environmental information related to the device used (base of shooting of the device), in order not to introduce external factors of disturb in the recognition of sex / age. The first activity carried out was therefore the application of the Face detection algorithm: giving an image input containing the face with other factors of disturb, we move to a new one concerning only the face. The Face detection algorithm appropriately calibrated for the considered dataset worked 100% on the entire data set considered. It was therefore possible to obtain a new dataset with only the faces.

## 4.2.3 Infrared Analysis

The third activity carried out was the test of the "gender recognition" algorithm on the fourth band, that of "near infrared" [90]. This type of analysis was performed in order to find out if in a different band from the normal one the images contain useful information for the task of the recognition of the sex. The recognition performed using (DWT + DCT + FFT) with AdaBoost, used for the previous experiment, produced 91% recognition performances. The performances are therefore slightly

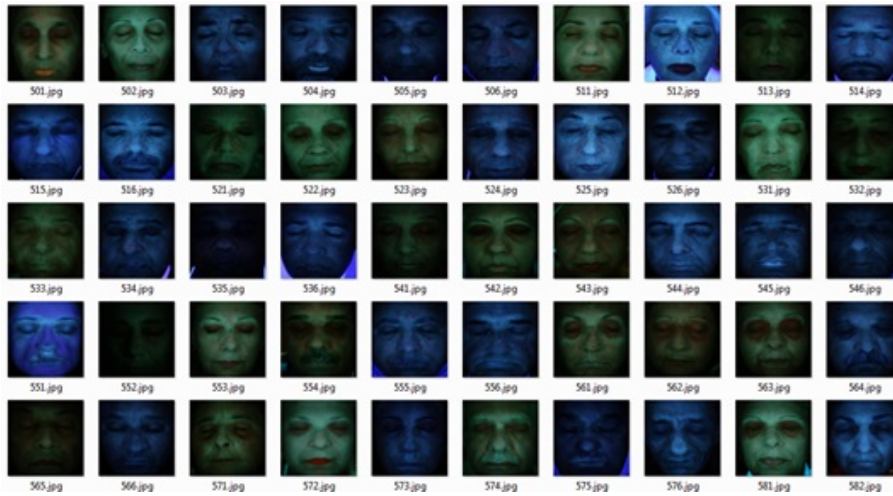


Figure 4.3: Extraction of the face.

lower than those obtained in the normal band. The fourth activity carried out was the test of the "age estimation" algorithm on the normal band. Age recognition is a problem addressed in literature [25] with methods such as Gabor, LBP and CLBP filters using SVM classifier. The results of these approaches on generic datasets do not reach the same performances as the recognition of sex (we are on an order of 70-90% accuracy), by virtue of the misclassification existing between adjacent age classes. A Training set (2/3 of the dataset) and Test set (1/3 of the dataset) subdivided into 4 bins (classes) were then created to conduct the age recognition on the dataset considered: 30-39, 40-49, 50-59, 60-70. In the division of the images between Training set and Test set, a fair division of male and female images was taken into account. The recognition performed using features (Gabor or LBP or combined filters) with SVM classifier produced 58% of the result. Unfortunately, this result is lower than in the literature (between 70% and 90%). This may be due to the low number of images that make up the training set or a misclassification present in the labeling of the dataset (we found 1 image that was wrongly labeled with 43 years instead of 31 years). Reducing the number of classes (from 4 bins to 3 bins: 30-39, 40-55, 56-70) is achieved at 73.17% of performance, since the major misclassification concerns the intermediate classes (40-55). The image of the underlying subject has been labeled with the label "43" years, when in reality the subject is thirty. It is an example of labeling error in the dataset that can lead to misclassification.

BGENERNO	COD. CH. CO.	DATA VENDITA	COD. CLIENTE	CART.	CART.	VASTI	DATA	IMPORTO	B. C.	DESC. B. C.	
16000005068116	6100	09/09/2016			115274	2261022000000		1.002	2.27	S	Vendita in BaseLine senza Premiati
16000005659993	6100	04/10/2016			137368	8003435005123		2.000	4.88	O	Vendita in Offerta senza Premiati
1600005463628	6100	26/09/2016	0000401312003817		134591	8002085001264		1.000	1.00	P	Vendita in Offerta con Premiati
1600004720808	6100	25/08/2016	0000401312023457		10651	8002580010181		1.000	.95	P	Vendita in Offerta con Premiati
1600006492403	6100	05/11/2016	0000401310047185		21816	2155120000000		.110	1.53	P	Vendita in Offerta con Premiati
1600007647718	6100	22/12/2016	0000401310049947		114758	8009115011024		1.000	2.29	T	Vendita in BaseLine con Premiati
1600004360896	6100	10/08/2016	0000401800056561		136095	8002910041991		1.000	.99	T	Vendita in BaseLine con Premiati
1600004523335	6100	18/08/2016	0000401312055908		112136	2188880000000		1.888	14.33	T	Vendita in BaseLine con Premiati
1600005213447	6100	15/09/2016	0000401800051559		130331	8051084002697		0	-4.00	P	Vendita in Offerta con Premiati
1600005383870	6100	01/10/2016	0000401312103357		54809	2522780000000		.642	4.43	P	Vendita in Offerta con Premiati

Figure 4.4: MBA

The fifth activity carried out was the test of the "age estimation" algorithm on the fourth band, that is the "near infrared" one. In a similar way to what has been done on "gender recognition", this type of analysis has been carried out in order to find out if in a different band from the normal one the images contain higher or lower information for the age recognition task. The recognition performed with the same algorithm used for the previous experiment produced 43% of performances. Unfortunately this result is lower than in the literature. Even if the normal band is combined with the 4th band, the recognition results do not improve. The activity carried out showed how the recognition of the sex dealt with by the algorithms present in the literature reaches remarkable results (over 90%). Different discourse concerns the recognition of age, as the performances obtained are much lower than those reported in the literature, also due to the wrong labeling of the images. The analysis carried out on the bands different from the normal ones did not bring any additional feedback for both the recognition tasks performed.

### 4.3 VMBA Evolution

In order to carry out a timely narration of VMBA, a dataset of loyalty cards and purchases was extracted, characterized by 1650984 of transactions relating to 4129 loyalty cards during the year on a point of sale (the same as the VMBA). These transactions were then crossed with those of the acquired videos (VMBA15) in order to faithfully reconstruct the route taken and the purchases made. As a result, the database of loyalty cards, in addition to serving to extract know-how in terms of traditional Market Basket Analysis, can be used in the VMBA framework.

It is therefore possible, following this KB, to detect the paths of the acquisitions (associated by the loyalty card), in order to reconstruct the order of picking / releasing. An evolution of this approach is to define a multiple challenge characterized by behavior analysis (DOOH methods), context analysis (VMBA methods), adding

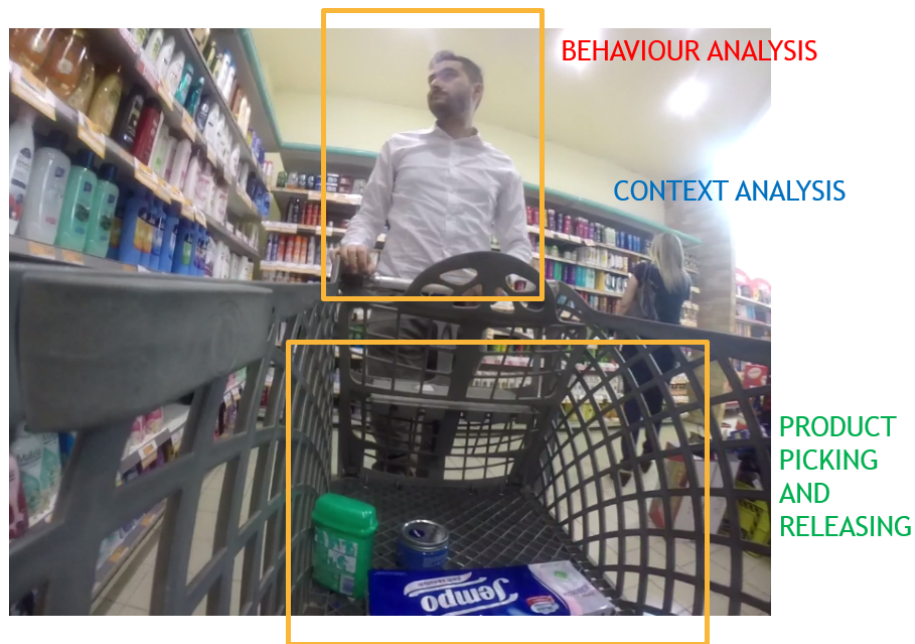


Figure 4.5: 3 ways approach

product picking and releasing. In fact, this picking / releasing information is important to better reconstruct the consumer shopping experience, which may have changed its mind when choosing a product.

To make this acquisition it is essential to insert a further camera on board the cart. For this experiment, acquisitions were made with side cameras at the top and central ones at the bottom of the cart.





Figure 4.6: Camera and cart

In the following photographs it is possible to view the picking and releasing activities performed by the side camera. The side camera at the top has an incomplete view of the carriage, however it is not affected by occlusions. The central camera of the trolley is strongly affected by occlusions due to the superposition of objects in front of the acquisition field. Among the future developments it is therefore useful to carry out the acquisition of a dataset to carry out the picking and realizing of the products to be crossed with the activities of VMBA and analysis of the associated receipt (MBA).



Figure 4.7: Product picking



Figure 4.8: Product releasing



Figure 4.9: Occlusions

A further experimentation concerns the use of thermal cameras, already used as anti-fraud technologies for the cold objects scenario. In fact, thanks to the chromatic mapping it is possible to detect the picking of objects from the fresh or frozen department, which stand out clearly from the rest of the context and are clearly visible in contact with the human body (given the considerable thermal difference). In this regard, an invention patent was filed and obtained on the use of computer vision applied to thermal objects for anti-shoplifting purposes (No. 102014902314973). This experimentation has its limits currently in the cost of the bolometer but it can also be useful to improve the performance of in-store analysis activities (e.g. monitoring of barriers or lane robots), and in the future at more affordable prices even on those on board the cart.





Figure 4.10: Thermal detection

## 4.4 Data Mining and Process Mining

To conclude, Market Basket Analysis and Visual Market Basket Analysis can be summarized as an application of Data Mining (MBA) and Process Mining (VMBA). [3] In fact, Market Basket Analysis is considered a Data Mining approach. Data Mining techniques are primarily used to find patterns in a large data sets. With data mining techniques it may be possible to find that certain categories of customers demand a certain product, or to find that the customers who most frequently buy product A are also the ones who just as often buy product B, or that the products placed on a specific location in the shop are also the ones that sell the best. Or in a medical analysis that patients that smoke are the most related to develop lung cancer, or that a large consume of alcohol increase the amount of depressed people. In this way is possible to understand important relationships to improve a business to plan more awareness against cancer. An other recent process oriented approach called Process Mining can be categorised as Business Intelligence that refers to techniques and tools used to analyse large amounts of digital data and retrieve valuable

business knowledge out of them. This purpose is as true for data mining techniques as process mining techniques, even if with different perspectives on the analysis and the results they produce. Both techniques are used to analyse large amounts of data, that it would be impossible to analyse manually and they produce information that can be used in business decisions. Process mining is not used to find relationship data patterns, but rather to find process relationships among data. Process relationship among data tries also to analyse the relationship between causes and effects among the data in a certain process. The input to the process mining analysis are event logs, audit trails, events. So, the analysis provides an overview of processes and activities. Process mining's perspective is not on patterns in the data but in the process events (Trnka, 2010). Visual Market Basket Analysis makes logs of customers by narrative carts, than logs can be used for an application of Process Mining, where we consider the process of buying the consumer at the point of sale. Process Mining is the 'missing link' between data mining and traditional BPM (Business Process Management). Data Mining provides valuable insights through analysis of data, but it does not generally concern processes. The scope of the two branches is to give a powerful instrument in order to better understand data and process and then to find a way to underline the data relationship of pattern and process to find out the weakness and try to improve the business. One of the most important tools for Datamining is Weka. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Market Basket Analysis can be easily implemented in Weka. In the following figures it is possible to see the run of Apriori algorithm in Weka.

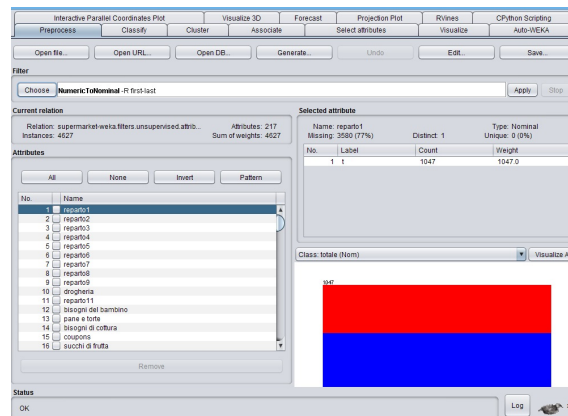


Figure 4.11: Weka MBA

```

=== Run information ===
Scheme: weka.associations.Apriori -M 50 -I 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -cc -1
Relation: supermarket-weka.filters.unsupervised.attribute.NumericalToNominal-First-last
Instances: 4627
Attributes: 217
[list of attributes omitted]
=== Associator model (full training set) ===

Apriori
=====
Minimum support: 0.1 (463 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 52
Size of set of large itemsets L(2): 634
Size of set of large itemsets L(3): 2598
Size of set of large itemsets L(4): 3950
Size of set of large itemsets L(5): 2470
Size of set of large itemsets L(6): 558
Size of set of large itemsets L(7): 20

Best rules found:
1. biscotti=t cibo surgelato=t snack per le feste=t frutta=t vegetali=t totale=high 510 ==> pane e torte=t 478 <conf:(0.94)> lift:(1.3) lev:(0.02) [110] conv:(4.33)
2. biscotti=t cibo surgelato=t formaggio=t frutta=t totale=high 495 ==> pane e torte=t 463 <conf:(0.94)> lift:(1.3) lev:(0.02) [106] conv:(4.2)
3. biscotti=t formaggio=t frutta=t vegetali=t totale=high 513 ==> pane e torte=t 479 <conf:(0.93)> lift:(1.3) lev:(0.02) [109] conv:(4.11)
4. biscotti=t formaggio=t biscotti=t snack per le feste=t frutta=t totale=high 557 ==> pane e torte=t 520 <conf:(0.93)> lift:(1.3) lev:(0.03) [119] conv:(4.11)
5. biscotti=t cottura=t formaggio=t frutta=t vegetali=t totale=high 519 ==> pane e torte=t 483 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.93)
6. cibo surgelato=t snack per le feste=t carta igienica=t frutta=t totale=high 518 ==> pane e torte=t 482 <conf:(0.93)> lift:(1.29) lev:(0.02) [109] conv:(3.92)
7. suocchi di frutta=t biscotti=t snack per le feste=t frutta=t totale=high 529 ==> pane e torte=t 492 <conf:(0.93)> lift:(1.29) lev:(0.02) [111] conv:(3.9)
8. biscotti=t formaggio=t frutta=t totale=high 554 ==> pane e torte=t 543 <conf:(0.93)> lift:(1.29) lev:(0.03) [122] conv:(3.9)
9. biscotti=t snack per le feste=t frutta=t vegetali=t totale=high 536 ==> pane e torte=t 554 <conf:(0.93)> lift:(1.29) lev:(0.03) [125] conv:(3.89)
10. biscotti=t cottura=t biscotti=t cibo surgelato=t frutta=t vegetali=t totale=high 561 ==> pane e torte=t 521 <conf:(0.93)> lift:(1.29) lev:(0.03) [117] conv:(3.84)

```

Figure 4.12: A priori

One of the most important tools used to conduct real Process Mining from logs is ProM, an extensible framework, written in Java, that supports a wide variety of process mining techniques in the form of plug-ins. The input of ProM is represented by logs, characterized by events, concepts and timestamps. Considering an OLTP log, events are represented by the tasks of the operators (concepts) and timestamps are the date and time records of the operations. The choice of the algorithm of process mining determines the different representation of the process analyzed. This is an important choice to focus the attention on the Process Gap Analysis.

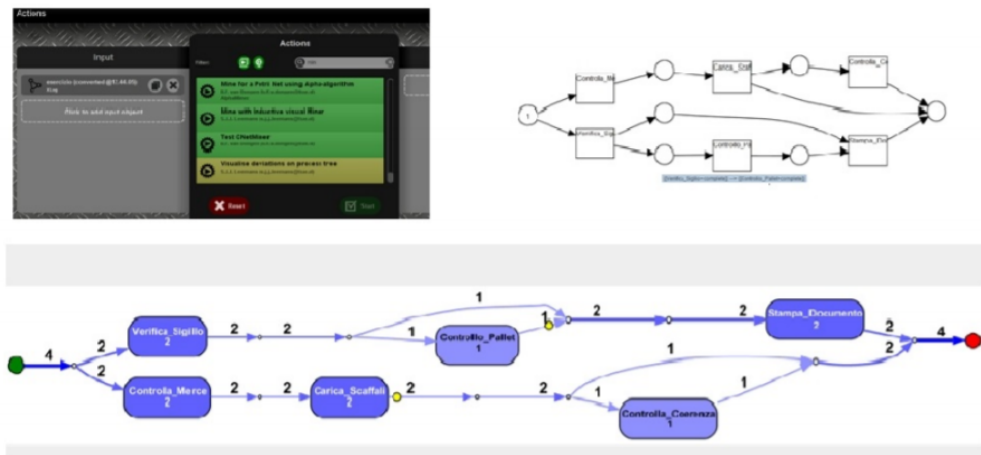


Figure 4.13: ProM

The evaluation of the process gap is the estimation of the distance between the actual process model and the expected process model. The last is a process known as conformance checking (Van der Aalst, et al, 2012). These models differ by nature as the former is merely descriptive whereas the latter is essentially prescriptive. In addition, they are usually developed with both different conceptual and practical tools and, as a consequence, they may also be represented in different ways and formats. In general, moreover, the models can be encoded in multiple representations to serve different goals. When comparing two models using a practical computer-based procedure, obvious prerequisites include the comparability of their (ultimate) representations – which must be digital, formal, unambiguous – and the comparability at the conceptual meta-level. In this sense, the flexibility of ProM for what concerns the output format represents a valuable feature for a tool which aims to support automated process analysis. The intended process model is created in order to document how the actual business process should be carried out. To fulfill its prescriptive role, a hard or soft copy of the procedural description of the process is typically handed out to the operational stakeholders. Such a procedural specification is often represented in natural language – which is inherently subject to ambiguity – possibly accompanied by diagrams expressed in informal or semi-formal notations (e.g., UML activity diagrams or BPMN (White, Stephen, 2004)). In more concrete terms, the key question is the following one: how can we create – e.g., from process descriptions expressed in natural language – a model that can be used to

produce representations that can be effectively compared against the actual process description – e.g., as produced by a tool such as ProM? It seems reasonable to have the chosen approach provide for i) a common semantic layer to give name and meaning to process elements, ii) a welldefined notation (comprehensible and/or usable by business experts) for describing processes with clear semantic links, and iii) tools to analyse and compare process descriptions according to proper semantic rules. In practice, the choice of the modeling language is not easy because a tension exists between expressivity and analyzability. For example, a notation such as BPMN, while suitable for modeling, tends to produce diagrams that are not amenable to analysis – unless considering a proper BPMN subset or transforming BPMN diagrams to Petri nets (Kalenkova, et al, 2015). By the way, keeping a distinction between ‘external’ models (employed for process specification and human communication) and ‘internal’ models (used for analysis) may be valuable. Conformance checking is commonly implemented by replaying history (i.e., event logs) on the expected process model, which is typically represented as a transition system such as a Petri net. However, the initial model representation may be different. For instance, it is possible to load a BPMN diagram into ProM, which results in a BPMN-to-Petri-Net conversion, and then use the tool to analyse and enrich the model with conformance information (Kalenkova, et al, 2015). Then, concluding, for marketing applications Process Mining applied to Visual Market Basket Analysis could be useful to detect the process gap that exists between the path made by the consumer and the path provided by the categories, buyers and marketing specialists by virtue of the type of customer (obtained also thanks to the feedback from DOOH), promotions, seasonality and assorted range of products. In this way, it would be possible to obtain useful indications for a more pertinent exhibition and proposal from the real process.



# Bibliography

- [1] Farinella G., Farioli G., Battiato S., Leonardi S., Gallo G., ‘Face Re-Identification for Digital Signage Applications’, VAAM 2014, 2014.
- [2] Camporesi A., Cascini E., Langone S., Santarcangelo V., ‘Metodo per la localizzazione di oggetti in un’area delimitata’, UIBM Patent N. 0001415239, 2012.
- [3] Santarcangelo V., Giacalone M., et al., ‘Big Data Process Analysis: From Data Mining to Process Mining’, CLADAG 2017, 2017.
- [4] Santarcangelo V., Farinella G.M., Furnari A., Battiato S., ‘Market basket analysis from egocentric videos’, Pattern Recognition Letters, Volume 112, Pages 83-90, 2018.
- [5] Santarcangelo V., Farinella G.M., Battiato S., ‘Egocentric Vision for Visual Market Basket Analysis’, EPIC ECCV, 2016.
- [6] Muller J., Exeler J., Buzeck M., Kruger A., ‘Reflective Signs: digital signs that adapt to audience attention’, LNCS, vol. 5538, pp. 1724. Springer, Heidelberg, 2016.
- [7] Distante C., Battiato S., Cavallaro A., ‘Video Analytics for Audience Measurement’, VAAM , 2014.
- [8] Politecnico di Milano, ‘Convegno dell’Osservatorio del Politecnico di Milano on Mobile Advertsing data’, 2016.
- [9] Kotler P., ‘Marketing 3.0 - Value Driven Marketing’, Seminar Kuwait, 2011.
- [10] Camporesi A., Camporesi M.C., Santarcangelo V., et al., ‘FROM ”RULES OF THE THUMB” TO DIGITAL ERA MANAGEMENT APPROACH.

- THE INNOVATION AND DEVELOPMENT PATH TAKEN BY A CONSUMER GOODS RETAIL DISTRIBUTION SOUTHERN ITALIAN COMPANY: CASE HISTORY AND PERSPECTIVES', 5th International Conference Economy Business - ISE International Scientific Events, At Elenite Holiday Village, Bulgaria, Volume: Journal of International Scientific Publications - Volume 10, 2016 - ISSN 1314-7242, 2016.
- [11] GFK, 'GFK Value Scope', 2016.
- [12] GFK , 'Tech Trends 2016', 2016.
- [13] Ravnik R., Solina F., 'Interactive and Audience Adaptive Digital Signage Using Real-Time Computer Vision', IJARS , 2013.
- [14] Santarcangelo V., Farinella G., Battiato S., 'Gender recognition: methods, datasets and results', VAAM, 2015.
- [15] Kim H., Lee S., Lee D., Choi S., Ju J., Myung H., 'Real-Time Human Pose Estimation and Gesture Recognition from Depth Images Using Superpixels and SVM Classifier', SENSORS, 2015.
- [16] Camporesi A., Santarcangelo V., 'Sistema per la misurazione della variazione dello stato comportamentale di un interlocutore', UIBM Patent P. 102014902291114, 2014.
- [17] Miller R.B., Heiman S.E., 'The New Strategic Selling', Warner Books, 1994.
- [18] Elliot A., Maier M., 'Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans', Annual review of psychology, 2013.
- [19] Grosu E., Grosu V., Preja C., Iuliana B., 'Neuro-linguistic Programming Based on the Concept of Modelling', Procedia, 2014.
- [20] E.Fazl-Ersi, M.Mousa-Pasandi, R.Laganiere, M.Awad, 'Age and gender recognition using informative features of various types', International Conference on Image Processing, 2014.

- 
- [21] H. Liu, Y. Gao, C. Wang, 'Gender identification in unconstrained scenarios using self-similarity of gradients features', International Conference on Image Processing, 2014.
- [22] M.Borgi, M.ElArbi, C. BenAmar, D. Labate, 'Face,gender and race classification using multiregularized features learning', International Conference on Image Processing, 2014.
- [23] H. Ren, Z. Li, 'Gender recognition using completely-aware local features', ICPR, 2014.
- [24] M. Heiman, 'THE NEW STRATEGIC SELLING', 2005.
- [25] A.Torrise, G.Farinella, G.Puglisi, S.Battiato, 'Selecting discriminative clbp patterns for age estimation', VAAM,2015.
- [26] R. Ravnik, F. Solina, 'Interactive and Audience Adaptive Digital Signage Using Real-Time Computer Vision', IJARS,2013.
- [27] B. Batagelj, R. Ravnik, F. Solina, 'Computer vision and digital signage', Tenth International Conference on Multimodal Interfaces, 2008.
- [28] J. Muller, J. Exeler, M. Buzeck, A. Kruger, 'Reflective Signs: digital signs that adapt to audience attention', LNCS, vol. 5538, pp. 1724. Springer, Heidelberg, 2009.
- [29] H.Kim, S. Lee, D. Lee, S. Choi, J. Ju, H. Myung, 'Real-Time Human Pose Estimation and Gesture Recognition from Depth Images Using Superpixels and SVM Classifier', SENSORS, 2015.
- [30] A.Elliot, M. Maier, Color Psychology, 'Effects of Perceiving Color on Psychological Functioning in Humans', Annual review of psychology, 2013.
- [31] E.Grosu, V.Grosu, C.Preja, B.Iuliana, 'Neuro-linguistic Programming Based on the Concept of Modelling', Procedia,2014.
- [32] A. Lanitis, C. Taylor, T. Cootes, 'Toward automatic simulation of aging effects on face images', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002.

- 
- [33] P. Viola, M. Jones, 'Robust Real-time Object Detection', *International Journal of Computer Vision*, 2004.
- [34] A. M. Martinez, R. Benavente, 'The AR Face Database', *CVC Technical Report 24*, 1998.
- [35] J. Wen, Y. Xu, J. Tang, Y. Zhan, Z. Lai, 'Joint video frame set division and low-rank decomposition for background subtraction', *IEEE Transactions on Circuits and Systems for Video Technology*, 2014.
- [36] C. E. Thomaz, G. A. Giraldi, 'A new ranking method for Principal Components Analysis and its application to face image analysis', *Image and Vision Computing*, 2010.
- [37] P. J. Phillips, H. Moon, P. J. Rauss, S. Rizvi, 'The FERET evaluation methodology for face recognition algorithms', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.
- [38] K. Ricanek Jr, T. Tesafaye, 'MORPH: A Longitudinal Image Database of Normal Adult Age-Progression', *IEEE 7th International Conference on Automatic Face and Gesture Recognition*, 2006.
- [39] G. Huang, M. Ramesh, T. Berg, E. Learned-Miller, 'Labeled faces in the wild: a database for studying facerecognition in unconstrained environments', 2008.
- [40] T. Danisman, I. M. Bilasco, C. Djeraba, 'Cross-database evaluation on normalized raw pixels for gender recognition under unconstrained settings', *International Conference on Pattern Recognition*, 2014.
- [41] B. Boser, I. Guyon, V. Vapnik, 'A Training Algorithm for Optimal Margin Classifiers', *Proceedings of the fifth annual workshop on Computational learning theory*, 1992.
- [42] Y. Freund, R. Schapire, 'A Short Introduction to Boosting', *Journal of Japanese Society for Artificial Intelligence*, 1999.
- [43] E. Eiding, R. Enbar, T. Hassner, 'Age and gender estimation of unfiltered faces', *IEEE Transactions on Information Forensics and Security*, 2014.

- 
- [44] S. Jia, N. Cristianini, 'Learning to classify gender from four million images', *Pattern Recognition Letters*, VOL 58 NO.1 35-41, 2015.
- [45] P. Carcagni, M. Del Coco, P. Mazzeo, A. Testa, C. Distanto, 'Features descriptors for demographic estimation: a comparative study', *VAAM*, 2014.
- [46] R. E. Shapire, 'A brief introduction to boosting', *International Joint Conference on Artificial intelligence*, 1999.
- [47] Y. Andreu, J. Lopez-Centelles, R. Mollineda and P. Garca-Sevilla, 'Analysis of the effect of image resolution on automatic face gender classification', *International Conference on Pattern Recognition*, 2014.
- [48] A. Gallagher, T. Chen, 'Clothing Co-segmentation for Recognizing People', *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [49] MPLAB, 'The MPLab GENKI Database, GENKI-4K Subset', <http://mplab.ucsd.edu>, 2011.
- [50] A. Gallagher, T. Chen, 'Understanding Groups of Images of People', *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
- [51] S. Ming, X. Siyu, Y. Fu, 'Genealogical Face Recognition based on UB KinFace Database', *IEEE CVPR Workshop on Biometrics* (2011)
- [52] Miller, R. B., Heiman, S. E., Sanchez, D., Tuleja, T., 'The new strategic selling: the unique sales system proven successful by the world's best companies', *Kogan Page Publishers*, 2004
- [53] S. Mann et al, 'An introduction to the 3D workshop on egocentric (first-person) vision', *Computer Vision and Pattern Recognition Workshops*, 2014.
- [54] P. Agrawal, J. Carreira and J. Malik, 'Learning to see by Moving', *IEEE International Conference on Computer Vision*, 2015.
- [55] A. Furnari, G.M. Farinella and S. Battiato, 'Recognizing Personal Contexts from Egocentric Images', *International Workshop on Assistive Computer Vision and Robotics (ACVR)*, in conjunction with *International Conference on Computer Vision*, 2015.

- 
- [56] Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F., Damiani, E., ‘Privacy-aware Big Data Analytics as a Service for Public Health Policies in Smart Cities’, *Sustainable Cities and Society*, 2018.
- [57] Torra, V., ‘Data Privacy: Foundations, New Developments and the Big Data Challenge’, Springer International Publishing, 2017.
- [58] Terry, N., ‘Existential challenges for healthcare data protection in the United States’, *Ethics, Medicine and Public Health*, 3(1), 19-27, 2017.
- [59] D’Acquisto, G., Naldi, M., ‘Big Data e Privacy by design (Vol. 5)’, G Giapichelli Editore, 2017.
- [60] Bertino, E., Ferrari, E., ‘Big Data Security and Privacy’, *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (pp. 425-439), Springer, 2018.
- [61] Cohen, M., ‘Fake news and manipulated data, the new GDPR, and the future of information’, *Business Information Review*, 34(2), 81-85, 2017
- [62] McDermott, Y., ‘Conceptualising the right to data protection in an era of Big Data’, *Big Data Society*, 4(1), 2017.
- [63] Beckett, P., ‘GDPR compliance: your tech department’s next big opportunity’, *Computer Fraud Security*, 2017(5), 9-13, 2017
- [64] Mittal, S., Sharma, P., ‘General Data Protection Regulation (GDPR)’, *Asian Journal of Computer Science And Information Technology*, 7(4), 2017
- [65] Y. Poleg, C. Arora and S. Peleg, ‘Temporal Segmentation of Egocentric Videos’, *International Conference on Computer Vision and Pattern Recognition*, 2014.
- [66] P. Tan, M. Steinbach and V. Kumar, ‘Introduction to Data Mining’, Chapter 6, Addison-Wesley Companion Book Site, 2006.
- [67] S. Wang, S. Fidler and R. Urtasun, ‘Lost Shopping! Monocular Localization in Large Indoor Spaces’, *IEEE International Conference on Computer Vision*, 2015.

- 
- [68] A. Dosovitskiy, P. Fischer and V. Golkov, ‘RFID Based Smart Shopping and Billing’, IEEE International Conference on Computer Vision, 2015.
- [69] A. Dosovitskiy, P. Fischer and V. Golkov, ‘FlowNet: Learning Optical Flow with Convolutional Networks’, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 12, 2013.
- [70] B. Xiong, G. Kim, L. Sigal, ‘Storyline Representation of Egocentric Videos with an Application to Story-based Search’, International Conference on Computer Vision, 2015.
- [71] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera and S. Battiato, ‘Representing Scenes for Real-Time Context Classification on Mobile Devices’, Pattern Recognition, Vol. 48, Issue 4, 2015.
- [72] J. Platt, N. Cristianini, J. Shawe-Taylor, ‘Large Margin DAGs for Multiclass Classification’, 547–553, MIT Press, 2000.
- [73] Z. Lu, K. Grauman, ‘Story-Driven Summarization for Egocentric Video’, International Conference on Computer Vision and Pattern Recognition, 2013.
- [74] Y. Poleg, T. Halperin, C. Arora, S. Peleg, ‘EgoSampling: Fast-Forward and Stereo for Egocentric Videos’, Conference on Computer Vision and Pattern Recognition, 2015.
- [75] Y. Lee, J. Ghosh, K. Grauman, ‘Discovering Important People and Objects for Egocentric Video Summarization’, Conference on Computer Vision and Pattern Recognition, 2012.
- [76] D. Damen, ‘You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video’, British Machine Vision Conference, 2014.
- [77] A. Fathi, J. Rehg, ‘Modeling Actions through State Changes’, International Conference on Pattern Recognition, 2013.
- [78] A. Fathi, X. Ren, J. Rehg, ‘Learning to Recognize Objects in Egocentric Activities’, Conference on Computer Vision and Pattern Recognition, 2011.

- 
- [79] Q. Xu et al., 'A wearable virtual guide for context-aware cognitive indoor navigation', International Conference on Human-Computer Interaction with Mobile Devices and Services, 2014.
- [80] T. Starner et al., 'Visual Contextual Awareness in Wearable Computing', International Semantic Web Conference, 1998.
- [81] A. Kendall, M. Grimes, R. Cipolla, 'PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization', International Conference on Computer Vision, 2015.
- [82] Veho, 'Veho Muvi Cam', [www.vehomuvi.com](http://www.vehomuvi.com), April 2016
- [83] M. Sahidullah , G. Saha, 'Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition', pp 543-565, Speech Communication, 2012.
- [84] J. Barron, D. Fleet, S. Beauchemin, 'Performance of Optical Flow Techniques', pp 43-77, International Journal of Computer Vision, vol 12, ISSUE1, 1994.
- [85] C.Wu, 'Towards Linear-time Incremental Structure from Motion', pp 145-175, International Conference on 3D Vision, 2013.
- [86] Y. Bengio, 'Representation Learning: A Review and New Perspectives', IEEE Pattern Analysis and Machine Intelligence, vol 35, issue 8, 2013.
- [87] Krizhevsky, A., Sutskever, I., Hinton, G., 'Imagenet classification with deep convolutional neural networks', Advances in neural information processing systems, 2012.
- [88] Muda, L., Begam, M., Elamvazuthi, I., 'Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques', Journal of Computing, Vol. 2, Issue 3, 2010.
- [89] Oliva, A., Torralba, A., 'Modelling the shape of the scene: a holistic representation of the spatial envelope', International Journal of Computer Vision, vol 42, issue 3, pp 145-175, 2001.



- 
- [90] A. Ross, C. Chen, 'Can Gender Be Predicted from Near-Infrared Face Images?', Proc. of International Conference on Image Analysis and Recognition (ICIAR), 2011.
- [91] D. Yi, Z. Lei, S. Liao, and S. Z. Li, 'Learning face representation from scratch', Eprint Arxiv, 2014.
- [92] S. Jia, N. Cristianini, 'Gender Classification by Deep Learning on Millions of Weakly Labelled Images', 16th International Conference on Data Mining Workshops (ICDMW), 2016.
- [93] G. Antipova, S. Berrania, J. Dugelayb, 'Minimalistic CNN-based ensemble model for gender prediction from face images', Pattern Recognition Letters, 2015.
- [94] M. C. Santana, J. Lorenzo-Navarro, and E. Ramon-Balmaseda, 'Descriptors and regions of interest fusion for gender classification in the wild', Image and Vision Computing, Volume 57, Pages 15-24, 2017.
- [95] Z. Qawaqneh, A. Mallouh, B. Barkana, 'Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model', Arxiv, 2017.
- [96] H. Hosseini, B. Xiao, R. Poovendran, 'Google's Cloud Vision API is Not Robust to Noise', 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017.
- [97] A. Marinos and G. Briscoe, 'Community Cloud Computing', CloudCom 2009: Cloud Computing pp 472-484, 2009.
- [98] L. Prieto, Z. Oplatková, 'Comparing the Performance of Emotion-Recognition Implementations in OpenCV', Cognitive Services, and Google Vision APIs, WSEAS Transactions on Information Science and Applications, Volume 14, pp. 184-190, 2017.
- [99] D. Cardone, A. Merla, 'New Frontiers for Applications of Thermal Infrared Imaging Devices: Computational Psychophysiology in the Neurosciences', Sensors 17, 2017.

- 
- [100] R. Ribeiro, J. Fernandes, A. Neves, 'Face Detection on Infrared Thermal Image', SIGNAL, 2017.
- [101] D. Apiletti, E. Baralis, et al., 'Frequent Itemsets Mining for Big Data: A Comparative Analysis', Big Data Research, Volume 9, Pages 67-83, 2017.
- [102] L. Wu, K. Gong, Y. He, X. Ge, J. Cui, 'A Study of Improving Apriori Algorithm', 2nd International Workshop on Intelligent Systems and Applications, 2010.
- [103] M. Kaur, U. Grag, 'ECLAT Algorithm for Frequent Item sets Generation', International Journal of Computer Systems, Vol. 01, Issue, 03, 2014.
- [104] J. Han, J. Pei, 'Mining Frequent Patterns by Pattern-Growth', Computer Science, Volume 2 Issue 2, Pages 14-20, 2000.
- [105] A. Trnka, 'Market Basket Analysis with Data Mining methods', International Conference on Networking and Information Technology, 2010.
- [106] M. Aydogdu, V. Celik, M. Demirci, 'Comparison of Three Different CNN Architectures for Age Classification', IEEE 11th International Conference on Semantic Computing (ICSC), 2017.
- [107] J. Shao, C. Qu, J. Li, 'A performance analysis of convolutional neural network models in SAR target recognition', SAR in Big Data Era: Models, Methods and Applications (BIGSAR DATA), 2017.
- [108] M. Hedjazi, I. Kourbane, Y. Genc, 'On identifying leaves: A comparison of CNN with classical ML methods', 25th Signal Processing and Communications Applications Conference (SIU), 2017.
- [109] C. Lee, H. Kim, K. Oh, 'Comparison of faster R-CNN models for object detection', 16th International Conference on Control, Automation and Systems (ICCAS), 2016.