University of Modena and Reggio Emilia
XXX cycle of the international doctorate school in
Information and Communication Technologies
Doctor of Philosophy dissertation in
Computer Engineering and Science

Affective and Interactive Evaluation of Multimedia Data

Fabrizio Balducci

Supervisor: Prof. Costantino Grana
PhD school director: Prof. Sonia Bergamaschi

Modena, 2018

Review committee composed of:

Prof. Lamberto Ballan
Prof. Giuseppe Serra

Contents

Chapter 1

Introduction

The general topic of this work consists in the analysis and evaluation of multimedia data that, by definition, are heterogeneous as the perception of human senses: hearing with sound and music, view with images and video, sense of touch with devices that feature multiple interaction modes.

The thesis deals with two macro-topics (interactive multimedia and medical images) that find a union in the chapter dedicated to the serious gaming, which combines fun and interaction with study and learning. The expression "Interactive multimedia" refers to videogames, the only software that amalgamates the different multimedia sources to achieve a goal hard to be uniquely defined since the various levels and shades that every person can give to it: fun.

At a higher level, a videogame represents an interactive *virtual world* that, thanks to its extreme ductility and to the modern technological advances, can be been used in fields other than pure entertainment as in the educational sector or in the scientific research as a controlled and cheap simulation environment or as a source of synthetic experimental data.

It is undoubted that the production of videogames has contributed significantly to the technology evolution, especially in the 90s with the introduction of GPUs in dedicated graphic cards and the increasing of the internet bandwidth used for the online multiplayer: from scarce pastime composed by few pixels on monochrome screens, videogames evolved to reach and overcome traditional entertainment industries such as cinema, literature, music, boardgames and comics books; moreover, often happens a

contamination between this media which leads to exploiting the intellectual properties and to interchanging authors and franchises.

The pervasiveness of the gaming platforms (home consoles, personal computers, mobile devices, web applications) and the availability of hardware resources that allow the exchange and storage of large data streams have helped to develop a market that moves large amounts of money in terms of production, promotion and sales; in this globalized market where the boundaries between success, budget balance or bankruptcy is very thin, the stakeholders are increasingly inclined to reach the best results with products that suits everyone, forgetting that every person has specific needs and expectations; it must also be considered that gamers tend to not end videogames since they soon lose interest due to repetitiveness and lack of new challenges so, in this way, a multimedia product becomes obsolete and off-market in very few time causing waste of time and production resources.

Considering an ideal mode of use like the *Dynamic Difficulty Adaptation* that only the videogame medium can allow, the Cap. 2 after a brief historical introduction and the analysis of fundamental concepts such as *genre* and *gameplay* introduces the term *Affective Ludology* with a wide and exhaustive corpus of experiments and studies known in the scientific literature: this *science of the game* in fact, applies the scientific method to emotional and cognitive interactions and modalities, also referring to theories of different scientific fields such as psychophysiology and Human-Computer Interaction.

To perform a scientific study, data that can define and identify emotions and affective states in the most objective way are needed: an example of that are physiological data such as brainwaves, whose values are captured by electroencephalogram (EEG) from the Emotiv EPOC headset and managed by a specially developed synchronization tool.

With the theoretical definition of the affective states of "Flow" and "Boredom", the role-playing videogame Neverwinter Nights 2 represents the ideal testing environment for formal design guidelines with which develop events and activities that, in a transparent way for the player, can dynamically highlight and recognize the emotional changes and induce and manipulate some affective states; to reach this goal experiments have been conducted proposing an evaluation method that compares machine learning techniques performed on the EEG data pre-classified by the Emotiv helmet.

The interest in the analysis of medical images, and more specifically the dermoscopic ones for skin lesions, has seen an increase in the computer vision field thanks to the new methods based on machine learning, useful to

develop proactive agents for the automatic diagnosis of malignant lesions (melanomas): this is the topic of Cap. 3 with a further study related to the Convolutional Neural Networks and the deep learning approach in the following chapter.

To build a dataset that can be used as ground-truth in a Medical Image Management System (MIMS), a tool for high-quality manual annotation has been developed by following principles and theories of Human-Computer Interaction; moreover, to inquire about its usability an experimental survey has been conducted with non-IT subjects.

After extraction of the primary features as colors and pixels of a manual annotation, functions and algorithms to extract derived features such as contours, intersections, shapes and numerical values have been developed. The process of *segmentation* consists in identifying and isolating the pixels of a dermatological lesion and so, in this chapter, standard image processing techniques are exploited with workarounds that refine results; therefore a pre-processing phase is necessary to eliminate artifacts such as thick hairs that makes hard the edge and contour identification and the proposed algorithm exploits the *general morphological closure* operator and the pixel masks associated to color channels to identify the structures that must be removed in the least invasive way.

Two segmentation methods are proposed: the first uses the *thresholding* by exploiting the peculiar structure of a dermoscopic image while the second uses *color clustering* combined with "tolerance masks" for those pixels that are ambiguously classified between simple skin or lesion; the experimental study compares the evaluation metrics of the two techniques, applied on two datasets (with and without hair removal) whose images are considered "hard" due to their peculiar features and artifacts.

The chapter 5 is dedicated to the design of a multimedia software that combines the analysis of medical images with the interactive learning represents the union of the topics addressed in this work: theoretical principles of serious games, already experimented in literature, are exposed with the aim to develop the prototype of the *Annote* serious videogame, also supplying implementation and technological details and the gameplay choices and assumptions.

Chapter 2

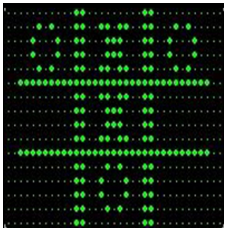
Interactive Multimedia

A commonly accepted definition of *Game* is provided by Lindley [50] as "a goal-directed and competitive activity conducted within a framework of agreed rules": given this definition, it is often said that *to play* involves to learn and master the internal mechanics.

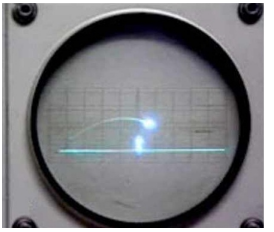
Nowadays the videogame industry requires high skilled workers [108] and budgets that rivals with the classical entertainment and cultural industries: their production has become a large business [127] estimated in 93$ billion in 2013 [134], for example the development cost of *Call of Duty Modern Warfare 2* was between 40 and 50$ million, with a marketing and release budget of $200 million [141]. Other aspects that highlight the market widening are the intersection between toys and games [146], the phenomenon of in-game advertising [9] and the importance of press and player critics [160]. The peculiar aspect that distinguishes this relatively novel (mass)medium from others as cinema, music, literature and comics is the *interactivity* that, together with all other multimedia stimuli, is able to engage the users (more specifically "players" or "gamers") in deep and unexpected ways.

In over 30 years gaming platforms evolved moving from large old cabinets to smart-watches that easily permit to play in mobility, for example allowing learning [23], large data exploration [81] or archaeological visits [13]; also the academic world also considers videogaming as a multidisciplinary teaching field [21] and research [36] [89].

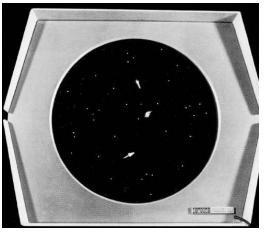
The interest about interactive entertainment software begins in the early 50's in the american academic faculties and research labs (Fig. 2.1) while



(a) OXO by A.S. Douglas (1952).



(b) Tennis for Two by W. Higinbotham (1958).



(c) Spacewar! by S. Russell (1961).

Figure 2.1: First experiments for interactive software.

from the late 70's it starts to become a business.

Producing videogames involves the tacit knowledge of designers whose target is to maximize the *player experience*, i.e. the set of feelings and opinions that come from an activity which has to capture player's senses; since each person is different due to factors like age, education, country and personal history results very hard to design and develop a product that suits each potential player.

2.1 Gameplay and Genres

The design of a videogame depends on its *Gameplay* and *Genre* that mutually affect each other and that represent the most creative part of a game project: a *game designer* has to master this key factors and must be a complete "director" with a wide multidisciplinary culture that ranges from the technical studies to the literary ones (examples are S. Miyamoto and Hideo Kojima, respectively the creators of the Super Mario and Metal Gear sagas).

The Gameplay represents the *internal rules* of the game, the set of mechanics that characterize its interactive dynamics and defines how narrative and ludic elements guarantee that players don't quit the game [43]. The Genre exposes design solutions and constant features coming from previous products widely accepted by the sell-market and usually the term is used to classify and categorize entities by highlighting their common and uncommon features, as it happens for books, movies and species: in

videogame context it refers to which gameplay a gamer prefers to satisfy his fun requests and so, very often, happens that he refers to a videogame citing its genre or by a direct comparison stating that *game X is like Y* (already known) [110].

In the work of Machado *et al.*[100] is presented a *player modeling taxonomy* with several game platforms that can be used by researchers while Lewis *et al.*[93] propose a taxonomy for game bugs and Pinelle *et al.*[124] consider how genres concern usability and design issues. Khaleghi and Lugmayr [84], and Scavarelli and Arya [134], introduce respectively surveys and a *game ethics framework* (tested on *Fable* and *Super Mario Bros.*) to classify according to genre, technologies and gameplay features.

Compiling a taxonomy with a clear genres division involves much subjectivity due to the extreme variety of gameplay: the works of Berens and Howard [22] and Carlá [34] suggest classifications based on historical period and technology platforms.

An interpretation of this classifications is the follow list:

Adventure: featuring environment exploration, riddles, cinematics and strong narrative plot.

• *textual* (Zork) • *graphic* (Monkey Island) • *slide-show* (Myst)	• *dynamic* (Tomb Raider) • *survival* (Resident Evil)

Action: characterized by an immediate and dynamic gameplay with fast movements, jumps, fights and weapons.

• *platform* (Super Mario Bros.) • *fight or beat'em up* (Street Fighter) • *first-person shooter or FPS* (Doom)	• *third-person shooter* (Max Payne) • *light gun shooter* (House of the Dead) • *stealth* (Metal Gear Solid)

Simulative: based on the simulation of real activities, vehicles or machinery, characterized by complex control systems.

• *flight (MS Flight Simulator)* • *drive* (Gran Turismo)	• others like trains, mech robots, space, fishing, etc.

Strategic: based on long-term decision strategies with the management and conquest of resources and equipments.

• *management* (Sim City) • *real-time strategy or RTS* (Warcraft)	• *real-time tactics* (Total war) • *turn-based strategy* (Civilization)

God: contains strategic and simulative features; the player controls the environment and the social dynamics of a independent living population.

• *people god* (Populous)	• *life simulator* (The Sims)

Role-Playing: based on both strategy and narrative elements with presence of action; relationships with allies and enemies are of great importance to foster player engagement and empathy.

• *computer role-playing* (Neverwinter Nights) • *japanese RPG* (Final Fantasy)	• *action RPG* (Diablo) • *massively multiplayer online or MMORPG* (World of Warcraft)

Puzzle: featuring a fast and immediate gameplay (suitable for mobile devices), they propose logic puzzles with increasing difficulty.

• *shapes* (Tetris) • *physic* (Angry Birds)	• *cards* (HearthStone)

Sport: based on both simulation and action, features real statistics, athletes, sponsors, marks and sites; often they are linked to real tournaments or events.

• *soccer and others* (FIFA, NBA, ...) • *drive* (SBK)	• *manager* (Football Manager)

Musical: features a gameplay based on sounds and choreographies; often uses dedicated hardware peripheral.

• *dance* (Just Dance) • *rhythm* (Guitar Hero)	• *sing* (Karaoke)

Educational: they exploit the multimedia and interactive features of videogames with the aim to teach something.

• *grammar and math* (Brain Training) • *exergame* (Wii Fit)	• *serious games*

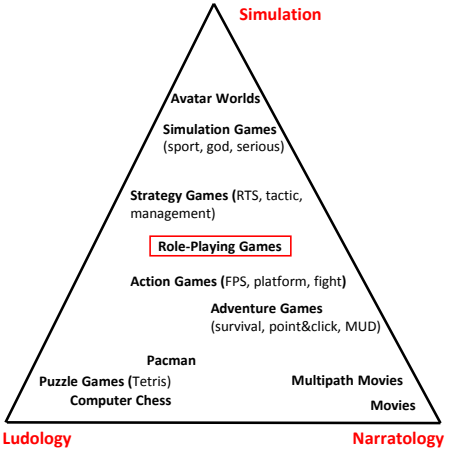


Figure 2.2: The triangular scheme about the polarity of game genres, according to three metrics from which the gameplay is characterized.

Lindley [50] defines a triangular scheme (an interpretation is in Fig. 2.2) where each vertex denotes the polarization of game genres towards the metrics of *Simulation*, *Ludology* and *Narratology*. Privileging the Simulation metric implies the reconstruction of physical laws and real world constraints (or particular aspects of them), denoting a slow gameplay based on high longevity and the tuning of a lot of parameters; conversely the Ludology denotes an immediate frantic gameplay with progressive difficulty, in fact this games usually belong to the *arcade* genre characterized by few options but low longevity.

About the Narratology metric, Langer *et al.*[90] propose the *suspenseful design* to inquire narrative and storytelling aspects of a videogame and Frasca [61] considers the narrative features as an integration for other metrics: it offers ways to enhance involvement similarly to non-interactive media based on dialogues and descriptions (like movies and books).

Wei [153] addresses the *narrative embedding* by structuring contents in horizontal, vertical and modal embedding, while Calleja [32] studies the *narrative involvement*, introducing a model to understand the *player experiential narrative*; finally Zagalo and Torres [161] present an emotion

module from an authoring tool of interactive storytelling.

It is noteworthy that in the scheme in Fig. 2.2, Role-Playing genre is positioned exactly in the center, emphasizing the balance of the metrics. The metrics balancing is one of the reasons for choosing this particular kind of genre for the experimentations in this thesis and so, in Section 2.4, further analysis will bring to the proposition of formal guidelines to design game levels (or "scenarios") and their related tasks.

2.2 Affective Ludology

The concept of playing is strictly linked to our lives, in fact it is one of the first activities that a human experiences by discovering his surrounding world: a successful method to learn rules or how to safely execute a task is in fact to *gamify*, i.e. to see them like an activity designed not only for the pure entertainment but from which gain experience for future goals [62]. Wilkinson [156] introduces *Affective Educational Games* as a way to use emotions in game-based learning: in his work presents a review of currently available technologies, theories and models to recognize and modeling emotions and tasks.

Frasca [66], referring to Csíkszentmihályi's studies [52] (that will be presented in the following paragraphs), defines the modern concept of *Ludology* as the *science of the game* which uses research methods and theories from a wide range of scientific communities (such as Human-Computer Interaction and psychophysiology) with the aim to improve the methodologies to study both players and (video)games, understanding the design of an optimal *player experience*.

Affective computing, as a field of study, was captured by Rosalind Picard [123] while Nacke [112] introduces the concept of *affective ludology* referring to the investigations of affective player-game interaction to understand emotional and cognitive experiences: it must inquire about cognition, emotions, and goal-oriented behavior from a scientific perspective and establish rigorous methodologies (e.g. psychological player testing or physiological response analysis of players).

The costs for a *triple A* title (AAA is a market classification acronym used for videogames with very high development and promotional budgets) oblige editors to accept fewer risks, looking for a formula that will ensure *great games for everyone*, omitting that a gamer needs a certain personalized

gaming experience that suits its peculiar needs.

The proposition of modalities involved in the evaluation and customization processes are at the center of this thesis: while Paavilainen [120] and Korhonen *et al.*[86] propose and review many evaluation heuristics, it becomes crucial to define and manipulate complex concepts like *Fun* with innovative methods that consider player's feelings and preferences by retrieving objective and scientifically evaluable data.

In academic literature there is a great corpus of studies in which videogames (and more generally the virtual worlds) are linked to emotions and affective states: Noah *et al.*[115] measure brain activity with magnetic resonance imaging (MRI) using a clone of *Dance Dance Revolution* videogame to study how different sensory inputs influence the motor output, while Groenegress *et al.*[68] introduce a system for real-time physiological analysis and metaphorical visualization within a virtual environment, considering heart rate, respiration and galvanic skin responses.

Rawn and Brodbeck [131] use questionnaires and *Doom 3* to inquire about violence and aggressive interactions, Gilleade *et al.*[65] uses the affective state is used to manipulate the game session and Jacopin [76] analyzes internal data from *F.E.A.R.*, *Kill-Zone 3* and *Transformers 3* to develop intelligent *NPCs* (Non-Player Characters or bot) i.e. any character controlled by the computer with an Artificial Intelligence.

Dormann and Biddle [57] propose game design for affective learning by introducing in *Ico* an *affective walk-through* while Marczak *et al.*[105] develop feedback-based metrics to empirically express engagement using *Dead island* and *Bioshock 2*, in fact by examining audio-video feedbacks the nature of player's motivations will emerge. Rilling and Wechselberger [132] introduce computer game principles within an automation industry training scenario, Mattiassi [106] inquires human-fighting game interaction using neuroscience theories while Chanel *et al.*[39] analyze physiological signals as indicators for difficulty adaptation in *Tetris*.

Plotnikov *et al.*[125] apply machine learning techniques on brainwaves data to stress the concepts of boredom and flow using *Tetris* while in Canossa *et al.*[33] there are metrics to discover frustration factors by game data during gaming sessions of *Kane & Lynch 2: Dog Days*. Herrewijn *et al.*[73] use surveys to analyze the factor of Immersion in the videogame *Fallout:New Vegas*.

2.3 Detect Brain Signals

In human body there are biological factors that produce detectable activities that do not directly depend on the nervous system but often affect it: for example *ElectroOculography* (EOG) analyzes the movements and the closure of the eyes that cause, in the retina, detectable electrical potential variations; *Electromyography* (EMG) that measures the neuromuscular activities generated by the muscles contractions to provide information on peripheral nerves and skeletal muscles; *Electrocardiogram* (ECG) that measures the difference in electrical potential that the heart produces during the heartbeat contractions.

The human brain is extremely developed and specialized and can be divided into four areas or *lobes*: Frontal, Parietal, Temporal and Occipital (Fig. 2.3) [1].

Frontal lobe: at the front of the brain, it contains the cortical area (the rear) of the motor skills; here the higher psychic activities (thoughts and ideas) are elaborated and, moreover, this lobe participates in the processes of learning and memory where the words are formed and controlled.

Parietal lobe: located at the top of the brain contains the area affected by tactile, painful, pressurized and thermal stimuli. The left side is dominant and controls the understanding of spoken and written language, memory of words and mathematical abilities; the right side controls visuospatial activities such as reconstructing an image with the ability to orient it in space and make it rotate, the perception of the trajectory of a moving object and the body awareness.

Temporal lobe: located at the bottom of the cerebral hemispheres, it processes affectivity, relationships, instinctive reactions and behaviors,

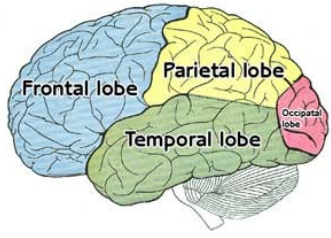


Figure 2.3: Lobes of the human brain [2].

visual recognition and auditory perception. The left temporal lobe includes the spoken language and chooses the words while the right allows to understand the pitch of the speech and the sequence of the sounds.

Occipital lobe: located at the back of the brain, its main activity is to process the visual information including those affecting posture and balance, in fact there are many neurons specialized in the recognition and processing of the details of an image.

Cerebral cortex: is a continuous layer of a few millimeters thick (about 24 mm) consisting of dendrites and synaptic connections that form the outer mantle of the cerebral hemispheres; it is the seat of superior brain functions such as conscience and the emotional and instinctual memory.

Manifestations of mental states can be measured: studies from Mandryk *et al.*[104] have successfully demonstrated how psychophysiological techniques evidence human emotions and cognitive activity during gaming. Andreassi [12] defines *psychophysiology* as "the science which inquires relations between the psychological manipulation and the resulting physiological effects", linking physiological measures with abstract psychological constructs like *Attention* and *Fun*; the work of Craveirinha and Roque [51] offers a theoretical overview of the nature of play activities, and studies how certain ludic elements are in relationship with the emotional spectrum.

A *Brain-Computer Interface* (BCI) is a device that measures brain electrical activities: an *electroencephalogram* represents their graphical description while the "International 10-20 System" (Fig. 2.4) is an internationally recognized method to apply the electrodes on the human scalp and it was developed to ensure standardized reproducibility of experiments and measures made with EEG. In this system there are 21 electrodes and their locations are determined by dividing the total frontback or rightleft distance of the skull perimeters into 10% and 20% intervals [78, 129].

The names that identify the position of an electrode are formed by one or two letters that allow to identify the basin exploration region (Fp: frontpole, F: frontal, C: central, P: parietal, T: temporal, O: occipital) and by a number that identifies the hemispheric (odd numbers: left, equal numbers: right, z: median line). In addition to the 21 electrodes of the international 10–20 system, intermediate electrode positions are also used and their locations and nomenclature are standardized by the *American Electroencephalographic Society guidelines* [117].

The electrical potential generated by a single brain neuron is too weak to

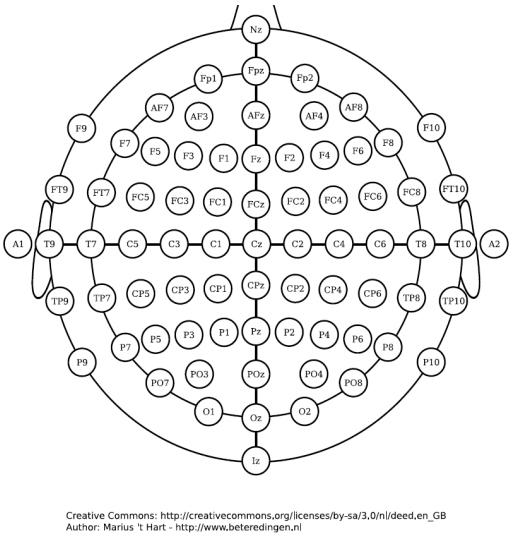
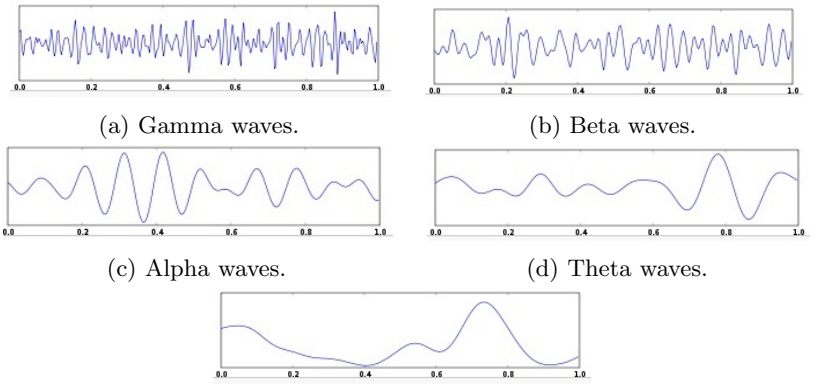


Figure 2.4: The International 10-20 system for the positioning of the EEG electrodes on the human scalp.

be detected by an electroencephalogram, and in fact it is the synchronized activity of millions of neurons having the same spatial orientation that is considered. For their characteristics, the pyramidal neurons on the cerebral cortex are considered the largest emitters of EEG signals producing different brainwaves characterized by their *frequency* [143, 3] (Fig. 2.5):

- *Gamma Waves* have the highest frequency range (30-80 Hz), involved in higher cognitive functioning like memory and information processing; states of anxiety and stress presents high levels of this
- *Beta Waves* known as "high frequency-low amplitude" waves (13-30 Hz), they denote the normal human brain activity and involve conscious thought, logical and critical thinking and socialization; their activity can increase with stimulants like caffeine
- *Alpha Waves* associated to calm and meditation with a regular and synchronized configuration (8-13 Hz); they are also called Berger waves in memory of the inventor of the EEG in 1929



(a) Gamma waves.

(b) Beta waves.

(c) Alpha waves.

(d) Theta waves.

(e) Delta waves.

Figure 2.5: Brainwaves frequency graphs.

- *Theta Waves* involved in daydreaming and REM sleep phase with a low frequency range (3.5-8 Hz); they denote a positive state of deep and raw emotions, intuition and creativity, with streams of consciousness near an hypnotic state
- *Delta Waves* the slowest recorded brainwaves (under 3.5 Hz) associated with deepest levels of sleep; an abnormal activity usually denotes brain injuries and learning problems

An useful and simple scheme to interpret the basic emotions is the *Circumplex model of affect* by Russell *et al.*[133, 126] (an interpretation of which is in Fig. 2.6): it consists of a two-dimensional spatial model with a set of human emotions defined on a circumference, related to the bipolar metrics of *Valence* (X axis) and *Arousal* (Y axis). Valence is an indicator which describes if an emotion is pleasant or unpleasant, while Arousal denotes the intensity (activation or deactivation), i.e. reactivity to stimuli that influence its detection.

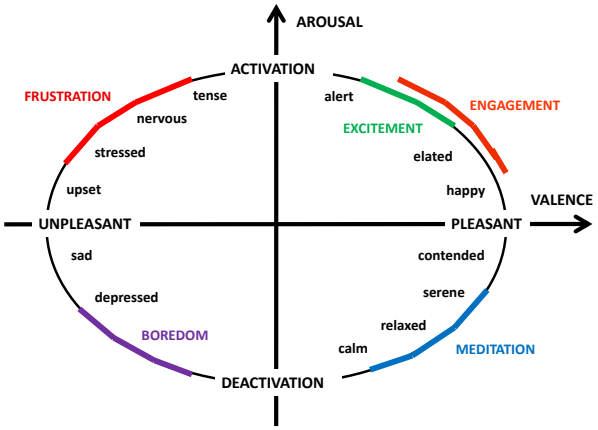
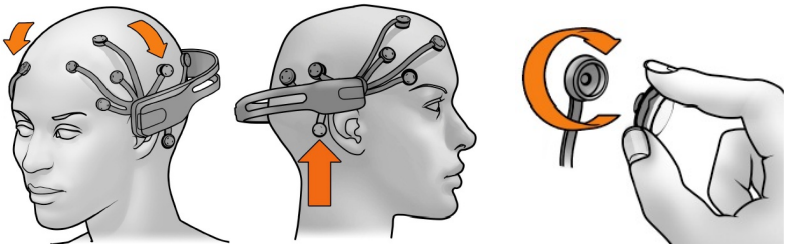


Figure 2.6: A circumplex model: Excitement, pleasant and activated, is located in the upper-right quadrant like Engagement; Meditation, pleasant but not activated, is located in the bottom-right one. Frustration and Boredom are unpleasant, but the first is activated and it is located in the upper-left quadrant while the second is in the bottom-left one.

2.3.1 The EPOC EEG Headset

In the field of science the BCI has been used with various purposes like measure mnemonic and cognitive efforts (Grimes *et al.*[67]) or like a real-time input device: Vachiratamporn *et al.*[148] measure player experience with EEG and heartbeat signals using a survival horror videogame as experimental environment, Liarokapis *et al.*[94] create a *Lego NTX* robot remotely controlled by them while Pour *et al.*[128] control a *Breakout* clone using a Brain-computer Interface device.

Lotte [98] considers BCI as an additional control channel to interact with virtual environments, but its results underline some limits according to which 20% of players cannot use it as a gaming device; adaptive virtual environment and emotion assessment methods are also in [35] while Coulton *et al.*[49] successfully use EEG headset with mobile games while Burke *et al.*[29] introduce game design guidelines for stroke rehabilitation using serious games and Kang *et al.*[82] create a 3D sensory gate-ball game system to improve both physical and mental health of the aged people.



(a) EPOC headset (b) Sensors

Figure 2.7: The Emotiv EPOC EEG headset scheme.

Vourvopoulos *et al.*[150] propose a reconstruction of the ancient Rome that can be explored using brainwaves and also Schwarz *et al.*[136] introduce a *Pong* clone controlled by two Emotiv headset together and Chumerin *et al.*[45] propose an application to navigate *The Maze* guided by brainwaves.

The Emotiv EPOC headset (Fig. 2.7) [5] is a wireless neuro-signal acquisition and processing device with 14 wet sensors (+2 of reference), capable of detecting brainwaves at 128 Hz sequential sampling rate; sensors are placed around the pre-frontal and frontal brain regions according to the international 10–20 standard system and this process is safe because the headset is a passive measuring device and less invasive than other physiological methods.

The terms *affect* and *emotion* can be used interchangeably and usually refers to a short-time emotional peak; conversely, a *mood* or *affective state* denotes a continuous lasting emotional trend which may involve more emotions that influence the global reactions [61].

In this work the BCI headset will be used to retrieve objective physiological data that must to be organized, evaluated and explained; moreover will be exploited the manipulation of emotions and on the definition of Flow and Boredom states by exploiting the five emotion detected by the *Emotiv EPOC* EEG headset which are Excitement, Engagement, Frustration, Meditation and Long-Term Excitement.

This five values are calculated by the internal Emotiv algorithms in a black-box mode on the raw signal of the headset sensors: the reliability of Emotiv EEG data is given in [15, 151, 7].

Engagement is associated to participation and attention and its increase denotes challenging tasks and pleasure to discover new aspects of the game; commonly it is used as an indicator of Fun. If the aim is to induce a Flow state, most of the game tasks must increase this emotion, while the contrary must be true for the *Boredom* induction, because it is a deactivated emotion interpretable as the opposite of Engagement.

Excitement is a positive indicator characterized by muscle tension and increased sweating and heartbeat: since evoked by short-time emotional peaks, it is better to consider it linked to *Long-Term Excitement* i.e. a global value which expresses how stable over time Excitement is. The manipulation of this emotions must work in the same way as seen for Engagement. Engagement and Excitement are closely related to the concept of Fun due to the mutual interaction between player's interest and involvement while facing an activity.

Inducing and analyzing *Frustration* and *Meditation* is difficult due to their high subjectivity and the negative valence. A meditative activity [44] can take a long time and a lot of training to be correctly performed by a motivated subject, also inducing frustration or boredom in the first attempts. Frustration might influence other emotions, in fact different players might perceive the same amount of it as a challenge or as a reason to quit the game: if the cause is internal (laziness, lack of confidence) it can be a motivating force but, if caused by external factors perceived to be outside individual control (i.e. task too hard) it can lead to powerlessness and eventually anger (Dollard *et al.*[56]).

2.4 Design for Affective Ludology

As seen above, the RPG gameplay is less immediate than others since it is based on statistics, skills and object inventory, with a great importance given to environmental exploration; when considering role-playing gameplay features, it will focus on the aspects shown in Lankoski [91]:

- strong story plot with moral choices
- progression of skills and abilities (statistics)
- presence of allies in a group (the "party")
- interaction based on dialogues (question driven)

- a world to explore (dungeons)

- challenges based on collaboration rather than action

For these reasons this genre can be considered a good experimental environment for the affective ludology experimentations and so two game levels will be created for the commercial RPG videogame *Neverwinter Nights 2* where the first will focus on the Boredom affective state and the other will stress the Flow one.

A game levels (or "Scenarios") representation mode is described in Park and Park [121] in terms of event/state/action graphs in order to minimize design anomalies while Vanhatupa [149] presents guidelines for RPG games and Noguiera *et al.*[116] consider to evaluate videogames dealing with game activities like *events* and *tasks*.

Karhulahti [83] defines mechanics and aesthetic principles to develop *Adventure* games while Thong [144] investigates effectiveness of role-playing videogames as a learning experience emphasizing *storyline* and *characters*; Horsfall and Oikonomou [74] shows that gamers prefer *strong storylines* and *character development* while also Tychsen *et al.*[145] find that *discovery & immersion* are the best motivational features to play them.

According to Brown and Cairns [26] in a videogame there are three participation phases (Engagement, Engrossment, Total Immersion) in which the concepts of Immersion and Flow result to be very close. About the design for Immersion, Calleja [31] states that *player involvement* is a prerequisite for it: a captivating story plot, for example, will not only influence the sense of *narrative involvement* but also the *affective involvement*, impacting on the quests and goals presented to the gamer (*ludic involvement*) resulting in a greater sense of "being there".

In Jennett *et al.*[79] Flow mood clearly overlaps with Immersion in the sense of *distorting time*: introducing the concept of *cognitive absorption*, Immersion is proposed as a gradual experience which involves the removal of the external environment (spatial, audio-visual and temporal) and appears evidently a precursor for the Flow. Multi-sensory virtual environments to increase Immersion are also in Chalmers *et al.*[38] while Wilcox [155] considers Immersion linked to the realism (of graphic and gameplay) reached by modern videogames.

Nacke *et al.*[112, 111] evaluate game levels for the first-person shooter *Half-Life* using a BCI and propose formal design guidelines that separate

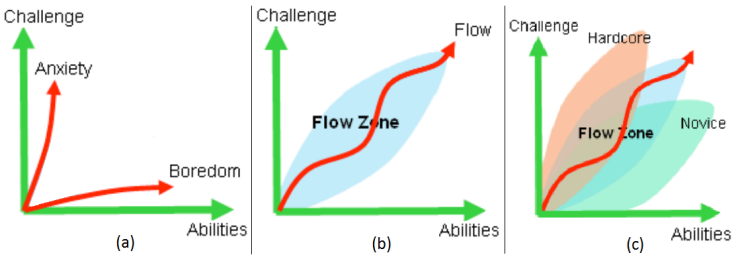


Figure 2.8: The Flow state: (a) absence of Flow for unskilled or over-skilled players; (b) Flow with challenge and abilities constant balancing; (c) different Flow zones for different players.

the *Immersion* and *Flow* by dividing the concept of *environment* from that of *combat* (which is the focus of FPS gameplay).

Considering these studies, for RPG genre Immersion can be considered as a *necessary condition* to reach the Flow state ($Immersion \subset Flow$): the importance of the visual assets becomes evident, in fact it can be considered a "theatrical stage" where characters are "actors" perform activities.

Our formalization of the following design guidelines wants to be re-usable, reproducible and independent from implementation by deliberately introducing variables and clauses easily customizable (as done in Nacke *et al.*[111]); generic definitions like "goal", "narrative/ludic element", "articulated dialogue" leave degrees of freedom that inevitably involve the experience of the game designer. The guideline sets are different but not the mere opposite, in fact each definition in one set is not always the negation of the corresponding one in the other but a variation of it.

2.4.1 A Game Level for the Boredom

Fisher [60] defines the boredom like "an unpleasant affective state with lack of concentration and difficulty during the execution of a task" while Csíkszentmihályi [53] denotes it like a state in which player's skills are greater than required (Fig. 2.8-a [40]).

This game level must be characterized by *linearity* and *repetitiveness* with poor challenge, minimal plot-story and weak visual assets; dialogues will be short and plain while allies will result unnecessary and weak.

After the previous considerations, a hypothesis of design guidelines is:

1. Given n sets of different assets like textures X_1, 3D models X_2, enemy types X_3, weapons X_4, then a game level $L = \{\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_n\}$, with $\widetilde{X}_k \subset X_k$ becomes less interesting if it contains few assets, so that $\forall \widetilde{X}_k \in L, |\widetilde{X}_k| \ll |X_k|$; the same should be for all other structural and visual assets like animations, visual effects, rooms, items, sounds and so on.

2. Assuming p as a spatial point that indicates a game level progression and given a set E of n types of enemy e characterized by constant strength (smaller than player strength), then the player challenge function $f_{ch}(e, p)$ for an encounter with an enemy of type e at a progression point p should remain constant $\forall e \in E, \forall p \in L$.

3. For each progression point p the player should be constantly rewarded with n items i that refresh the player health value and its ammunition value so that the full ammunition supply value $A = \sum_{a=0}^{n} i_a$ and the full health value $H = \sum_{h=0}^{n} i_h$ are steadily reached.

4. No real and clear gaming goals G_k are given to the player so that the victory condition goal $G_v = G_1 \wedge G_2 \wedge G_3 ... \wedge G_n = false$ doesn't allow any reward r_v.

5. Considering n characters a featuring skills of a specific class $c \in C$ ={warrior, wizard, thief, animal, unskilled} and given that a_c can join the player's ally party A, then L becomes less interesting if $|A| = n$ and L contains much less elements of A so that $L = \{a_1, ..., a_m\} \subset A, m \ll n \wedge \exists a_c, c = unskilled$.

6. Given n generic dialogues $d \in D = \{d_1, d_2, ..., d_n\}$ that a player can have with the NPCs (Non-Player Characters) of the game level L, then L becomes less interesting if contains much less elements of D so that $L = \{d_k, ..., d_m\} \subset D$ $m \ll n$; given also n questions $q \in Q = \{q_1, q_2, ..., q_n\}$ that the player can choose during an articulated dialog $d_a \in D$, then L becomes less interesting if the dialogues contain much less elements of Q so that for the most part of D is $d_a = \{q_1, ..., q_m\} \subset Q$, $m \ll n$

.

2.4.2 A Game Level for the Flow

The Flow state must be characterized by constant balance between challenge and skills (Fig. 2.8-b): if challenge becomes higher then the activity will result overwhelming generating *anxiety* while the contrary will provoke *boredom* so, to remain in a Flow state, the player must continue to learn new skills during the gaming session (Broin [25]); Flow results a mental state with total involvement and attention where skills fully meet the challenges and the player is fully absorbed by them. The notion of *Flow* is closely related to that of *Fun* but it doesn't coincide with it: Nakamura *et al.*[113] observed chess and sport players, noting that their enjoyment derives by the mere fact to accomplish their activity independently by other rewards.

Juul [80] studies the factor of *difficulty* perceived by players, arguing that nowadays videogames have become more easy; Zagal and Altizer [159] analyze mechanisms for *character progression* as a fundamental element of an RPG, while James *et al.*[77] faces the concept of *reward* in games, introducing a differentiation between personal/material/competitive. Warpefelt [152] proposes believable NPCs for player attention while Burelli [28] investigates how *camera behaviors* impact on the gaming experience.

A factor which influences the *Flow zone* is the gamer expertise which may require gameplay variations: in order to design a game for broader audiences, the in-game experience has to be not linear and static but instead it needs to offer a wide coverage of potential experiences to fit different players (i.e. the *expansion of the Flow zone* Fig. 2.8-c).

Due to all the previous considerations and with a game design perspective, there are some core elements needed to evoke the Flow:

- the game is intrinsically rewarding and the player desires to play
- the game offers the right challenge to match player's abilities
- the player feels a sense of personal control on the activities and user–interface, with immediate feedback on each action performed
- clear goals to reach and unambiguous activities to perform
- sense of alienation from the "outside world" and loss of the time-conception

In this game level, challenges involve complex dialogues and multiple goals; at least one ally helps player to accomplish profitable activities

while the level structure includes narrative elements that encourage the environmental exploration and the goals achievement.

Considering this features, a hypothesis of formal design guidelines is:

1. Given a set of indoor level parts I and a set of outdoor level parts O, the game level L should be a set union of outdoor and indoor level parts $L = \{I, O\}$.

2. Given n sets of different assets like textures X_1, visual effects X_2, animations X_3, sounds X_4, then L becomes more atmospheric and fosters imagination if $L = \{X_1, X_2, ..., X_n\}$; the same should be for all other structural and visual assets like enemy types, items, 3D models, rooms, weapons and so on.

3. Assuming a spatial point p to indicate a game level progression and given a set E of n type of enemies e, then the player challenge function $f_{ch}(e, p)$ for an encounter with an enemy e at progression point p has to progressively increase $\forall e \in E$, $\forall p \in L$.

4. For a spatial point p, after a set of progression points p_k in a game level $L = \{p_1, p_2, ..., p_n\}$, a reward type r_k from a set $R =$\{ammunition, health pack, experience points, money, spell, magic item, weapon, armor\} should be given to the player.

5. There's at least one main goal G_k so that the victory condition goal $G_v = G_1 \wedge G_2 ... \wedge G_n = true$; each achievement must lead to one or more reward r_k from a set of n rewards $R = \{r_1, r_2, ..., r_n\}$ in order to gratify the player efforts.

6. Considering n characters a featuring skills of a specific class $c \in C =$\{warrior, wizard, thief, animal\} and given that a_c can join the player's ally party A, then L becomes more interesting if $L \subseteq A$.

7. Given n meaningful dialogues $d \in D = \{d_1, d_2, ..., d_n\}$ that a player can have with the NPCs of the game level L, has to be $L \subseteq D$; given also n questions $q \in Q = \{q_1, q_2, ..., q_n\}$ that the player can choose during an articulated dialog $d_a \in D$, assuming that $\forall q_k \in Q, an_k$ is an answer, then the game level L becomes more interesting if for the most part of D is $d_a \subseteq Q \wedge an_k$ is a *narrative or ludic element* for reaching a victory condition V, achieve a reward r_k or obtain meaningful information.

(a) Single fight in a poor environment.

(b) The two weak enemies of the game level.

(c) A dialogue in plain textual mode.

(d) The tedious unskilled ally.

Figure 2.9: Boredom game level tasks.

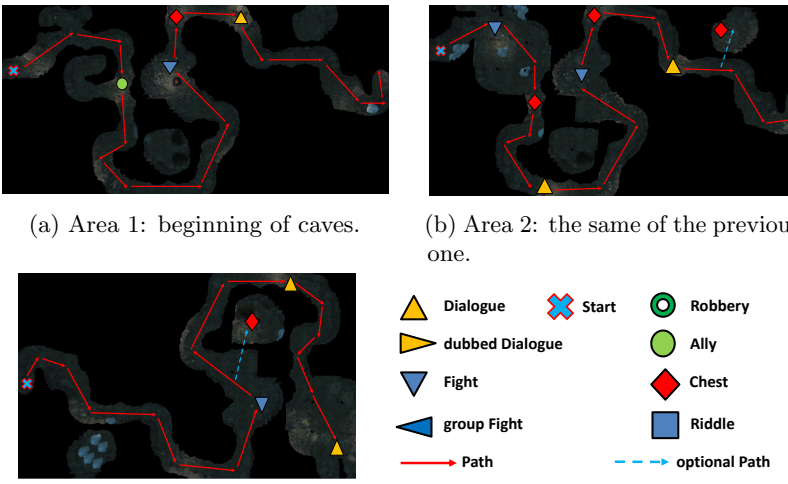
2.5 Develop for Affective Ludology

This section, referring to the formalities of the previous one, shows one of the possible implementations of the game levels (consisting of smaller pieces called "areas"); maps in Fig. 2.10 and Fig. 2.12 show player paths (also optional ones) and provide symbols to indicate the gaming activities.

2.5.1 The Boredom Game Level

Three areas have been designed by editing the *2311_tunnels* original game area that has linear guided paths and poor ambient structures.

This level consists of only one indoor cave environment, characterized by cold texture colors to emphasize ambient chilliness and repetitiveness (Fig. 2.9-a); there are no visual effects, no background music or battle sounds and every NPC has the same 3D model.



(a) Area 1: beginning of caves.

(b) Area 2: the same of the previous one.

(c) Area 3: the reverse of the previous two.

(d) Symbols and meanings.

Figure 2.10: Boredom game level areas and description of the icons.

The story plot offers a simple goal to accomplish without gratifications and the player can't upgrade its character due to the lack of experience points; challenge is minimal with constant difficulty: enemies lack resistance and combat skills and they don't attack in groups (Fig. 2.9-b). Only two weapons are available without magical skills and player's health and munitions are always kept at the maximum value: every dead enemy releases an health potion and every opened chest contains munitions and weapons (always the same two).

The unskilled ally (Fig. 2.9-d) is non-interactive (the player can not control him), unnecessary and tedious, in fact he only follows the main character around the screen using a scripted path-find algorithm without ever helping; NPCs have all the same 3D model and just some of them interact with the player using only textual dialogues (no animations or camera changes, Fig. 2.9-c) that are unbranched and useless, filled by out-of-context sentences and with few answers to choose from that do not add anything to the story plot.

(a) Group fight with allies and visual effects.

(b) A dialogue with animations and dubbed voices.

(c) Use of magic spells.

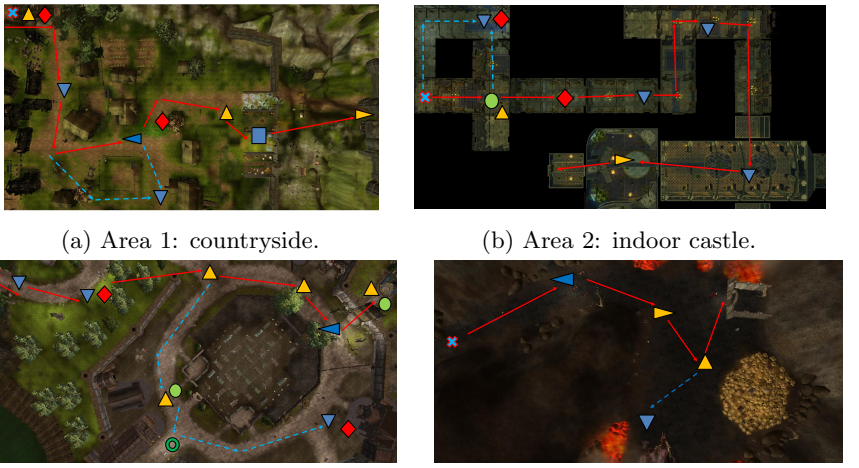
(d) Customizing skills with wearable items.

Figure 2.11: Flow game level tasks.

The first Area (Fig. 2.10-a) is linear and empty, introduces the ally, and the only dialogue creates false expectations about the prosecution of the adventure. The second one (Fig. 2.10-b) follows the same path of the previous one, increasing the repetitiveness and lowering the player's expectations. The final area (Fig. 2.10-c) forces the player to trace back along the same path as before leaving him disorientated and free to roam without any reward or real conclusion.

2.5.2 The Flow Game Level

Based on the formal guidelines, the level has been designed by linking and editing the *3010_highcliff*, *3000_castle_never*, *3063_merchant* and *3032_th_canyon* original game areas. Linking the areas the original enemies and dialogues have been replaced to create a new coherent story plot with the simplification of the paths while three original dubbed dialogues has been preserved (Fig. 2.12).



(a) Area 1: countryside.

(b) Area 2: indoor castle.

(c) Area 3: haunted village.

(d) Area 4: volcanic cave.

Figure 2.12: Flow game level map constituted by four sequential areas.

Environmental structures are diversified by indoor and outdoor areas, each of which rich of 3D models, items, visual effects (lights, animated water and trees), weapons, musics and sounds, alternation between day and night; optional paths and tasks introduce variations usually appreciated from completist gamers.

A new articulated and progressive story plot offers well-defined and satisfactory goals (rescue, enrollment, battles), while challenges steadily increase due to the variety of enemy types (with progressive better weapons and skills): previous encountered enemy types become harder to defeat and can perform group attacks and ambushes (Fig. 2.11-a) and it is possible to use spells and magical abilities (Fig. 2.11-c). Rewards are balanced with skills and challenge: only few enemies (harder to defeat) release health potions while the chests with magic objects, ammo and weapons are located only before or after the more challenging sections; wearable objects increase player's skills that can grow using the experience points (Fig. 2.11-d).

There is a party of skilled and useful allies (wizard, warrior, thief) with different ways to enlist them (automatic, dialogue, payment): they are interactive with a proactive AI and, moreover, additional non-interactive

NPCs like guards and peasants can also help the player to find secrets and accomplish profitable secondary goals.

Dialogues are useful and articulated providing useful information like memories, opinions and narrative elements (Fig. 2.11-b) while some of them influence the prosecution of the story with moral entanglements and riddles; they have animations and camera shots and three of them use professional voice-dubbing to stimulate player involvement.

An opening dialogue introduces the story assigning a main goal to the player: the first area (Fig. 2.12-a) features two combats with an helping NPC; there is a "riddle task" where the player has to interpret a sentence and choose the right path to receive a reward useful for the future. The second area (Fig. 2.12-b) presents voice-dubbed dialogues and an indoor castle setting; the background music has a fast pace and a wizard ally automatically joins the party to help the player. Area3 (Fig. 2.12-c) provides a large optional path (where it is possible to recruit the thief ally), a group fight, and a dialogue with moral entanglements with the boss enemy (that with the right ludic interaction can enjoy the player party). The next area (Fig. 2.12-d) is even more challenging and ends with a voice-dubbed moral dialogue about player's greed, where the wrong choice will lead to an impossible fight against a giant dragon (negative ending). The last fifth area is the smallest, placed in the King hall where presents the positive ending and the player's reward.

This game level has different length from the previous one and denotes more gaming time: the reason is that the only way to allow the player's abilities to grow is by offering a progressive story-plot with challenges that require time to evolve in a natural way to the player's eyes; moreover the narrative/ludic framings (as a guideline for Immersion) have been neglected by Nacke works while this work wants to include them by exploiting the RPG genre.

2.6 Tools

The videogame *Neverwinter Nights 2* belongs to the RPG genre and its choice has been taken by considering:

- available assets rich of quests, characters and ludic events
- large community with support, tools and mods

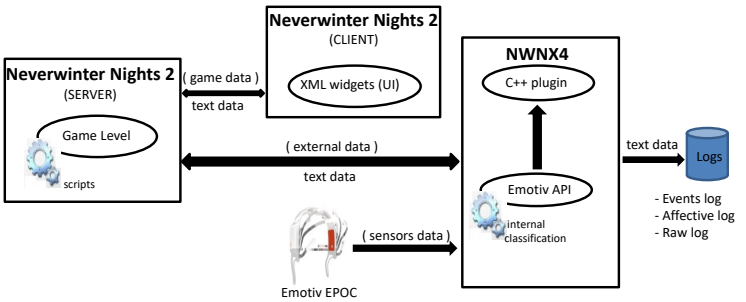


Figure 2.13: The architecture of the software system.

- a visual editor (*Electron*) for dialogues, NPCs, AI, triggers and assets
- powerful C-like internal scripting language (*NWScript*)
- XML widgets to customize the game interface

There are two types of data to extract: *internal* (game data) and *external* (EEG data). The NWNX4 [142] tool allows to deploy a C++ plugin useful for the communication between Neverwinter Nights 2 and the Emotiv headset: it is open-source and enables to implement internal software modules. Emotiv provides an SDK with programming APIs to interface the headset with an external software that shares the same programming environment: in this way a software like NWNX that wants to use affective features simply needs to include the *edk.dll* and use a *EmoEngine* to manage the *EmoState* events and data structures [5].

In Fig. 2.13 is depicted the architecture of the developed system: its aim is to automatically synchronize the internal data (environmental and character variables, gaming events, map coordinates) with the affective ones (the five punctual emotions recognized by the EPOC headset). The NWNX tool is the center of the computation: with an internal *C++ plugin* is able to interface the headset sensors (using Emotiv API) with the videogame (low–level message passing). The link between a gaming session (client side) with the NWNX plugin is the *server side* of the videogame which exploits a *NWScript* that runs inside a customized game level to recover the in-game data and to send them as a simple text message.

The final output will be the textual logs in Fig. 2.14:

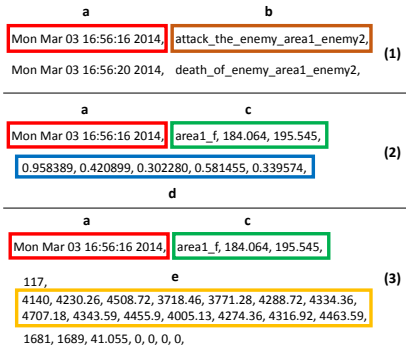


Figure 2.14: The data logs. (1) timestamp (box a) and its related event (box b); (2): timestamp, map coordinates (box c) and five affective values (box d); (3): timestamp, map coordinates and raw sensor values (box e).

Figure 2.15: The custom XML widget to print the timestamp on the screen.

- events log: shows the timestamp and the tags of the occurred gaming events (Fig. 2.14-1)
- affective log: features the name of the area, the XY coordinates and the punctual values of the emotions, pre-classified by the Emotiv algorithms in a a range between 0 and 1 (Fig. 2.14-2)
- raw log: collects the punctual values of the 14 sensors (named AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) plus the control values (Fig. 2.14-3)

A XML *UIPane* widget can be useful for screen captures since NWNX can print in real-time on the screen the same timestamp that synchronizes the data (Fig. 2.15). In addition, that logs determine a useful *implicit annotation* for these videos.

2.7 Experimental Setup

In this study about affective ludology is used a *virtual world* as research environments since it is cheap and easy to monitor and control all the variables but the proposed methodology can be extended to real-world evaluation tasks.

The goal is twofold: on the one hand, I want to demonstrate how it is possible to influence the emotions and affective states of a subject that has to perform certain formally designed tasks in a specific environment; on the other it is proposed an evaluation methodology that uses objective physiological data like the pre-classified Emotiv emotions with the aim to verify if they are valid to evaluate those manipulations.

This experiment has a two-treatment (boredom/flow) within-subject design, with "game level" as independent variable factor while the dependent variables are the five Emotiv EPOC emotions evaluated and controlled by our method (proposed in the next section). Laboratory experimentation involved 19 academic students (13 male, 6 female) which played both game levels; each subject was tested in a range hour between 10.00 AM and 07.00 PM on weekdays and the time duration of an experimental session was about one hour and half in which subjects were in a silent, empty and comfortable office room, seated on an adjustable chair and assisted by an experimenter.

A feature of the RPG is to allow the customization of the characters (or Avatar) controlled by the player both on physical level (race, clothing, hair color, sex, ...) and at capacity level by assigning scores to a set of parameters (strength, attack, dexterity, constitution, talents ...): since this was beyond the scope and timing of the experimentation, all subjects played using the same game character, choosing a non-specialized class easy to control with all statistics set to an average value to be suitable for any type of gamer; it is up to the allies implement specializations like Wizard, Thief and Warrior. In this way the variable "amount of experience and confidence with RPG videogames" can be minimized for a subject.

The first step is the *Setup*, consisting in headset and sensors placing with signal tuning; after this, a brief explanation about modalities of the study is provided followed by the compilation of a pre-questionnaire about gaming preferences and anonymous information of the subject.

The next step consists in a *Tutorial* where an experimenter helps the player to familiarize with the user-interface and to perform tasks like:

• move the character	• rotate and zoom the view
• talk with a NPC	• open a chest and take the items
• fight an enemy	• open the inventory and use items
• equip an armor	• steal an object
• equip a weapon	• open a door and enter
• update some skills	• take the control of an ally
• launch a spell	• rest to recover the health

The tutorial session and the questionnaire-fillings are considered as steps to induce relaxation and a neutral initial affective mood (similar to [148, 136, 119]) before starting a gaming session; after each session a post-questionnaire about the gaming experience has to be compiled and the same happens after the second one. To not exceed experimentation times, when a player dies the session is interrupted, but this happened few times only in advanced points of the Flow game level.

2.7.1 Description of the EEG Data

The game level can be considered as the sequence of activities that the player faces during his gaming session: in order to interpret EEG data which greatly change over time, analyzing a gaming session using its entire logs results impractical due to the their length and, furthermore, each game activity has different duration in which can be very hard to identify the precise moment where an affective stimulus appears.

Our approach is to split the logs identifying limited log sequences for each activity (where *events* represent automatically activated accidents and *tasks* are voluntary performed acts): by removing rows that do not belong to an activity designed with the guidelines, it is possible isolate them as *semantic units* more easily treatable. The data values are retrieved by crossing the *events log* with the *affective log*: the first provides the identifier and the initial time of the activity (Fig. 2.14-1) while the second one supplies its affective values, synchronized by the same timestamp (Fig. 2.14-2).

The gaming activities to be considered are:

• *dialogue* • *dubbed dialogue* • *riddle dialogue* • *chest open*	• *single fight* • *group fight* • *skills upgrade* • *stealing task*

For an activity, the *affective amount* of its time-duration is measured by the sequence of entries between the initial timestamp and the one that precedes the next activity tag (Fig. 2.16); moreover the recorded videos are a further useful help to identify the time edges, especially for dialogues.

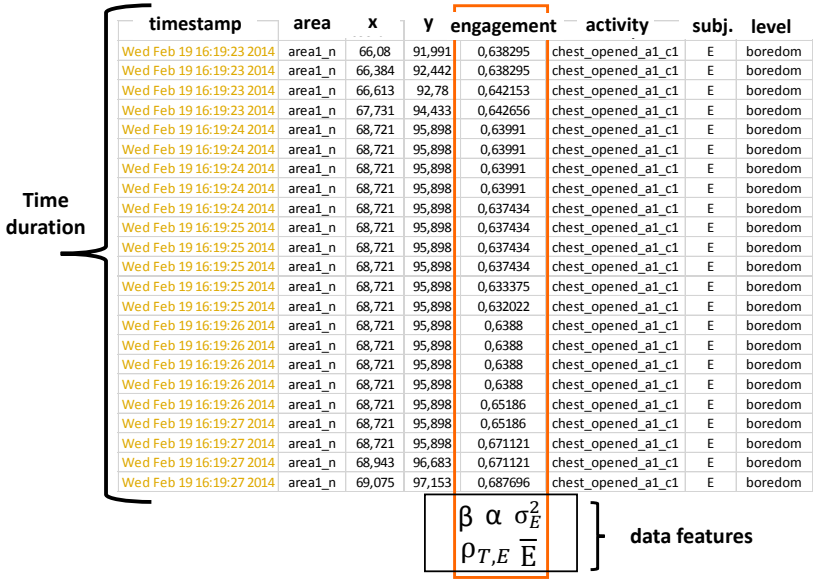


Figure 2.16: Affective log of an activity featuring the sequence of entries for its time duration and five *features* for the pre-classified "Engagement".

	a1	a2	a3	Tot.
subj.1	4	5	3	12
subj.2	5	6	4	15
subj.3	5	6	4	15
subj.4	5	6	4	15
subj.5	5	6	4	15
subj.6	5	6	4	15
subj.7	5	6	4	15
subj.8	5	7	4	16
subj.9	5	5	3	13
subj.10	5	6	4	15
subj.11	5	6	3	14
subj.12	5	8	4	17
subj.13	5	6	4	15
subj.14	5	6	5	16
subj.15	5	7	4	16
subj.16	4	4	2	10
subj.17	5	5	5	15
subj.18	5	8	2	15
subj.19	4	6	3	13
Tot.	92	115	70	277
Avg.	4.8	6.1	3.7	14.6

Table 2.1: Event count by areas and subjects: Boredom game level.

Tables 2.1 and 2.2 show the amount of activities in the two game levels for each subject: the Boredom game level presents an average of 14 activities while the Flow one offers about the double, having more event types and a different length compared to the first.

It is evident that not all subjects have experienced the same number of activities in the same areas since there may be repeated or avoided activities and the gaming sessions can have different time duration or also they can stop prematurely with a game over.

In the Boredom game level players performed on average 4.8, 6.1, and 3.7 activities for the three areas, respectively while in the Flow one they

	a1	a2	a3	a4	a5	Tot.
subj.1	6	9	13	4	0	32
subj.2	6	8	6	5	0	25
subj.3	6	10	13	5	0	34
subj.4	7	9	7	4	1	28
subj.5	6	10	8	3	1	28
subj.6	7	8	6	5	0	26
subj.7	7	9	7	1	0	24
subj.8	7	9	12	0	0	28
subj.9	7	9	7	5	0	28
subj.10	6	10	14	4	0	34
subj.11	6	9	6	4	1	26
subj.12	5	10	12	5	0	32
subj.13	6	8	7	0	0	21
subj.14	6	5	5	0	0	16
subj.15	9	7	7	4	0	27
subj.16	6	10	6	4	0	26
subj.17	10	8	5	0	0	23
subj.18	9	7	8	3	0	27
subj.19	8	8	8	4	0	28
Tot.	130	163	157	60	3	513
Avg.	6.8	8.6	8.3	3.2	0.1	27

Table 2.2: Event count by areas and subjects: Flow game level.

performed on average 6.8, 8.6, 8.3, and 3.2, activities for the four areas, respectively. Only 3 subjects (average 0.1) reached the Flow level final area with the positive ending.

2.8 Machine Learning Approach

To handle the recorded EEG affective data, three machine learning techniques that perform a supervised classification will be chosen, with the aim to classify each activity into the two states of Boredom or Flow; to train the model will be used the label associated to each activity considering the game level from which it comes.

With this method it is possible to investigate if the brain-recorded data are characteristic enough to differentiate (separate) among the two game levels: if this occurs for a substantial proportion of activities, it can be argued that player's emotions have been well manipulated during the gaming sessions and so the experimental goal will be achieved.

The first classifier is a Support Vector Machine (SVM) with a *linear kernel*: it deals with a set of points (feature vectors) x_i along with their categories y_i, for some dimension d, the $x_i \in R^d$ and the $y_i = \pm 1$; we are looking for the best separating hyperplane, defined by a possibly small set of support vectors. The equation of a decision hyperplane is:

$$< w, x > +b = 0 \tag{2.1}$$

where $w \in R^d, < w, x >$ is the inner (dot) product of w and x, and b is a real valued bias term.

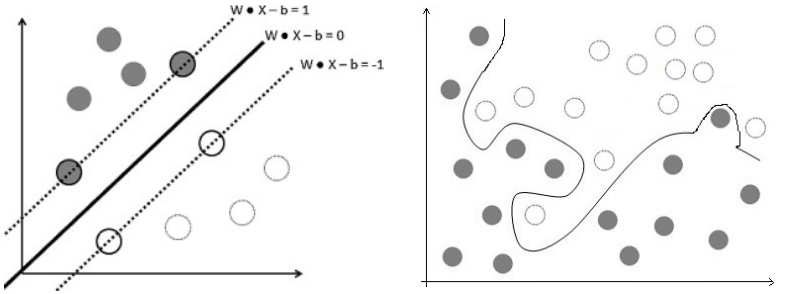
The following problem defines the best separating hyperplane: find w and b that minimize $||w||$ such that for all data points (x_i, y_i)

$$y_i(< w, x_i > +b) \geq 1 \tag{2.2}$$

The support vectors are the x_i on the boundary, those for which $y_i(< w, x_i > +b) = 1$.

A linear kernel separates the cases of Boredom and Flow categories with a plain surface in a n-dimensional feature space (Fig. 2.17-a in a two-dimensional space) but some classification problems present distributions that do not permit a simple hyperplane as a separating criterion (Fig. 2.17-b): for those problems there is a variant that retains the simplicity of SVM by applying the kernel trick, to fit the maximum-margin hyperplane in a transformed feature space: the resulting algorithm is formally similar, but every *dot product* is replaced by a *non-linear kernel* function.

In this way the transformation may be *non-linear* and the transformed space high dimensional so, although the classifier is a hyperplane in the



(a) Linear classification with the two margin split lines.

(b) Non-linear classification.

Figure 2.17: Binary classification examples in a two-dimensional feature space.

transformed feature space, it may be nonlinear in the original input space. From the theory of reproducing kernels comes a class of functions $G(x, y)$ with the property that there is a linear space S and a function ϕ mapping x to S such that the dot product takes place in the space S.

$$G(x, y) =< \phi(x), \phi(y) > \tag{2.3}$$

Some common kernels included in this class of functions are the *Polynomials* ($G(x, y) = (1 + x'y)^d$, for some positive integer d) and the *Gaussian Radial Basis Function* ($G(x, y) = e^{-\frac{1}{2\sigma^2}(x-y)'(x-y)}$, for some positive number σ).

A *Decision Tree* (or D-tree) is a classifier based on the data structure of the tree that can be used for supervised learning with a predictive modeling approaches: each internal node (split) is labeled with an input feature while the arcs that link a node to many others (children) is labeled with a condition on the input feature that determines the descending path that leads from the root node to the leaves (nodes without children).

In a *binary* tree a node can have almost two children and each leaf is labeled with a class name in a discrete set of values or with a probability distribution over the classes that predict the value of the target variable.

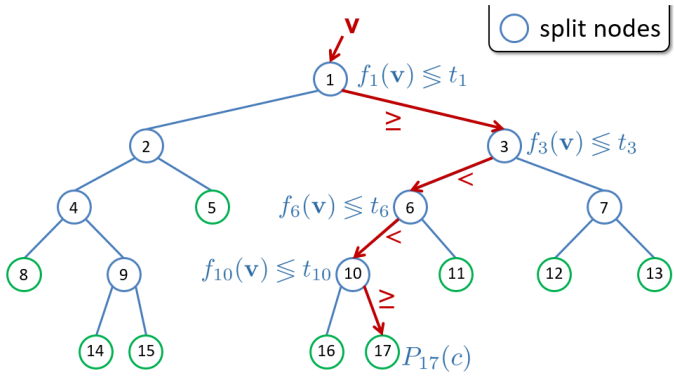


Figure 2.18: The decision process in a binary decision tree.

In this way, the decision tree classifier results characterized by:

- *nodes* (root/parent/child/leaf/split) and *arcs* (descending directed)
- no cycles between nodes
- feature vector $v \in R^n$
- split functions $f_n(v) : R^N \rightarrow R$
- thresholds $t_n \in R$
- a set of classes (labels) C
- classifications $P_n(c)$ where c is a class

Also for this classifier type the *split functions* are very important and their use is similar to what seen with SVM but each split node can also use a different one [63].

Classification (or clustering) tree analysis consists in the prediction of the class to which the data belongs ($P_n(c)$) with a process called *recursive partitioning* repeated recursively on each splitted subset (Fig. 2.18). The algorithms that navigate and build decision trees usually work top-down by choosing a value for the variable at each step that best splits the set of items.

To decide which feature to split at each node in the navigation of the tree, it is used the *information gain* value (IG) [157]; it is based on the concept of *Entropy* that from information theory is defined as:

$$H(T) = I_E(p_1, p_2, ..., p_n) = \sum_{i=1}^{n} (p_i \log_2 p_i) \tag{2.4}$$

Given p_i as the probability of an item with label i of being chosen in the split step, p_1, p_2, ... are fractions that add up to 1 and represent the percentage of each class present in the child node that results from a split in the tree.

In this way the Information Gain for a node a of the tree T is:

$$IG(T, a) = H(T) - H(T|a) \tag{2.5}$$

where $H(T)$ is the entropy for the parent node of a and $H(T|a)$ represents the weighted sum of the entropy for its children.

2.8.1 Features for Machine Learning

As seen in the design (2.4) and the develop (5.2) sections, tasks and events can be very different even if they are of the same "type" (for example a dialogue) since they depend on the game level (and area) where are located.

The five emotions pre-classified by the Emotiv headset can be useful to describe the affective mood of an activity and to use it in a machine-learning classifier: this can be done by computing for each of them a numerical *feature*, that is a single value which summarizes the entire sequence of entries in the *affective log* recognized during the time duration (Fig. 2.16) of the activity.

The features calculated for each of the five punctual emotions are:

- the *angular coefficient* β and the *intercept* α of a regression line, calculated considering the time T as a positive variable which constantly increases on the x-axis; it gives indications on the presence of a trend for the emotion E

  $$E_i = \alpha + \beta T_i + \mu_i, i : 1..n \tag{2.6}$$

  where μ_i is the statistical error

- the *Pearson product-moment correlation coefficient* between time T and the emotion E: a value $r_e > 0.7$ evidences a strong local correlation that can be direct (positive sign) or inverse (negative sign).

$$\rho_{T,E} = \frac{cov(T,E)}{\sigma_T \sigma_E}, -1 \leq \rho_{T,E} \leq +1 \tag{2.7}$$

  In particular, it is used the *Pearson correlation coefficient* r_e calculated on the n affective values contained in the time interval:

$$r_e = \frac{\sum_{i=1}^{n}(T_i - \bar{T})(E_i - \bar{E})}{\sqrt{\sum_{i=1}^{n}(T_i - \bar{T})^2}\sqrt{\sum_{i=1}^{n}(E_i - \bar{E})^2}}, \quad -1 \leq r_e \leq +1 \tag{2.8}$$

- the *arithmetic mean* $\overline{E}$ and the *variance* σ_E^2 that give quantitative information on the affective values of emotion E

The fact that each game level has been designed and developed with specific formal guidelines permits to label as "Flow" or "Boredom" each of their activity depending on the membership: in this way the dataset results implicitly annotated and can be used as a *ground-truth* for a classifier.

Considering the 19 subjects (13 male and 6 female) results a 790x25 predictors matrix M_0 where rows represent the labeled activities (277 Boredom and 513 Flow) and columns their relative 25 features (five for each emotion); since the ground-truth is unbalanced, the Boredom cases must be increased by randomly sampling 236 examples from the original ones obtaining the final M_F with dimensions 1026x25.

Since the ground-truth cases are still too few for a machine learning approach, the classifiers will use the *leave-one-out* setting which trains the system n times iteratively using $n-1$ cases and testing it with the one left out.

2.8.2 Classification Results

The use of non-linear kernel function in a SVM classifier brings a marked performance improvement since accuracy increases from 70% linear to 92.5% rbf (radial basis function), highlighting a strong polarization between the two classes (see Tables 2.4 and 2.4) while all metrics of rbf results over 90% with *f1-score* almost the same for both classes.

	Boredom	Flow	Tot.
original events	513 (ext.)	513	1026
classified events	507	519	1026
true positives	492	494	986
false negatives	25	15	40
recall	95.2%	97%	
precision	97%	95.2%	
f1-score	96%	96%	
accuracy			96%

Table 2.3: Decision Tree classification results.

	Boredom	Flow	Tot.
original events	513(ext.)	513	1026
classified events	542	484	1026
true positives	489	460	949
false negatives	24	53	77
recall	95.3%	89.7%	
precision	90.2%	95%	
f1-score	92.7%	92.3%	
accuracy			92.5%

Table 2.4: SVM-rbf classification results.

The Decision Tree method has the best performance (Table 2.3): it is able to correctly separate 96% of the gaming activities (accuracy) and all the other metrics are over 95% for both classes.

From this percentages it results confirmed the correctness of the design and development phases by following the guidelines and, moreover, the validity of the proposed methodology applied with the features calculated on the five emotions pre-classified by the Emotiv EPOC headset.

	Boredom	Flow	Tot.
original events	513 (ext.)	513	1026
classified events	612	414	1026
true positives	409	310	719
false negatives	104	203	307
recall	79.7%	60.4%	
precision	66.8%	74.9%	
f1-score	72.7%	66.9%	
accuracy			70%

Table 2.5: SVM-linear classification results.

2.8.3 Classification for Game Level Areas

This classification process considers to which areas the correctly classified activities belong: by looking at what are the level parts that have the majority of the recognized events, it is possible to understand what are the best-designed areas of a game level designed for affective ludology.

In the following, for each of the three classification methods are still reported all the results but, for brevity, only the results of the best classifier (Decision Tree in Table 2.6 and 2.9) are analyzed. Results for SVM-rbf are in in Table 2.7 (Boredom) and in 2.10 (Flow), while for the SVM-linear are in Table 2.8 (Boredom) and in 2.11 (Flow).

The over 92% good classification for boredom areas suggests that they were recognized as repetitive and tedious: from Area1 to Area2 there is a predictable increment due to the repetitiveness. It is remarkable that while the first two areas have the same structure, the third reduces the perceived boredom introducing small but significant variations; moreover, the structure of Area2 features the largest number of activities (211) while the third area has the fewest (126).

For the flow game level, the classification performance varies but it is always above 95%: it grows from 95.4% for the first area up to 100% for the fourth one; since the third area is the largest and contains many optional activities, can be observed a slight decrease with respect to the typical progression associated to the Flow affective state (in results of the other classifiers the growing is more noticeable). The fifth smallest area has only one dialogue and is not significant since it was reached by only 3 subjects.

	total (ext.)	correct	rate
Area 1	176	170	96.6%
Area 2	211	206	97.6%
Area 3	126	116	92.1%
Tot.	513	492	95.9%

Table 2.6: Mood-congruent events by area: Boredom game level (D-tree)

	total (ext.)	correct	rate
Area 1	176	169	96%
Area 2	211	201	95.3%
Area 3	126	119	94.4%
Tot.	513	489	95.3%

Table 2.7: Mood-congruent events by area: Boredom game level (SVM-rbf)

	total (ext.)	correct	rate
Area 1	176	154	87.5%
Area 2	211	167	79%
Area 3	126	88	70%
Tot.	513	409	79.7%

Table 2.8: Mood-congruent events by area: Boredom game level (SVM-linear)

	total (ext.)	correct	rate
Area 1	130	124	95.4%
Area 2	163	157	96.3%
Area 3	157	151	96.2%
Area 4	60	60	100%
Area 5	3	2	66.7%
Tot.	513	494	96.3%

Table 2.9: Mood-congruent events by area: Flow game level (D-tree)

	total (ext.)	correct	rate
Area 1	130	111	85.4%
Area 2	163	143	87.7%
Area 3	157	145	92.4%
Area 4	60	58	96.7%
Area 5	3	3	100%
Tot.	513	460	89.7%

Table 2.10: Mood-congruent events by area: Flow game level (SVM-rbf)

	total (ext.)	correct	rate
Area 1	130	54	41.5%
Area 2	163	99	60.7%
Area 3	157	108	68.8%
Area 4	60	43	71.7%
Area 5	3	3	100%
Tot.	513	307	59.8%

Table 2.11: Mood-congruent events by area: Flow game level (SVM-linear)

2.8.4 Classification for Activity Types

Another useful analysis is to inquire what are the best classified activity types, i.e. if there are activities better characterized (designed and developed) by others.

Also in this section, for each of the three classification methods are reported all the results but only those for the Decision Tree classifier (in Table 2.12 and 2.15) are discussed. Results for SVM-rbf are in in Table 2.13 (Boredom) and 2.16 (Flow), while for the SVM-linear are in Table 2.14 (Boredom) and 2.17 (Flow).

In the boredom game level, dialogues are about 1/3 more numerous than other types and results the best classified (97.3%); the classification rating is always over 94% and evidences a correct development phase and sound guidelines supported by the poorness and repetitiveness.

The flow game level denotes activities specially designed and developed for it and so *stealing action*, *skills upgrade*, *riddle dialogue* and *dubbed dialogue* are 100% but the first three have a low number of examples due also to their optionality. In this game level the *single fight* (96.1%) is the best classified followed by the simple *dialogue* (95.6%) that has a number comparable to it; it is remarkable that all classification ratings are over 94%.

Regarding the results, it must be noticed that they significantly varies depending on the chosen classifier and even the difference of 3.5% that is between Decision Tree and SVM-rbf is enough to change the results: for example considering the boredom game level, Area1 results the best classified for SVM-rbf while for the D-tree the best is Area2; moreover considering the flow game level *group fight* results one of the best classified activity for SVM-RBF classifier while it results the worst for the D-tree one.

	total (ext.)	correct	rate
dialogue	224	218	97.3%
single fight	136	129	94.8%
chest opened	153	145	94.8%
Tot.	513	492	95.9%

Table 2.12: Mood-congruent events by type: Boredom game level (D-tree)

	total (ext.)	correct	rate
dialogue	224	217	96.9%
single fight	136	125	91.9%
chest opened	153	147	96.1%
Tot.	513	489	95.3%

Table 2.13: Mood-congruent events by type: Boredom game level (SVM-RBF)

	total (ext.)	correct	rate
dialogue	224	182	81.5%
single fight	136	107	78.7%
chest opened	153	120	78.4%
Tot.	513	409	79.7%

Table 2.14: Mood-congruent events by type: Boredom game level (SVM-linear)

	total (ext.)	correct	rate
dubbed dialogue	46	46	100%
riddle dialogue	18	18	100%
dialogue	159	152	95.6%
single fight	153	147	96.1%
fight vs. a group	37	35	94.6%
chest opened	85	81	95.3%
skills upgrade	11	11	100%
stealing action	4	4	100%
Tot.	513	494	96.3%

Table 2.15: Mood-congruent events by type: Flow game level (D-tree)

	total (ext.)	correct	rate
dubbed dialogue	46	46	100%
riddle dialogue	18	18	100%
dialogue	159	145	91.2%
single fight	153	130	85%
fight vs. a group	37	35	94.6%
chest opened	85	71	83.6%
skills upgrade	11	11	100%
stealing action	4	4	100%
Tot.	513	460	89.7%

Table 2.16: Mood-congruent events by type: Flow game level (SVM-RBF)

	total (ext.)	correct	rate
dubbed dialogue	46	40	89.9%
riddle dialogue	18	15	83.3%
dialogue	159	91	57.2%
single fight	153	89	58.2%
fight vs. a group	37	22	59.5%
chest opened	85	40	47.1%
skills upgrade	11	9	81.8%
stealing action	4	4	100%
Tot.	513	310	60.4%

Table 2.17: Mood-congruent events by type: Flow game level (SVM-linear)

2.9 Inferential Statistics

To give a better interpretation of results that takes into account not only the quantitative observations, it is conducted an inferential statistic test with the aim to prove if is the choice of the classifier that determines improvements. For each classifier, each gaming activity is assigned to its experimental subject: tables 2.18, 2.19 and 2.20 show for each subject, the sum of the activities (Boredom+Flow), the sum of the congruent classified and its percentage.

From previous sections we can consider valid the use of the pre-classified Emotiv emotions since each classifier gave remarkable numerical results.

For the statistical test we must inquire:

- H0 (null hypothesis): variations in results are not influenced by the choice of the classifier (equal medians)
- H1: reject H0, differences between classifiers are significant (because of their choice)

To perform the statistical test it is used the non-parametric *Wilcoxon signed rank-sum*, featuring independent observations paired with each other and no hypothesis about the distribution of their values; the parameters for this test are h and p: if h=0 then H0 is *accepted* while at the default 5% significance level, if $p < 0.05$ then H0 is *rejected*.

Taking account of the three classifier used, the cases and the correspondent results are:

• SVM-linear vs. SVM-rbf	• h=0,p=0.08→not reject H0
• SVM-linear vs. D-tree	• h=1,p=0.02→reject H0
• SVM-rbf vs. D-tree	• h=0,p=0.19→not reject H0

The inference test states that for Case2 (SVM-linear vs. Decision Tree) with a statistical validity the classification differences are determined by the choice of the classifier: this confirms the validity of introducing the Decision tree in this study despite having already the SVM-rbf an increase of 20% from the SVM-linear one.

Subject	Classified	Congruent	%
subj.1	48	46	95.8%
subj.2	49	48	98%
subj.3	59	56	94.9%
subj.4	61	59	96.7%
subj.5	52	50	96.1%
subj.6	55	54	98.2%
subj.7	48	46	95.8%
subj.8	56	53	94.6%
subj.9	52	51	98.1%
subj.10	64	62	96.9%
subj.11	55	55	100%
subj.12	66	62	93.9%
subj.13	49	45	91.8%
subj.14	41	39	95.1%
subj.15	62	59	95.2%
subj.16	47	45	95.7%
subj.17	48	47	97.9%
subj.18	57	53	93%
subj.18	57	56	98.2%
Tot.	1026	986	

Table 2.18: Classification for subjects: Decision tree.

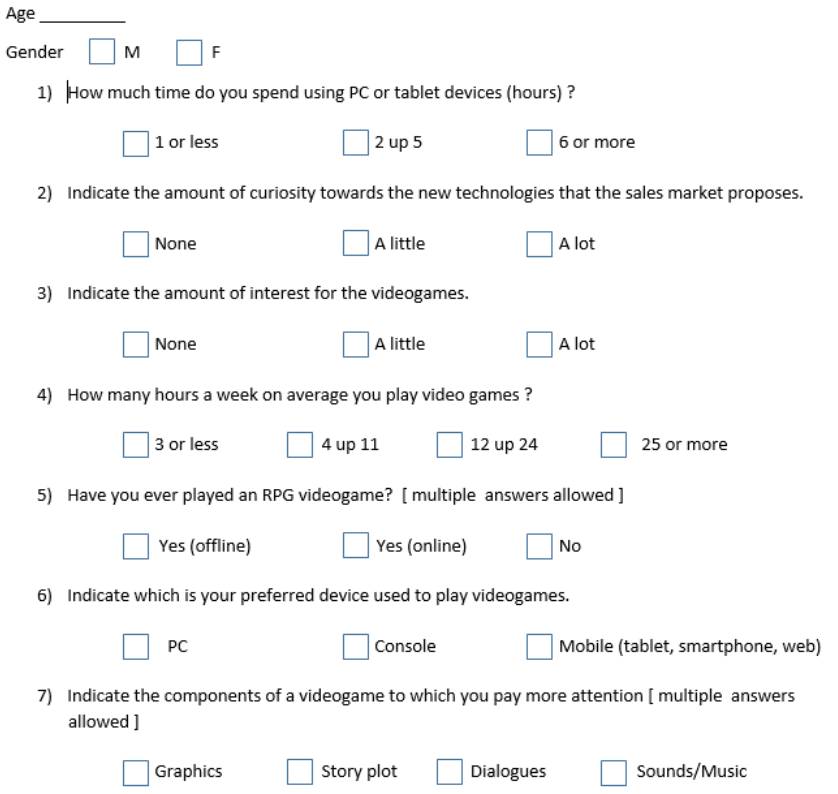
Subject	Classified	Congruent	%
subj.1	48	42	87.5%
subj.2	49	42	85.7%
subj.3	59	58	98.3%
subj.4	61	52	85.2%
subj.5	52	50	96.1%
subj.6	55	53	96.4%
subj.7	48	46	95.8%
subj.8	56	56	100%
subj.9	52	48	92.3%
subj.10	64	60	93.7%
subj.11	55	53	96.4%
subj.12	66	59	89.4%
subj.13	49	48	98%
subj.14	41	39	95.1%
subj.15	62	54	87.1%
subj.16	47	44	93.6%
subj.17	48	44	91.7%
subj.18	57	53	93%
subj.19	57	48	84.2%
Tot.	1026	949	

Table 2.19: Classification for subjects: SVM-rbf.

Subject	Classified	Congruent	%
subj.1	48	35	72.9%
subj.2	49	29	59.2%
subj.3	59	48	81.4%
subj.4	61	47	77%
subj.5	52	41	78.8%
subj.6	55	41	74.5%
subj.7	48	34	70.8%
subj.8	56	48	85.7%
subj.9	52	37	71.1%
subj.10	64	49	76.6%
subj.11	55	34	61.8%
subj.12	66	43	65.1%
subj.13	49	37	75.5%
subj.14	41	30	73.2%
subj.15	62	39	62.9%
subj.16	47	25	53.2%
subj.17	48	28	58.3%
subj.18	57	41	71.9%
subj.19	57	33	57.9%
Tot.	1026	719	

Table 2.20: Classification for subjects: SVM-linear.

2.10 Subjective Data: Questionnaires


Age _____

Gender ☐ M ☐ F

1) How much time do you spend using PC or tablet devices (hours) ?

☐ 1 or less ☐ 2 up 5 ☐ 6 or more

2) Indicate the amount of curiosity towards the new technologies that the sales market proposes.

☐ None ☐ A little ☐ A lot

3) Indicate the amount of interest for the videogames.

☐ None ☐ A little ☐ A lot

4) How many hours a week on average you play video games ?

☐ 3 or less ☐ 4 up 11 ☐ 12 up 24 ☐ 25 or more

5) Have you ever played an RPG videogame? [multiple answers allowed]

☐ Yes (offline) ☐ Yes (online) ☐ No

6) Indicate which is your preferred device used to play videogames.

☐ PC ☐ Console ☐ Mobile (tablet, smartphone, web)

7) Indicate the components of a videogame to which you pay more attention [multiple answers allowed]

☐ Graphics ☐ Story plot ☐ Dialogues ☐ Sounds/Music

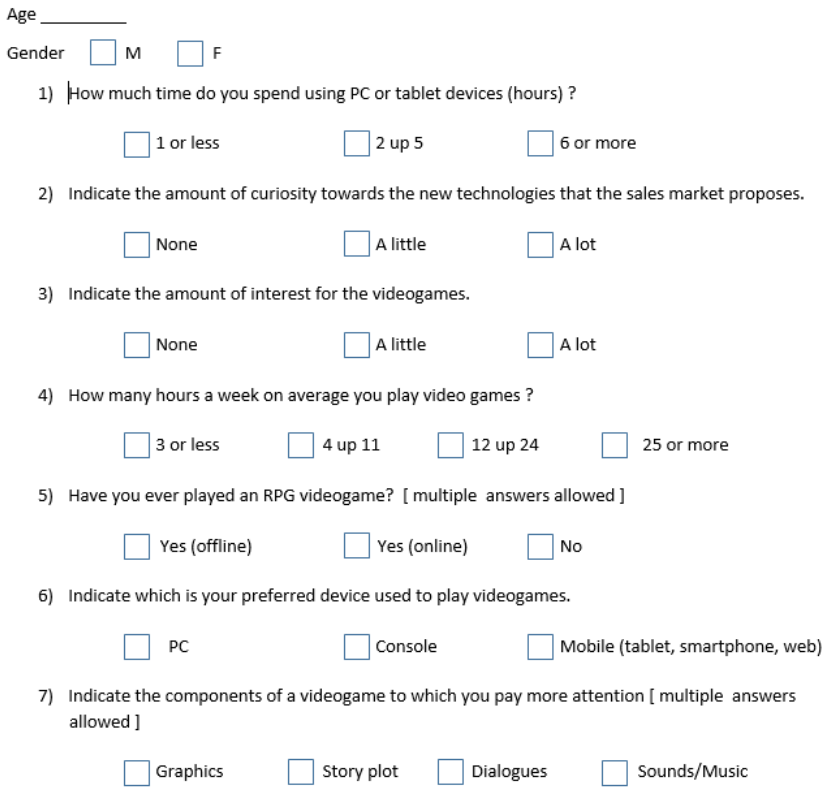
Figure 2.19: The pre-questionnaire module for gaming preferences.

Questionnaires give a subjective contribution (opposite to the objective ones given by the physiological data) to this study: they have a non-standard format specially developed to report a complementary view for this study, with a quantitative analysis of the player experience based on the frequency of the closed answers.

The pre-questionnaire (Fig. 2.19) presents general information about the subject and his gaming preferences: it results that 52.6% of subjects

uses PCs for 2-5 hours daily and 36.8% for more than 6 hours while the time spent to play videogames is between 4-11 and 12-24 hours per week for this group. Interest about videogames is *a lot* for 84.2% of subjects and 78.9% of them has previously played an RPG one (offline or online).

Considering the gaming platforms, 68.4% prefers PC, 21% game consoles and 10.5% mobile platforms; about technical aspects, *plot* is important for 84.2%, 52.6% prefers *graphic*, 31.6% *dialogues* and only 21% *audio* assets (multiple answers were allowed).


Age _____

Gender ☐ M ☐ F

1) How much time do you spend using PC or tablet devices (hours) ?

☐ 1 or less ☐ 2 up 5 ☐ 6 or more

2) Indicate the amount of curiosity towards the new technologies that the sales market proposes.

☐ None ☐ A little ☐ A lot

3) Indicate the amount of interest for the videogames.

☐ None ☐ A little ☐ A lot

4) How many hours a week on average you play video games ?

☐ 3 or less ☐ 4 up 11 ☐ 12 up 24 ☐ 25 or more

5) Have you ever played an RPG videogame? [multiple answers allowed]

☐ Yes (offline) ☐ Yes (online) ☐ No

6) Indicate which is your preferred device used to play videogames.

☐ PC ☐ Console ☐ Mobile (tablet, smartphone, web)

7) Indicate the components of a videogame to which you pay more attention [multiple answers allowed]

☐ Graphics ☐ Story plot ☐ Dialogues ☐ Sounds/Music

Figure 2.20: The post-questionnaire module about the player experience.

2.10.1 Post-Questionnaires

The post-questionnaire (Fig. 2.20) whose data are in Tables 2.21 and 2.22 presents closed-answer questions with only one choice allowed and has been compiled by each subject two times (after each gaming session).

About player's general *Satisfaction*, for Boredom game level 78.9% of subjects is polarized towards low values (*a little*) while for the Flow one 52.6% expresses *a lot*, 47.4% *little* and nobody chooses *none*. This results reflect expectations for the first level and are reasonable for the Flow one.

The proposed *Allies* are evaluated in Question 2: the 94.7% of subjects dislikes the one in Boredom game level while, for the second game level they are divided between positive and negative judgment (56.6% versus 47.4%) bringing the need to improve this aspect in the development phase.

Question 3 evaluates *Dialogue* satisfaction: subjects dislike them for Boredom game level (47.4% *none* and 52.6% *little* acceptance) while, for the Flow one is almost the opposite with 57.9% *a lot* and 42.1% *a little*.

By considering the perceived *Boredom*, in Question 4 subjects express *enough* (52.6%) and *a lot* (10.5%) for Boredom game level while for the other one 21% expresses *none* and 57.9% *a little*: this polarization confirms the goodness of the proposed guidelines.

About the perceived *Frustration*, for Boredom game level subjects are almost divided between positive and negative judgment (57.9% versus 42.1%) while in the Flow one frustration results absent or very low for 89.5% of them: by considering also Question 4 results, this emphasizes a connection between perceived boredom and the frustration that comes when a gamer is forced to play something that dislikes.

	None	A little	Enough	A lot
Satisfaction	2	15	-	2
Allies satisf.	10	8	-	1
Dialogues satisf.	9	10	-	0
Boredom perc.	0	7	10	2
Frustration perc.	5	6	7	1
Involvement perc.	3	11	5	0

Table 2.21: Post-questionnaires - Boredom game level

	None	A little	Enough	A lot
Satisfaction	0	9	-	10
Allies satisf.	1	9	-	9
Dialogues satisf.	0	8	-	11
Boredom perc.	4	11	4	0
Frustration perc.	3	14	2	0
Involvement perc.	0	4	14	1

Table 2.22: Post-questionnaires - Flow game level

Question 6 treats the *Involvement* perceived by players, for which the majority is polarized between negative values for Boredom game level (73.7%) and positive for the Flow one (78.9%): this results fully confirm the decisions taken in both design and development phases.

Chapter 3

Interactive Tools and Medical Image Analysis

The interest of biomedical and computer vision communities in acquisition and analysis of epidermal images has been increased during the last decades: the possibility to automatically classify early melanomas is investigated because the malignant melanoma is one of the most common and dangerous skin cancer and 100,000 new cases with over 9,000 deaths are diagnosed every year by only considering the USA [55, 47]; in this context automated system for fast and accurate melanoma diagnosis are well accepted, also considering technical approaches from standard machine learning methods up to the Convolutional Neural Networks (CNN).

In this context, a key element is the availability of user-friendly annotation tools that can be used by non-IT experts to produce well-annotated and high-quality medical data; systems that allow fast and accurate image acquisition and lesion investigation and classification are well accepted in the biomedical community considering that the diagnostic accuracy with trained clinicians has reached about 75-84% (Kittler *et al.*[85]).

High-resolution images and their annotated data, combined with analysis pipelines and machine learning techniques, represent the base to develop diagnostic recommendation systems that are intelligent, automated and proactive. Traditionally, clinical experts manually categorize and examine printed medical images so it becomes a very time consuming task; to avoid

this issue, advanced user-friendly annotation tools must be developed, in order to improve the digital libraries in biomedical and related fields in size and quality of annotated data that will be processed by machine learning and pattern recognition techniques.

The usability of the tools, with also the assurance that their hardware devices are medical-compliant, is a fundamental element due to user's lack of deep IT-skills.

3.1 Approaches to Skin Image Analysis

A complete and rich survey about lesion borders is in Celebi *et al.*[37]; the first step for skin inspection is the acquisition of digital images: the main techniques involve *Epiluminence Microscopy* (ELM, or dermoscopy), *Transmission Electron Microscopy* (TEM) and the acquisition through standard RGB cameras (Maglogiannis *et al.*[102]).

In last decades, standard video devices are commonly used for skin lesion inspection systems, in particular in the telemedicine field (Singh *et al.*[140]). However, these solutions present some issues, like low camera spatial resolution (melanoma or other skin details can be very small) and distortions caused by the camera lenses. Moreover, variable illumination conditions can strongly deteriorate the quality of acquisitions (Ali *et al.*[11]).

After the digital acquisition, the second step consists in the analysis and investigation of the epidermal images acquired: several works in literature have been proposed for automated epidermal image analysis, in order to support biomedical experts; most of them are based on Computer Vision approaches, typically combining low-level visual features representation, image processing techniques and machine learning and pattern recognition algorithms as in Codella *et al.*[47].

The work of Seidenari *et al.*[137] provides an overview about the finding of melanomas by image analysis, Wighton *et al.*[154] present a general model using supervised learning and MAP estimation that is capable of performing many common tasks in automated skin lesion diagnosis; Celebi *et al.*[58] treat lesion border detection with thresholding methods while Fan *et al.*[59] that use saliency combined with Otsu thresholding.

The Peruch *et al.*[122] study faces lesion segmentation through Mimicking Expert Dermatologists Segmentations (MEDS) and in Liu *et al.*[96] propose an unsupervised sub-segmentation technique for skin lesions. Codella *et*

al.[46] manually pre-segment images, already cropped around the region of interest, used in conjunction with hand-coded and unsupervised features to achieve state-of-the-art results in melanoma recognition task with a dataset of 2,000 skin images; specifically, a combination of sparse coding, deep learning and SVM learning algorithms are exploited.

Mendonca [109] and Batara *et al.*[19] compare several machine learning methods like SVMs and K-nearest neighbors (kNNs), based on color, edge and texture descriptors while learning approaches and deep learning techniques have been exploited by Schaefer *et al.*[135] and Rastgoo *et al.*[130] in literature; the combination of hand-coded features extractors, sparse-coding methods, HOGs and SVM (Bakheet [17]) and deep learning techniques are also used focusing on melanoma recognition and segmentation tasks in dermoscopy and dermatology domain [47].

In 2016 a set of challenges has been presented in the paper *Skin Lesion Analysis toward Melanoma Detection* [69]: the aim is to use one of the most complete dataset of medical images, collected by the *International Skin Imaging Collaboration* (ISIC) and obtained with the aggregation of dermoscopic datasets coming from multiple institutions, to test and evaluate techniques for segmentation and diagnosis. The best scores in classification and segmentation has been achieved using deep learning approaches (He *et al.*[71], Long *et al.*[97]): the ISIC database has been exploited in Yuan [158] with a 19-layer deep convolutional neural network while classification and segmentation goals are achieved using CNN approaches also in Majtner [103] and Baweja [20].

Before the ISIC challenge, the exploitation of deep learning techniques was partially bounded by the limited size of datasets present in literature: the amount of training data and high quality annotations are in fact key aspects for deep learning approaches (Krizhevsky *et al.* [87]).

3.2 MIMS: a Medical Image Management System

In Fig. 3.1 it is proposed an architecture for a heterogeneous integrated system to store and manage medical images and the derived data.

The system core is a *Digital Library* which stores and organizes the dataset acquired from extern hospital equipment: it consists of 436 dermoscopic skin images in standard JPEG format with a high spatial resolution of 4000x2664 or 3000x4000 pixels. The *dermatoscope* is a medical device which consists of a magnifier, a non-polarised light source and a liquid medium between the instrument and the skin, and that allows inspection of skin lesions avoiding surface reflections.

The mobile-oriented *Annotation Tool* represents the first step that has to be designed and developed: the aim is to enable the annotation of images by domain experts like dermatologists, allowing the storage of *primitive features* 3.2 like strokes, colors and pen sizes where each color can denote specific semantics given by dermatologists.

In this way will be possible to extract the *derived features* like contours, shapes, intersections, color features and numerical values, as shown in Figure 3.3. that can be used in the *Machine Learning module* that supports a *Recommendation Module* for diagnosis that distinguish between benign and malignant (melanomas) lesions. Finally, in the architecture there are a *Information Visualization module* for various and dynamic data visualization and a specific *Query module* to permit search facilities [10].

A consolidated standard standard for handling, storing, printing, and transmitting information in medical imaging is the DICOM (Digital Imaging and Communications in Medicine, ISO standard 12052:2006) of which the National Electrical Manufacturers Association (NEMA) holds the copyright. This standard has been widely adopted by hospitals and other physician offices: it includes a file format definition and a network communications protocol which uses TCP/IP to communicate between systems and enables the integration of medical imaging devices like scanners, servers, workstations, printers and network hardware from different manufacturers. A DICOM object consists of attributes (such as name, Id, datetime) with a special one containing the pixel data: in this way the medical image contains for example the patient ID within the file so that the image can never be separated from this information (similarly to the JPEG format

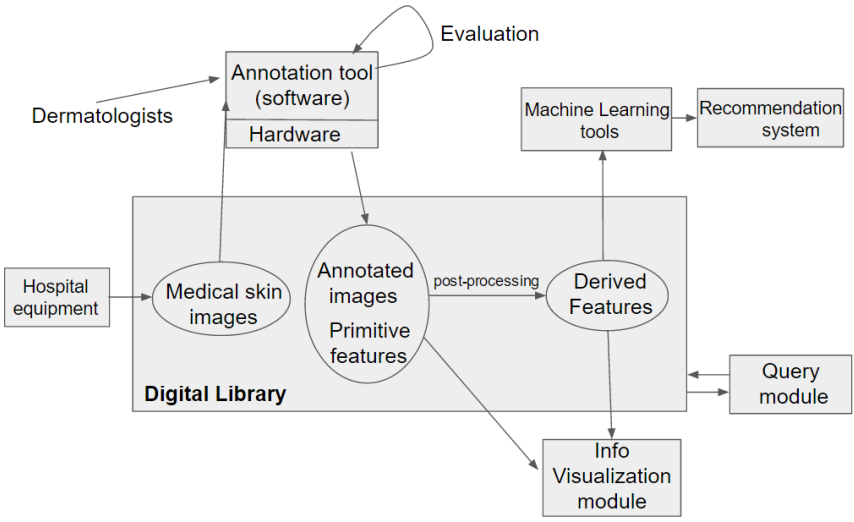


Figure 3.1: The proposed architecture for MIMS.

that embeds tags to describe an image).

The main goal is the retention of an annotations in the form of a new enriched images by indexing each drawn stroke whose points correspond to the image pixels. The organization of this data is made in form of specific folders and file names, composed to incorporate and identify the resources by using the annotator username with the path and the name of the original JPEG file.

The primitive features are:

1. a new color image, featuring all the draw strokes (primitive features)

2. a binary (black and white) image for each color used to drawn the strokes (each image assembles the strokes of the same color)

3. a text file with the features of each stroke (list of coordinate points, color code and pen size): this file is used to rebuilt and reload a saved annotation

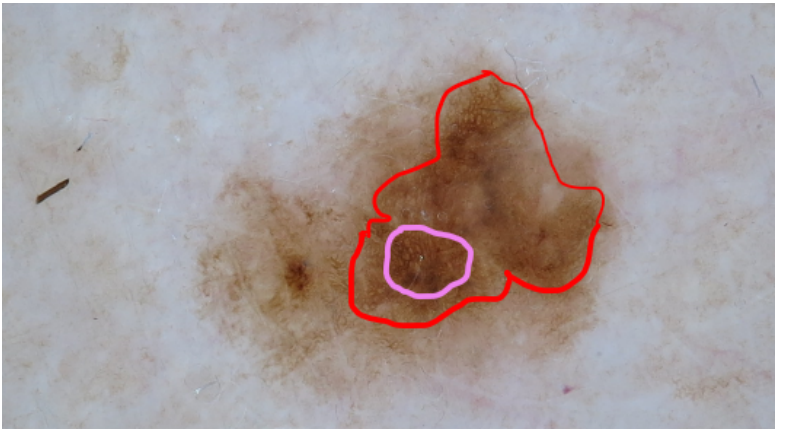


Figure 3.2: A dermoscopic skin image annotated with strokes of different colors and width (primitive features).

3.2.1 Developing the Annotation Tool

Since the tool must be mobile-oriented, as hardware platform has been chosen the *Microsoft Surface Pro 3* (Fig. 3.4), a portable tablet device powerful but less invasive than a modern PC that can be used in mobility into a medical environment like an hospital. Dermatologists can annotate with the assurance that the only strokes recognized on the touch screen are those which come from using the specific *Surface Pen*, avoiding unwanted strokes coming from touch gestures or oversight movements.

To acquire the necessary data by a large group of non-IT subjects, several principles of usability and Human-Computer Interaction have to be followed bearing in mind the following [139]:

- final users are domain experts that could be unfamiliar with technical tool or data organization and analysis
- physician are usually overworked so they do not have much time to skill themselves on externals tools
- physician and specialists like dermatologists have peculiar working protocols and pipelines, so the tool must be non-invasive with the aim to not impact on their daily activities

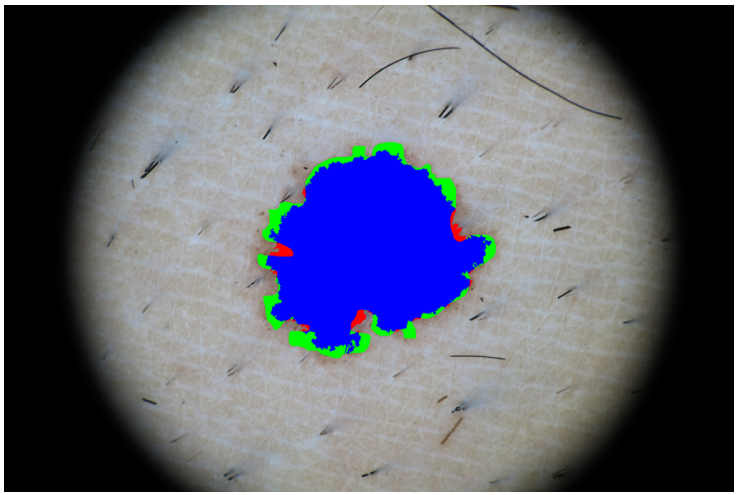


Figure 3.3: Examples of derived feature: an automatic area (red) made by the system and the intersection area (blue), obtained by crossing with the one (green) constructed by a manual annotation.

- the image annotation task must be as much as possible fast and user-friendly for a dermatologist, imitating what they would do naturally
- the medical environment has peculiar safety and security requirements so, in addition to the software tool, the hardware introduced in this areas must be considered

The task of *annotate* with a precise technological pen on a glass screen is natural and intuitive and shows the characteristic of *Affordance* [64]: in the field of psychology the term includes all actions that are physically possible on an object or environment (like what a physician wants to performs on a real medical printed photo). In a general way, when the concept is applied to design and develop activities it refers to those action possibilities that the user is aware of, also considering the available or permissible peripherals that allow such interactions. The *perceived affordance* refers not only to the user physical capabilities but also on his goals, beliefs, and past experiences [118] so that an object (real or virtual) can naturally 'suggests' how to interact with it.

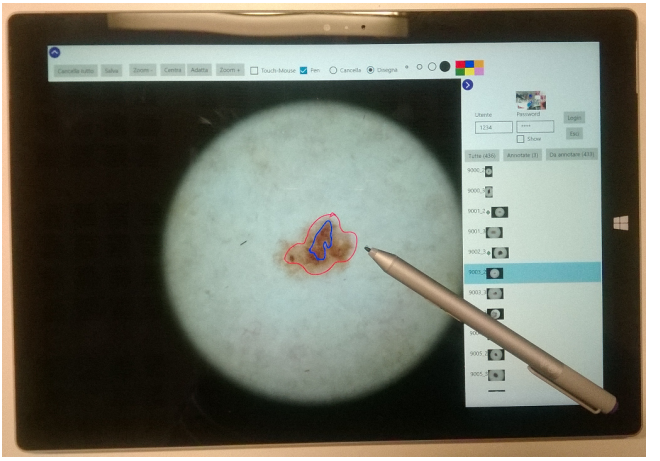


Figure 3.4: The *Microsoft Surface Pro 3* is the hardware platform chosen for the annotation tool that allows portability and error-free annotations.

Another usability aspect to take account is the *direct manipulation* that, formally, is an *interaction style* in which users act on displayed objects of interest using physical, incremental, reversible actions whose effects are immediately visible on the screen [138]: gestures on a touch screen are *to pitch* (zoom-in and zoom-out operations) and *to drag* (scrolling) which are alternatives to the button widgets that are provided by the user-interface.

In his works, Shneiderman identifies several interaction style principles applicable to this:

- continuous representation of the objects of interest: while performing an action user can see its effects on the state of the system
- physical actions instead of complex syntax: in contrast with command-line interfaces, actions are invoked physically via input peripherals, visual widgets (button, menu,) and gestures.
- continuous feedback (Fig. 3.5) and reversible incremental actions: it results easy to validate each action and fix mistakes
- rapid learning: users learn by recognition instead of remember complex syntax commands

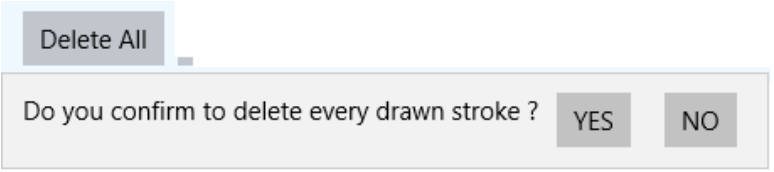


Figure 3.5: Example of a feedback to avoid unwanted irreversible actions while using a "Delete" button.

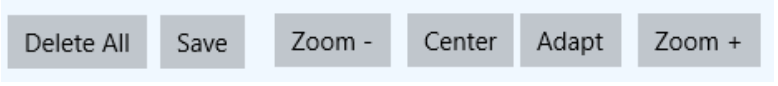


Figure 3.6: The top panel of the annotation tool (part 1): the user can save or delete the annotations and manipulate the image display.

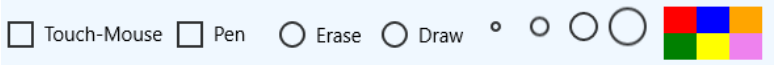


Figure 3.7: The top panel of the annotation tool (part 2): the user can set the interaction and input modalities and manage the stroke features.

The interface is minimal but functional and presents two retractable panels at the top (separated in Fig. 3.6 and 3.7) and at the right (Fig. 3.8) part of the screen, with the aim to take full advantage of the whole screen space. The top panel contains a set of functionalities regarding image and annotation processing:

1. delete or store the annotation strokes on the device (Fig. 3.6)
2. manipulate the image display (zoom in/out, center on the screen, scale adaptation) (Fig. 3.6)
3. select the input device (mouse/touch, pen or both) (Fig. 3.7)
4. select the interaction modality (draw or erase strokes on the screen) (Fig. 3.7)
5. manage the stroke features like color and width (Fig. 3.7)

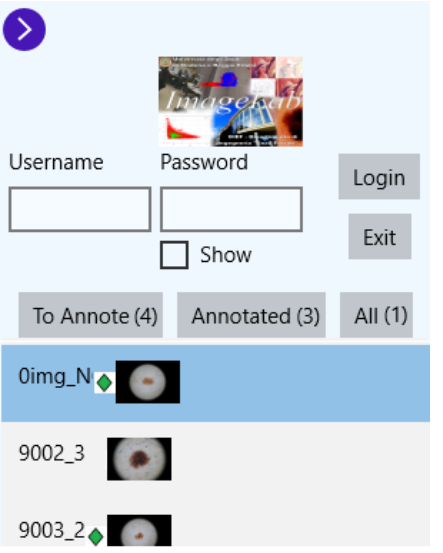


Figure 3.8: The right panel of the annotation tool: it contains the login module, the list of images (with a thumbnail) and the tabs that dynamically filter them.

The right panel presents two horizontal sections: the superior one permits the login and the exit from the application (exploiting the internal users database that also allows the meta-data organization previously described). The lower panel allows the image selection and loading by showing a list with a small preview (thumbnail) of them and, if an image has previously been annotated, a green rhombus appears next to its name).

In the previous panel there are three tabs that have a dual functionality: they show respectively the total number of images, the number of the annotated images and the number of the images that shall be annotated; moreover, when the user selects one tab, a *dynamic filter* updates the image preview list based on the selected preference.

Strongly Disagree			Indifferent			Strongly Agree
1	2	3	4	5	6	7

Figure 3.9: The 7-degree *Likert* scale used for answer to a question.

3.2.2 Experimental Survey

An evaluation was performed to test the annotation facilities and to understand if the user-interface and its functionalities were easy to understand and perform: six non-medical subjects, after a small brief on the capabilities and the target of this tool, were encouraged to explore the tool and observed while carrying some annotation tasks on a subset of 25 images.

After each experimentation phase, a questionnaire made by following Human-Computer Interaction guidelines and consisting by 20 questions was given to the subjects [99]. The evaluation for each question is given by a 7-degree ordered *Likert* scale as in Figure 3.9 and asks to rate the agreement with the statements ranging from "strongly disagree" to "strongly agree". Questions are divided into four sections concerning *Usefulness*, *Ease of Use*, *Ease of Learning* and *Satisfaction*.

The questionnaire considers that users evaluate primarily using the dimensions of *Usefulness*, *Satisfaction*, and *Ease of Use* which are used to discriminate between interfaces. Partial correlations calculated using scales suggested that *Ease of Use* and *Usefulness* influence one another, such that improvements in Ease of Use improve ratings of Usefulness and vice-versa; while both drive to *Satisfaction*, the Usefulness is relatively less important with internal systems that users are required to use. Users are more variable in their Usefulness ratings when they have only limited exposure to a product while, as expected from the literature, Satisfaction is strongly related to the real usage (actual or predicted).

Questions and answers (evaluated by the frequency of votes in a *Likert* scale from 1 up to 7) are reported in Table 3.1: the majority of answers expresses positive agree as for question j (Both occasional and regular users would like it) or question f (It is easy to use). Few low votes in interval [3-5] considering question t (I feel I need to have it) and question k (I can recover from mistakes quickly and easily).

In Figure 3.10 is shown a plot which summarizes and visualizes all

↓ Question - Rating →	1	2	3	4	5	6	7
a) It gives me more control over the activities	0	0	0	0	0	3	3
b) It makes the things I want to accomplish easier	0	0	0	0	0	3	3
c) It saves me time when I use it	0	0	0	0	2	4	0
d) It meets my needs	0	0	0	1	0	4	1
e) It does everything I would expect it to d	0	0	0	0	2	1	3
f) It is easy to use	0	0	0	0	0	2	4
g) Using it is effortless	0	0	0	0	0	2	4
h) I can use it without written instructions	0	0	0	0	0	3	3
i) I don't notice any inconsistencies as I use it	0	0	0	0	2	3	1
j) Both occasional and regular users would like it	0	0	0	0	0	1	5
k) I can recover from mistakes quickly and easily	0	0	1	0	2	1	2
l) I learned to use it quickly	0	0	0	0	0	2	4
m) I easily remember how to use it	0	0	0	0	0	2	4
n) It is easy to learn to use it	0	0	0	0	0	3	3
o) I quickly became skillful with it	0	0	0	0	0	2	4
p) I am satisfied with it	0	0	0	0	2	4	0
q) I would recommend it to a friend	0	0	0	0	0	5	1
r) It is fun to use	0	0	0	0	1	3	2
s) It works the way I want it to work	0	0	0	0	1	3	2
t) I feel I need to have it	0	0	1	1	2	0	2

Table 3.1: The Questionnaire.

the questionnaire results: it consists in an histogram featuring various dimensions:

1. X-axis: question code

2. Y-axis: the overall score in percentage that expresses a polarization towards a "strongly agreement" by considering a weighted vote for each given answer with the rating [1 up 7]

3. Color: the blue palette expresses dislike, the yellow neutrality and the green a positive mood

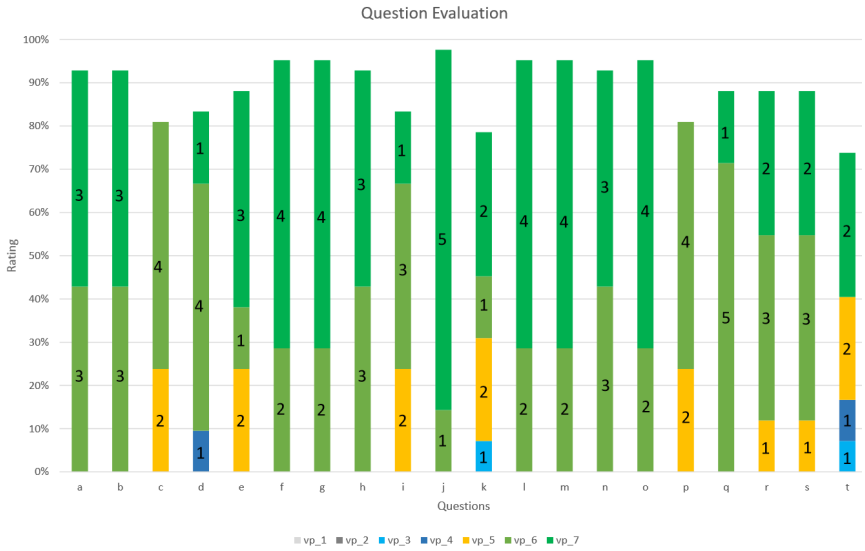


Figure 3.10: The histogram about questionnaire answers.

4. Number (into a bar): answer frequency, how many subjects answered to the question with the same rating

5. Sub-bar: its width expresses the weight that have a specific answer on the overall score (100% would be all subjects scoring the highest vote of 7)

Observing the plot it becomes evident that, by considering the overall frequencies score in percentage, for all questions subjects expresses a positive liking mood (the height of all the bars exceeds 80%).

Since a sub-bar height is related to the liking mood, expressing a lot of low ratings prevents that the global bar height could exceeds the one of a question with higher scores: a notable case is the comparison between question j and question q since they have the sub-bar with a numerousness one the opposite of the other. By exploiting this visualization mode the differences between the answers become evident, for example between question g and question h: they differ for only one subject so the first bar is higher than the second since in the second one the subject answered value

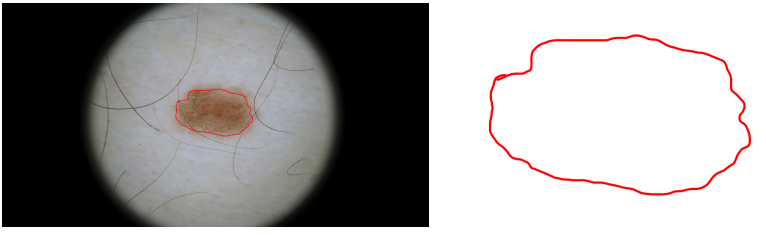
'6' instead of '7' (the number '3' in the question g sub-bar and its clearer color evidence a majority of frequency but a lower valence compared with the bar of darker green color).

3.3 Extraction of Derived Features

Once the dataset has been created, it is possible to apply image processing functions and algorithms to extract *derived features* like contours, shapes, intersections, color features and numerical values, as anticipated in Fig. 3.3; referring to the architecture of Fig. 3.1 this stage focuses on the exploit of state-of-art reliable methods for extracting the visual features needed, also introducing workarounds to improve results.

The main target is to manage the high-resolution image dataset and to exploit low-level visual features with image processing pipelines and machine learning techniques: in this way an intelligent and proactive diagnostic system can be trained using this ground-truth for the task of *segmentation*, learning how to isolate lesion areas and borders by mimiking the knowledge of many dermatologists.

In Fig. 3.11-a there is a dermoscopic image where a dermatologist annotated the principal lesion contour (red color): the primitive meta-data directly extracted by the tool is in Fig. 3.11-b: the annotation will be used to evaluate how many lesion pixels are automatically detected by the system but, before this, a pre-processing phase is necessary to remove the thick hairs since these artifacts influence shape and contour extraction [162].



(a) Dermoscopic annotated skin lesion.

(b) Extracted annotation pixels (primitive feature).

Figure 3.11: An annotated image with its primitive features.

3.3.1 Pre-processing: Thick Hair Removing

To accomplish the task it has been reworked the *DullRazor* [8] pipeline consisting of:

1. *Detection* step that locates into an RGB image the slender and elongated structures that resemble the hairs on the skin by making an hair pixel mask

2. *Replacement* step that replaces each retrieved hair pixel with the interpolation of two lateral pixels chosen from a line segment built on a straight direction

The software will be developed in C++ to reach best performances; the *Detection* phase exploits the *generalized grayscale morphological closing* operation G_c: for each color channel c (Red-Green-Blue) the operator makes a set of morphologic closing by using different kernels with the aim to compare (by choosing the highest value c_p), for each pixel, which kernel better approximates a potential hair shape.

The value of G_c for each pixel p is calculated as:

$$\forall c \in r, g, b, \forall p, G_c = |b_c(p) - max(c_p)| \tag{3.1}$$

where $b_c(p)$ is the actual image pixel value and c_p measures how many pixels of the kernel are verified as hair structure for an image pixel p.

The kernel represents a sort of 'skeleton' that reconstructs pixel-to-pixel the hair shape (elongated, slight, weakly curvilinear) on the kernel-closed image and so the kernel structure is a very critical variable; it have been used four kernels for each of the possible directions (11 pixels horizontal at 0°/180°, 11 pixels vertical at 90°/270°, 9 pixels oblique at 45°/225° and 9 pixels oblique at 135°/315°) from which a potential hair could spread from a single pixel (located et the center of the kernel).

The final hair mask M_h is the union of the resulting pixel masks is

$$M_h = M_r \cup M_g \cup M_b \tag{3.2}$$

where each mask is obtained by a threshold on the generalized closing operator value previously calculated for each pixel.

Before the *Replacement* phase it must be verified that each candidate hair pixel belongs to a valid thick and long structure so, for each direction

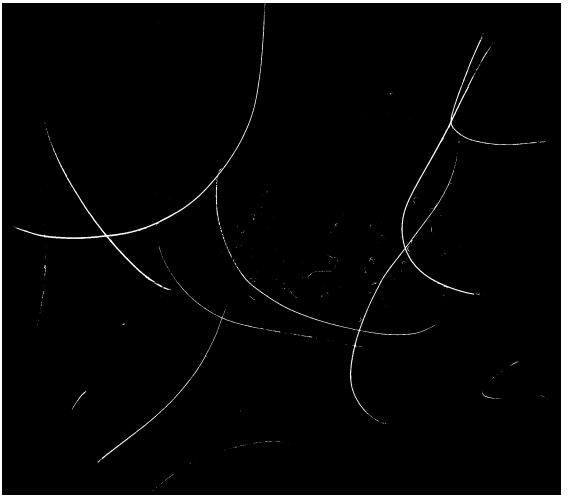


Figure 3.12: The output thick hairs mask (M_h) for the image in Fig. 3.11.

previously described, a path is built having the pixel at the center until the non-hair regions are reached. The longest path is used to take the two interpolation pixels, selected on both the sides, perpendicular to the directions, at a fixed distance from the hair structure borders.

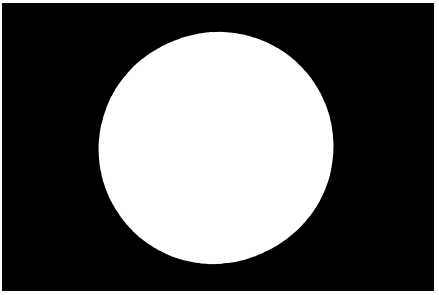
This pipeline is greatly affected by the variables used at each step, for example the kernel shape, the G_c threshold for the hair masks and the distance for the choose of the interpolation pixels.

3.3.2 Skin Lesion Detection

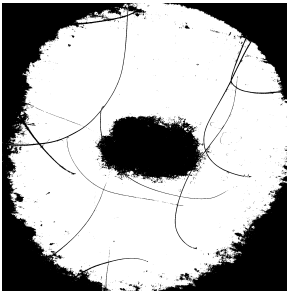
After removing the most of hairs, it is possible to design methods for the skin lesion segmentation by exploiting standard image processing techniques in a fine-tuned *work pipelines* that filters and refines results:

1. Thresholding pipeline with pixel mask
2. Color Clustering pipeline with pixel mask and cluster tolerance

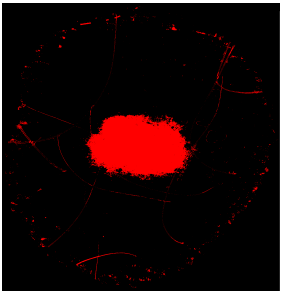
By considering the peculiar structure of a dermoscopic image, a binary mask M_0 (Fig. 3.13-a) is built with a morphologic *erosion* on the original



(a) External black frame mask (M_0).



(b) Threshold with mask

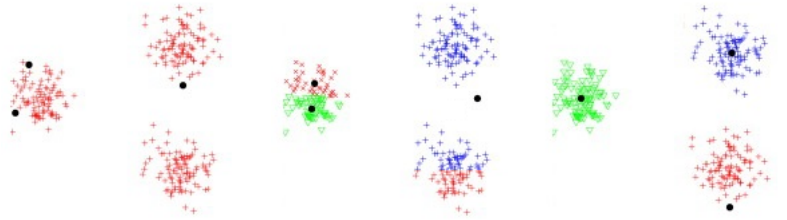


(c) Threshold with mask, hairs removed.

Figure 3.13: Thresholding pipeline with pixel mask.

blurred image (kernel size 201x201) with the aim to approximate the large black frame which surrounds the bright circular skin that contains the lesion. The *Thresholding pipeline* uses the Otsu algorithm calculated considering only the pixels not in M_0 (Fig. 3.13-b); to reach better performance it must be considered that it is enough "to break" the hair structures so that it becomes harder to create wrong closed paths: Fig. 3.13-c shows the pixels of the thresholded lesion when hairs have been removed (for future tests and comparisons pixels are in red color).

The *Color clustering pipeline* is based on the K-nearest neighbors clustering (kNNs) on pixel colors that iteratively measures a *distance* according one or more features (for example the RGB value) on objects mapped in a n-dimensional feature space; the algorithm randomly chooses k *centroids*



(a) First step, random centroids.

(b) Second step, new distances.

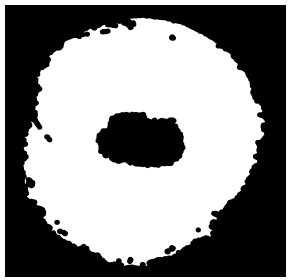
(c) Third step, convergence.

Figure 3.14: Example of iterative clustering ($k = 3$).

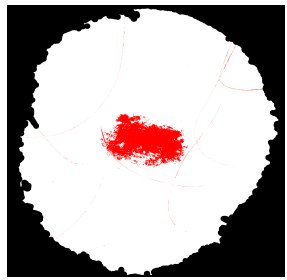
and calculates the distances from each object of the dataset: in this way each of them will be labeled the same as the closer centroid giving origin to the first k clusters. Next it is calculate the average of all the objects of each cluster, and the point that represents this value becomes the new centroid of the cluster, leading to repeat the distance evaluation process until the convergence with the aim to build more compact and accurate clusters (Fig. 3.14).

For this clustering pipeline it is needed an heuristic to differentiate if a cluster belongs to the bright skin or to the lesion skin by assign the label *skin* or *lesion* to each cluster; moreover it must be considered that a cluster can easily intersect the two area types (some pixels on the lesion and others on the skin) especially on the borders. To deal with this ambiguity a pixel *toleration area* is developed: it exploits a further morphologic *closing* and *erosion* on the thresholded image by using two different kernels to obtain an *enlarged mask* that builds "safety areas" around the lesion borders (side effects are that scattered hair pixels are removed but groups of them may be enlarged).

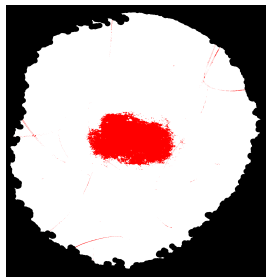
The algorithm is computed using *OpenCV* functions and, as seen for the threshold pipeline, by only considering pixels not in M_0 mask; the number of clusters (K=10) has been chosen empirically after a set of experimental sessions. Moreover the enlarged pixel mask in Fig. 3.15-a has been divided into two sub-masks M_1 and M_2: the first represents the largest connected component that from now will replace M_0, while M_2 is the second largest connected component thar approximates the tolerance area for the central lesion pattern and that must contain the bigger and uniform cluster.



(a) The pixel toleration areas: M_1 (black frame) and M_2 (central).



(b) Final lesion global cluster, no hairs removed.



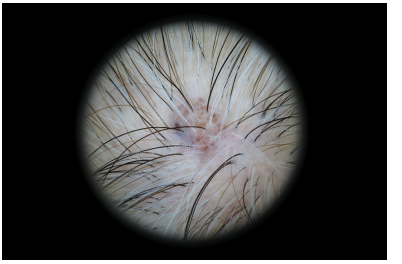
(c) Final lesion global cluster, hairs removed.

Figure 3.15: Color clustering pipeline.

For each color cluster are counted the pixels considered 'skin' or 'lesion' by using the M_2 masks so if a cluster has more than 10% of its pixels out from this tolerance area it will be considered as simple skin (all of its pixels labeled 'skin'). It must be noticed that M_1 must necessarily be used to exclude clusters that compose the concentric halo near the border between the bright skin and the black frame.

Examples of the global color cluster consisting by all the pixels labeled 'lesion' are in Fig. 3.15-b and Fig. 3.15-c respectively for the original image and for the one with hairs removed (for the future comparisons the pixels are in red color).

(a) Well-detailed lesion, average hair presence.



(b) Nuanced lesion, massive hair presence.

Figure 3.16: Examples of hard dermoscopic images.

3.3.3 Experimental Study

To test which of the proposed segmentation pipelines reaches the best performance, from the dataset is built a group of *hard images* where lesions differ from each other for size, colors, patterns, type (pigmented or not) and moreover for the presence of thick hairs at various sizes (Fig. 3.16).

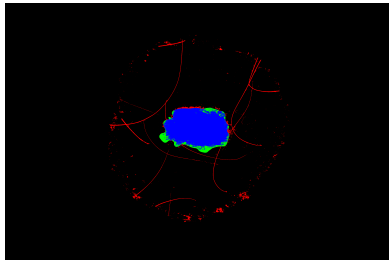
Each of the 17 chosen images are differentiated in 'original' version (Dataset 1) and 'hair removed' version (Dataset 2) for a total of 34 images. The experiment has a within-subjects design with two treatments:

1. Dataset 1 - Threshold pipeline
2. Dataset 2 - Threshold pipeline
3. Dataset 1 - Color clustering pipeline
4. Dataset 2 - Color clustering pipeline

From the manual annotation it is extracted a green labeled area whose pixels must be considered as a *lesion ground truth* (Fig. 3.17-a). For each image in the datasets can be considered the resulting labeled pixels (previously seen in red color) coming from the two pipelines and intersect them with the ground-truth area to calculate measures like *Precision, Recall, f1-score* and *Accuracy*.

To compute the four measures we consider as *global goodness* of segmentation the value coming from the sum of the pixel of each classification, for each image of the dataset.

(a) Area derived from the manual annotation (ground truth).



(b) Intersection of annotations for comparison.

Figure 3.17: The annotation comparison.

An example of the comparison between the ground-truth and the Threshold pipeline (Fig. 3.15-b) is depicted in Fig. 3.17-b:

- true positives (TP): blue pixels (pixel intersection)
- false positives (FP): red pixel (that not intersect the green ones)
- false negatives (FN): green pixels (that not intersect the red ones)
- true negatives (TN): all the remaining black pixels

The metrics involved in the pixel comparison are:

• Precision $= \frac{TP}{TP+FP}$	• f1-score $= 2 * \frac{Precision*Recall}{Precision+Recall}$
• Recall $= \frac{TP}{TP+FN}$	• Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

Experimental results are in Tables 3.2 and 3.3 for the Threshold (Otsu) pipeline while Tables 3.4 and 3.5 shows results for the Color Clustering pipeline; it must be noticed that due to the peculiar image template and its size (pixel number) the *Accuracy* metric does not results significant, in fact the True Negative pixels (TN) represent the largest area (as the black frame) that is excluded for pixel comparison during the pipelines.

The results comparison between the two pipelines on each dataset evidences the detection improvement coming from the Color clustering

Image	TP	FP	FN	TN
9001_2	217579	68714	33936	10335771
9012_2	70030	96724	2253	10486993
9013_3	364014	71399	14687	10205900
9017_2	87322	120510	5082	10443086
9030_2	296949	165099	3822	10190130
9051_2	7570	130537	219452	10298441
9067_2	106601	138285	10404	10400710
9084_2	52677	84071	3436	11859816
9108_2	214603	208246	1024	11576127
9110_2	176420	124390	12708	11686482
9112a_2	59915	47201	9617	11883267
9121_2	77964	41491	6187	11874358
9131_2	43994	109632	11788	11834586
9143_2	97002	113956	1150	11787892
9149_2	109552	61622	80655	11748171
9178a_2	645856	69300	71720	11213124
9190_2	818771	21707	82220	11077302
Tot.	3446819	1672884	570141	188902156
Precision (global)	0.67			
Recall (global)	0.86			
f1-score (global)	0.75			
Accuracy (global)	0.99			

Table 3.2: Results: Threshold pipeline - Dataset 1 (original).

especially in terms of the *Precision* metric, in fact it increases from 67% to 94% for the dataset 1 and from 73% to 94% for the second one.

Also examining results between the two dataset for each pipeline we notice that all metrics always improve; only for the second dataset in Color Clustering pipeline, the *Precision* remains unchanged but *Recall* increases significantly: this demonstrates the need of hair removing treatment before each pipeline. The *f1-score* metric provides the best interpretation of the results: it increases in the threshold pipeline (from 75% for dataset 1 to 79% for dataset 2) and in the Color clustering pipeline (from 79% for dataset 1 to 83% for dataset 2) and, as it is evident, it is always better for the second

Image	TP	FP	FN	TN
9001_2	220376	43680	31139	10360805
9012_2	69964	80068	2319	10503649
9013_3	364010	66414	14691	10210885
9017_2	88283	106019	4121	10457577
9030_2	297509	73515	3262	10281714
9051_2	54816	129842	172206	10299136
9067_2	107926	116745	9079	10422250
9084_2	52832	57024	3281	11886863
9108_2	214525	158533	1102	11625840
9110_2	175242	92636	13886	11718236
9112a_2	59604	31391	9928	11899077
9121_2	77764	42813	6387	11873036
9131_2	44603	96025	11179	11848193
9143_2	96999	110616	1153	11791232
9149_2	114513	47094	75694	11762699
9178a_2	648944	36962	68632	11245462
9190_2	816557	8327	84434	11090682
Tot.	3504467	1297704	512493	189277336
Precision (global)	0.73			
Recall (global)	0.87			
f1-score (global)	0.79			
Accuracy (global)	0.99			

Table 3.3: Results: Threshold pipeline - Dataset 2 (hair removed).

pipeline at the dataset equality. Finally it must be considered that as previously explained we are considering images having highly characterized features that represents only a small part of the complete dataset.

Image	TP	FP	FN	TN
9001_2	126582	3625	111980	10413813
9012_2	38998	3192	25896	10587914
9013_3	299399	14928	64596	10277077
9017_2	0	0	84975	10571025
9030_2	116308	15577	170264	10353851
9051_2	0	0	215597	10440403
9067_2	88896	19563	19617	10527924
9084_2	0	0	50833	11949167
9108_2	190647	38220	10254	11760879
9110_2	133860	14066	44867	11807207
9112a_2	20617	1455	41831	11936097
9121_2	56131	739	20949	11922181
9131_2	0	0	49121	11950879
9143_2	86348	20269	3693	11889690
9149_2	82926	6518	97343	11813213
9178a_2	499044	17268	195287	11288401
9190_2	736431	7393	141939	11114237
Tot.	2476187	162813	1349042	190603958
Precision (global)	0.94			
Recall (global)	0.65			
f1-score (global)	0.77			
Accuracy (global)	0.99			

Table 3.4: Results: Color Clustering pipeline - Dataset 1 (original).

Image	TP	FP	FN	TN
9001_2	183634	6196	54928	10411242
9012_2	57091	1600	7803	10589506
9013_3	299163	10266	64832	10281739
9017_2	59410	1423	25565	10569602
9030_2	278039	24196	8533	10345232
9051_2	0	0	215597	10440403
9067_2	82996	17204	25517	10530283
9084_2	30803	1022	20030	11948145
9108_2	198896	94849	2005	11704250
9110_2	156684	30981	22043	11790292
9112a_2	28484	120	33964	11937432
9121_2	56452	491	20628	11922429
9131_2	23825	3110	25296	11947769
9143_2	87097	21348	2944	11888611
9149_2	80332	268	99937	11819463
9178a_2	520811	3489	173520	11302180
9190_2	736402	1201	141968	11120429
Tot.	2880119	217764	945110	190549007
Precision (global)	0.93			
Recall (global)	0.75			
f1-score (global)	0.83			
Accuracy (global)	0.99			

Table 3.5: Results: Color Clustering pipeline - Dataset 2 (hairs removed).

Chapter 4

Deep Learning for Skin Lesion Segmentation

As told in previously, the introduction of Convolutional Neural Networks (CNNs) with the deep learning approach have shown stunning results in computer vision and image classification topics: in 2012 Krizhevskys *et al.* [88] proposed *AlexNet*, a CNN used for the first time for image classification with which they won the *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) while, in 2014, Fully Convolutional Networks (FCN) were first used by Long *et al.* [97] to perform end-to-end segmentation of natural images by observing that the fully connected layers in a classification network can be viewed as convolutions with kernels that cover the entire input region.

Deep learning can learn sets of high level features from low level ones and gain high accuracy for classification applications without the need for extracting handcrafted features: the main goal consists in providing a very large number of *labeled samples* as a ground-truth to perform supervised learning by exploiting the architecture (layers, weights, neurons) and the operators (filter, pooling, ..) by choosing the appropriate parameter values.

In Cap. 3 have been introduced technical workarounds to create a ground-truth dataset: since an input image contains both healthy (normal skin) and lesion parts, in the the training of a CNN can be exploited *segmentation masks* as that produced by the two working pipelines previously proposed.

The deep learning approach takes advantage of a set of convolve-filters that can analyze various structures in the input images, from which the network automatically extracts visual and structural features consisting in some *general patterns*.

The general patterns that hopefully a lesion segmentation network has to automatically learn are that used by dermatologists for the classification of skin lesions: the most important are that coming from the *Pattern-Texture Analysis* and the *ABCD Rule* that investigates [102]:

1. Asymmetry: the lesion is bisected by two axes that are positioned to produce the lowest asymmetry possible with respect to a point under one or more axes

2. Border: the lesion is examined if there is a sharp, abrupt cutoff of pigment pattern at the periphery of the lesion or a gradual, indistinct cutoff

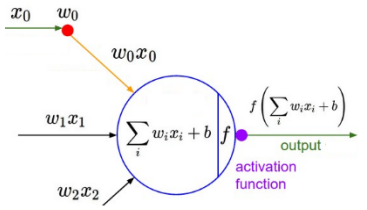
3. Color: color properties inside the lesion are examined and the number of colors present is determined

4. Differential structures: the structural components like pigment network, dots, globules, structureless areas

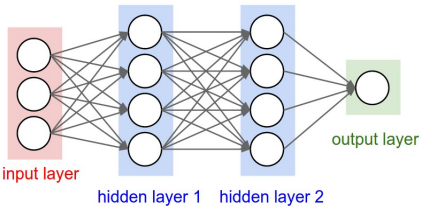
4.1 The Architecture of (Deep) CNNs

Deep Convolutional Neural Networks (ConvNets or CNN) are very similar to ordinary Neural Networks where *layers* are modeled as collections of *neurons* (Fig. 4.1-a) that have learnable weights and biases and are linked in an acyclic graph (Fig. 4.1-a) where into a single layer they do not share any connection [6]. Layers receive some inputs and transform them performing dot products optionally followed by a non-linear activation function: while the whole network expresses a single differentiable score function, the weights values are updated through *back-propagation*, a way of computing gradients of expressions through recursive application of chain rule to minimize a qualitative metric called *loss function*.

A CNN makes the explicit assumption that inputs are images and so it is possible to encode certain properties that mitigate the parameters number and the complexity (Fig. 4.2); unlike regular Neural Networks in fact, the layers have neurons arranged in 3 dimensions (width, height,



(a) In regular NN a single neuron can implement a binary classifier performing dot products.



(b) NN are modeled as layers of neurons that receives inputs and transforms it through learnable weights and biases.

Figure 4.1: A brain-analogous neuron scheme and a regular Neural Network.

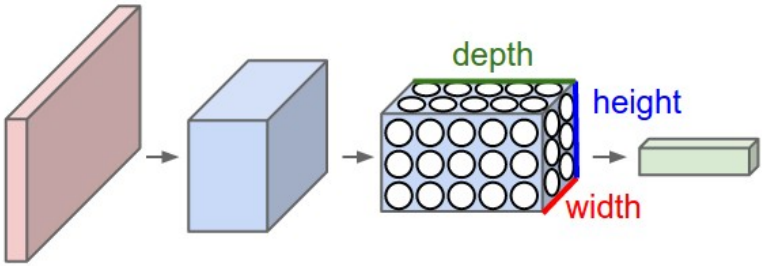


Figure 4.2: A ConvNet arranges its neurons in three dimensional layers that transform each 3D input volume into an output 3D volume.

depth) that are only connected to a small region of the layer before them: when dealing with high-dimensional inputs in fact, it is impractical to have all the neurons in a fully-connected way and so CNN layers implement a *local connectivity* that filters the input using a *kernel* that defines a sliding pixel window called *receptive field*.

The most common form of a ConvNet architecture stacks several *convolve layers* and *pooling layers* with a *fully connected layer* as the last one, repeating this pattern until the input image has been merged spatially to a small size. Each convolve layer is followed by a pooling one that operates independently on every depth slice of the input and performs a downsampling operation along the spatial dimensions (width, height) reducing the size of the output feature map: in this way the resized image

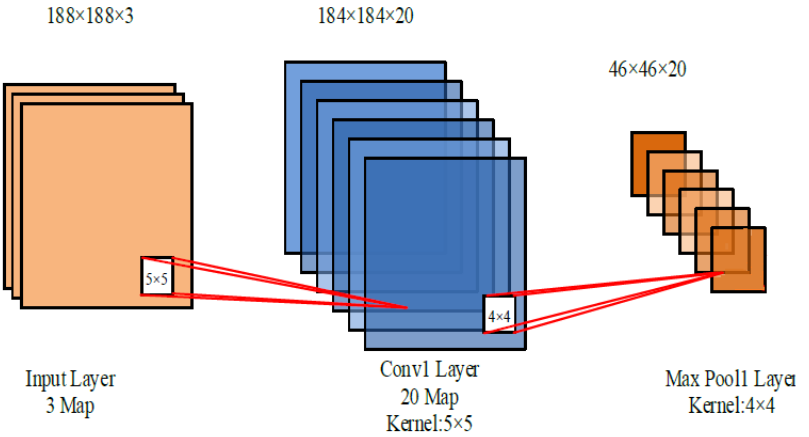


Figure 4.3: A scheme pf how the convolve and pooling layers work.

keeps some general patterns and also highlights new ones.

The use of local filters rests on the *Parameter sharing* scheme that reduces the net parameters by making the assumption that if one feature is useful to compute at some input spatial position, then it should also be useful to compute at a different position.

The hyper-parameters that control a conv. layer are:

- depth: the number of filters used
- stride: with which the filter slides on the input image
- zero-padding: allows to control the spatial size of the output by pad the input with zeros around the border

The scheme in Fig. 4.3 shows how the main layers work [114]:

1. input: the first layer consists of an image of 188x188 pixel size with 3 color channels (in that way result 3 input images, one for each channel)
2. conv. layer: uses 20 filters with a 5x5 kernel that extend through the full depth of the input, from the first input rise 20 new activation maps with the same size of the original

3. pool layer: uses a 4x4 kernel to reduce the size of each input image while the *MaxPool* operator takes the maximum pixel value in the kernel window: in this example the output image becomes 1/4 of the original (no lost pixels since there is no border padding)

Moreover there are also other types of layer to consider:

- fully-connected layer: as in regular Neural Networks neurons between two adjacent layers have full connections with that of the previous layer and their activations are computed with a matrix multiplication followed by a bias offset; in a CNN it computes the class scores and has a [1x1xC] size, where C is the number of the output class scores

- dropout layer: at each training stage, individual nodes are dropped out of the network with a certain probability improving the network ability to generalize and avoiding the overfitting

- activation layer: usually follows conv. and fully-connected layers and applies an elementwise non-linear activation functions (Sigmoid, Tanh, ReLU, ...) that enhance the network expression capability

It is remarkable that for the convolution operation the *backward pass* (for both the data and the weights) is also a convolution but with spatially-flipped filters while for the pooling layers the backpropagation has a simple interpretation as only routing the gradient to the input that had the highest value in the forward pass; moreover while conv. and fully-connected layers perform transformations that are function of not only the activations in the input but also of the parameters (the weights and biases of the neurons), the activation and pooling layers implement a fixed function.

An advantage in using CNNs is the reuse of their specialized architectures to reach similar goals as done for example with AlexNet that, despite being originally created for real world images, in Majtner *et al.*[103] has been adapted for melanoma recognition; moreover, if the train dataset is almost similar with the original it is possible to reuse also the weights, considering the net as a *pre-trained* CNN in which update its weight values calculating new gradients by passing the new images.

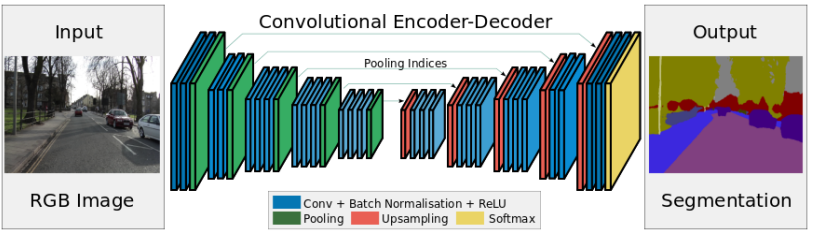


Figure 4.4: SegNet is one of the first architectures with an encoder network followed by a decoder one.

4.2 DeepLab as Segmentation Architecture

Unlike the classification goal, where the only end result of a CNN is the global class presence probability, a general semantic segmentation architecture results more complex since it is constituted by two main parts: an *encoder* network followed by a *decoder* network (a clarifying scheme is in Fig. 4.4 [16]).

An encoder is usually a pre-trained classification network like Oxford's *Visual Geometry Group net* (VGGNet) [95] or *Residual Network*(ResNet) [72], followed by a decoder network that is the part where each architecture differs since its goal is to semantically project the discriminative features (lower resolution) learned by the encoder onto the pixel space (higher resolution) to obtain a *dense classification*.

Pooling operations are useful to increase the field of view of a filter reducing the feature map resolution but have the side effect of losing spatial information (where output features are located) so it is needed a mechanism to recover the information lost; different architectures employ different mechanisms (skip connections, pyramid pooling etc) as a part of the decoding phase, which eventually turns into the goal of learning to execute an interpolation function.

ResNet-101 is a classification CNN with 101 layers divided into four groups and constitutes the pre-trained network encoder of DeepLab V2 [42, 41]: its architecture is an evolution of VGG, in fact the second step is the new main module while the first, third and fourth steps remain the same.

The VGG architecture (Fig. 4.5) consists of a sequence of five main groups (plus one fully-connected layer as output) where adjacent groups

Figure 4.5: The scheme of VGG-16 from which comes ResNet, the encoder network of the DeepLab architecture.

are connected via max-pooling layers: it only performs 3x3 convolutions and 2x2 pooling from the beginning to the end and each group contains a series of learnable conv. layers and their total amount could be 11, 13, 16, or 19 depending on their number in each group (in ResNet their number is different since the network depth could be 50, 101 or 152 and also are missed fully connected layers at the end of the network).

DeepLab architecture exploits a different approach from the encoder-decoder one, in fact instead of using upsampling and deconvolutional layers introduces innovative aspects like:

1. atrous/dilated convolutions (filters with spaces between each cell)
2. atrous spatial pyramid pooling (ASPP)
3. fully-connected Conditional Random Fields (CRF)

The problem of the reduced spatial resolution when using pooling layers is approached with the atrous convolution (the term is a shorthand for convolution with upsampled filters), originally developed for the efficient computation of the wavelet transform in the "algorithme a trous"; unlike the deconvolutional approach, the proposed one converts image classification networks into dense feature extractors by removing the downsampling operator from the last few max pooling layers and from subsequent convolutional layers by introducing a upsampling method that inserts holes (trous in French). The method offers an efficient mechanism to find the best trade-off between accurate localization (small field-of-view) and context assimilation (large field-of-view) allowing to compute the responses of any layer at any desirable resolution: the resulting feature maps computed at

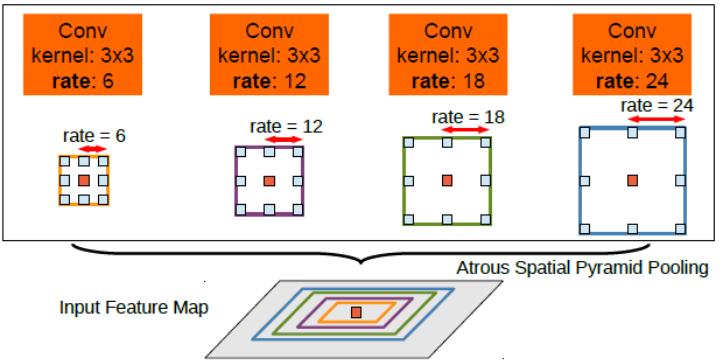


Figure 4.6: To classify the center pixel ASPP exploits multi-scale features by employing multiple parallel filters with different rates.

a higher sampling rate appear more dense, leading to effectively enlarge the filters field of view without increasing the number of parameters or the amount of computation.

Deep CNNs have the remarkable ability to implicitly manage the existence of objects at multiple scales by using training datasets that contain rescaled versions of the same image and then aggregating the score maps. To handle the scale variability in semantic segmentation DeepLab introduces the ASPP technique (Fig. 4.6), a computationally efficient scheme of resampling a feature layer at multiple rates prior to convolution by using multiple parallel atrous layers with different sampling rates; the features extracted for each sampling rate are then further processed in separate branches and fused to generate the final result.

The third improvement relates to the fact that an object classifier requires invariance to spatial transformations but, although score maps can predict the presence and rough position of objects, they cannot really delineate their borders. To address the spatial accuracy challenge the architecture employs CRFs: with the aim to overcome the limitations of short-range CRFs used to smooth noisy segmentation maps, DeepLab boost its ability to capture fine details by employing a fully connected Conditional Random Field model, that leads to fine-grained localization accuracy by recovering object boundaries at a level of detail that is well beyond the reach of existing methods.

4.2.1 The International Skin Imaging Collaboration (ISIC) project

The International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale publicly accessible dataset of dermoscopy images and currently the dataset contains more than 20,000 images from leading clinical centers internationally, acquired from a variety of devices [69]. The images are screened for both privacy and quality and the associated clinical metadata has also been vetted by recognized melanoma experts; broad and international participation ensures that the dataset contains a clinically representative distribution.

The ISIC dataset was the foundation for the first public benchmark challenge on dermoscopic image analysis in 2016: the goal of the challenge was to develop automated melanoma diagnosis algorithms across 3 parts of lesion analysis: 1) lesion segmentation, 2) lesion feature detection (4 classes) and 3) lesion classification (3 classes).

The metrics considered for the evaluation of the segmentation task are:

- Pixel-level Sensitivity (Recall) = $\frac{TP}{TP+FN}$
- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Pixel-level Specificity = $\frac{TN}{TN+FP}$
- Dice Coefficient = $\frac{2*TP}{2*TP+FP+FN}$
- Jaccard Index = $\frac{TP}{TP+FP+FN}$

For the segmentation task participants have to submit automated predictions of lesion segmentations from dermoscopic images in the form of binary masks: the training data include the original image with the manual annotation made by an expert that defines the lesion boundaries where the pixel value of 255 denotes the lesion area and the value 0 is normal skin.

In 2017 the top ranked participant achieved a *Jaccard Index* of 0.765, an *Accuracy* of 93.4 and a *Dice coefficient* of 0.849, using fully CNN and deep learning approaches; from the segmentation results emerges that the average Jaccard Index may not accurately reflect the number of images where automated segmentation falls outside an inter-observer variability so this metric could be adjusted by either using multiple segmentations per image to determine a failure threshold or by choosing a fixed threshold based on measurements.

4.2.2 Designing an Experimental Environment

Since all of the DeepLab code is publicly available and moreover, since it sets new state-of-art at the Pascal VOC-2012 semantic image segmentation task (79.7%) and advances the results with Pascal-Context, Pascal-Person-Part, and Cityscapes datasets, it can be considered to exploit the medical image segmentation goal. The fully convolutional structure has been adapted and executed using the Keras framework with Theano as backend [4] but the original 21 output classes must be reduced to 2 (binary pixel classification as lesion or skin); moreover, since the goal is similar but the training dataset and the output classes are very different from the originals, the architecture can be maintained but its weight values should be retrained from scratch.

As seen in [47] to design a good experiment the ISIC training data must be split into two partitions of 2000 for training and 150 for validation, so that the network parameters can be trained on the 90% and the loss function computed on the remaining while the test dataset consists in further 600 images.

The network is trained with a *categorical cross-entropy* loss function and the output represents a binary decision for each pixel classified as 255 or 0; three optimizers are available (Adam, SGD and RMSProp) while the learning rate and the momentum have to be adjusted from their initial starting values to 0.001 and 0.9 respectively during the training.

Since the dimensions of the original images are very large and due to computational performances, they are resized to 128x128 pixels and the output to 16x16; it is known that to reach best results with deep learning approaches a large number of training samples is needed, so it is performed a phase of *data augmentation*: for each original image there are 7 new ones, increasing the original dataset from 2000 to 160000.

On each original image are performed the following operations:

- one random flip, choosing between vertical, horizontal or both axes
- three random rotations, choosing angles in the set A={15, 30, 45, 60, 75, 90}
- three random rotations of the flipped image, choosing angles in the set A={15, 30, 45, 60, 75, 90}

Chapter 5

Serious Gaming and Medical Image Analysis

The *Gamification* process consists in the application of game-design elements and principles in non-game contexts [75]: it uses the game mechanics to improve skills and knowledge of a subject, also enhancing its engagement and excitement while performing a task that usually does not provides them. Referring to the Csíkszentmihályi [53, 113] and Chen [40] studies (seen in Cap. 2) the sense of fun is strictly connected with the *Flow theory* characterized by the constant steady and balance between the *challenge* offered to gamers and the *skills* developed while facing them.

An *Exergame* identifies games that are also a form of exercise and involves the creation of a context in which the subject can use certain tools to replicate a series of real movements or tasks: they are used to counteract a sedentary activity, medical rehabilitation and promoting an active lifestyle and they are designed to provide immediate feedback to the player with the possibility of monitoring behaviors and biological parameters.

With the gamification technique it is possible to develop *serious games* which are games designed not only for the pure entertainment: this game genre is focused on the *simulation* feature with pedagogical purpose, by exploiting fun and competition while used in environments like defense, education [13], scientific exploration, health care [54], emergency management [27], city politics [101, 14, 24].

The aim of this chapter is to apply the gamification process to the task introduced in Cap. 3 that represents one of the main activity in the daily work of dermatologists with which they make diagnoses and comparisons: the medical image annotation.

5.1 The Gamification Process

The idea to develop the serious (video)game "Annote" for medical teaching (taking also inspiration from exergames) comes from the previous works done for the development of the tool seen in Fig. 3.4, Cap. 3 and used by academic dermatologists to annotate skin lesion images and build the ground-truth for MIMS (Fig. 3.1, Cap. 3). During a preliminary test session, senior dermatologists and academic interns expressed interest towards innovative learning methods like serious gaming.

Referring to the Flow concept explained in Cap. 2 and revised by Chen [40], a videogame as an interactive multimedia can be considered composed by two essential parts:

1. content: the soul of a videogame, a specific experience to convey which it has been designed

2. system: the body of a video game, the interactive software structure that, exploiting also hardware devices, communicates the content to the players through visuals, audio and interactions

The affective state of Flow clearly requires a balance and harmony of the two parts: the *content* can make special a videogame by intriguing and involving the player, but when the *system* part is very well-designed it already includes mechanics that naturally induce Flow since any content becomes rewarding; this explains why people prefer certain games more than other games and how they become addicted towards them, from the simplicity of *Tetris* to the complexity of *Civilization*.

All the previous theoretical considerations find a practical aspect in the survey of Hamari *et al.*[70] that presents ten *motivational affordances* tested on gamification empirical studies which can guide the design of this serious game (for need of reading, this principles will be referred with the prefix of "h" referring to the author of the mentioned work):

• h1) *Score points* • h2) *Leaderboard* • h3) *Achievements* • h4) *Levels* • h5) *Challenge*	• h6) *Plot story/Theme* • h7) *Goals* • h8) *Feedback* • h9) *Rewards* • h10) *Progress*

The work of Coltell *et al.*[48] takes up the previous aspects (also there principles have the prefix of "c" referring to the author) and adds:

• c1) *Rules* • c2) *Safety* • c3) *Interaction*	• c4) *People* • c5) *Fantasy* • c6) *Exploration*

The *game objects* is the act of 'draw strokes' and involves the right use of *interactive tools*: the *repetitiveness* results a learning element that, differently from commercial videogames, reinforces behavior change and a progression of the player performances [147, 30].

The principles h8, c1, c2, c3 and c6 are met by considering the client-server architecture in Fig. 5.1 (for safety and privacy of data), the interactive tools (interaction) and the user-interface (feedback, rules); moreover there is also a separate section with a non-interactive tutorial to learn the interface interactions and the aim of the game. An imaginative story (h6, c5, h7) results hard to introduce if not with the plot of a *survival mission* like "Save as many lives as you can" or with goals as "Reach/Surpass the points of a colleague" and "Annotate X lesions spending only Y time".

The main types of *challenge* offered by the game (c1, h5, h7, h10) are:

1. border challenge (precision): the player has to draw a lesion border annotation that imitates the 'official' one (ground-truth) also considering the completion of already begun strokes

2. structures challenge (recognition): the player has to annotate not

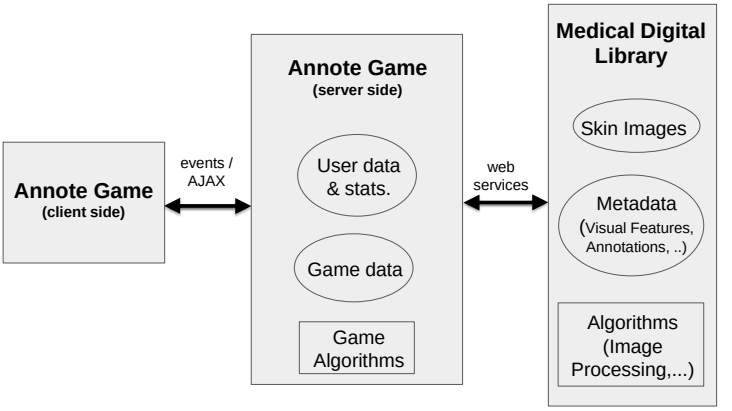


Figure 5.1: The *Annote* system architecture.

lesion borders but groups of skin *textures*, *clues* and lesion *patterns* (lines, circles, reticles, ...)

3. time challenge (pressure): a variant of the previous two where the player must annotate respecting a flowing timer for each image

4. lesion classification (quiz): the player looks an image and gives a diagnosis on the severity of a lesion by choosing from a ranked set (Likert scale)

It must be noticed that the ground-truth has been previously made for MIMS by the academics (or by the automatic algorithms used) so this guarantees the quality and the reliability of an annotation chosen for a gaming pixel comparison.

The variety of the game (h4, h5, h7, h10) is also given by the *difficulty* modes that can be proposed:

- the type and amount of images chosen by the teacher for a gaming session

- the activation of aid/impediment features

To expand and diversify the gameplay, forms of *rewards* are introduced (h1, h2, h3, h5, h10, c4):

- power-up or penalty: they grant or steal resources to the player (time, points), enable/disable features like the available tools (zoom, stroke width/color) or their performances (resolution/sharpness/size of the displayed image, mouse/pen speed)

- points and leaderboard: a centralized classification which emphasizes the desire to improve gaming (and learning) performances between students [107]; it also permits to know who are the other players/opponents

- personal profile: customizable, summarizes player informations and its gaming history

- badges and achievements: they are "titles" that appear in the player profile and in the leaderboard screen, depending by the number of accomplished tasks and by the amount of gained points

5.2 Developing the *Annote* prototype

This videogame allows to draw strokes on a dermoscopic image using different colors (8 choices) and pen sizes (4 choices); as previously seen, the dataset consists of 436 dermoscopic images in standard JPEG format with a resolution of 4000x2664 or 3000x4000 pixels used for the MIMS project.

Large part of design and technology is in common with the original annotation tool and the use of the .NET Framework (ASP.NET with C# the server side, XHTML with Javascript the client side) allows to neglect the problem of the *input mode*: since the client module reaches the server one by a standard web connection it is possible to interact with the *Surface Pen* if the client runs on a Surface device or with mouse/touch otherwise.

The Medical Image Management System interfaces with the game server using *web services* which allow communication between heterogeneous technologies; the game client permits the interaction with the user interface exploiting dynamic events and AJAX requests.

To create a gaming task, the following must be set:

• list of activated power-up(s) • points assigned for the goal • list of images to annotate	• type of challenge • time (if expected)

All gaming interactions and stroke collections are managed locally on the client and the data exchange with the server is limited to minimize the network resources and to separate the management of data and algorithms from the gaming module used by students to learn and to experiment on real medical data (that result protected for the privacy).

The exchanged data are:

• original image (input) • strokes (input/output)	• events (input/output) • textual settings (input): points, time, task messages

The Annote *server* manages data about points and achievements of the registered students with the *User data and stats.* module, the *game data* repository stores the settings of each gaming task while the *game algorithms* module performs evaluations (for example comparing an 'official ground-truth' annotations with that made by the player); it must be noticed that at each original dermoscopic image is associated a list of annotations that can be retrieved by the game algorithms.

At this stage, the game prototype lacks some advanced features and has a limited number of events and gaming tasks; the user interface is divided into three main sections (Fig. 5.2 and 5.3):

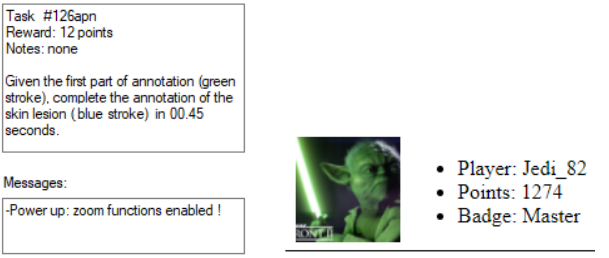
- upper section: shows the player profile (photo, nickname, points, badges, ...)
- middle section: tools to manage the image (Adapt, Center, Zoom) and to change the interaction mode (Erase or Draw) or the stroke features (color/width selection, annotation deleting)

Figure 5.2: The user interface during a gaming session.

- lower-left section: shows the main skin image with the superimposed strokes

- lower-right section: represents the interactive section of the game, in fact shows points and time (for time challenges), the setting of the task and the interactive messages occurred during the gaming session; moreover there are three buttons to commit, reload or change the gaming session

At the beginning of a game session, an XML configuration file is send by the server *Game data* module to set up the user interface (for example to enable/disable buttons and widgets, instantiate the timer, load the image or draw default strokes). To manage game dynamics like *power-up* or the *increasing difficulty*, the game client implements a simple *event manager* that sends asynchronous messages to the server which in turn raises appropriate counter-events: for example, the end of the time involves a game failure, a *zoom event* will enable/disable the corresponding buttons, a *time X event* will increase/decrease the timer by X seconds, a *speed X event* will increase/decrease the mouse/pen speed by an amount of X.



(a) The task section. (b) The player profile.

Figure 5.3: Sections of the User Interface.

When a student commits his work, the *Game algorithms* server module computes the corresponding rewards and updates the total points of the player in the leaderboard rank of the registered gamers. As seen before, the reward for a task is calculated with a pixel comparison between annotations: the first is that made by the player while the second is that selected by the system from MIMS; the evaluation metric used is the *f1-score* that, as seen in Cap. 3, results to be the most suitable to compare the dermoscopic annotations constituted by pixel masks.

Chapter 6

Conclusions

In this thesis has been presented a wide view about the affective ludology combining theory with practice by introducing data analysis strategies for sensor devices and interactive multimedia; after that has been faced the topic about dermoscopic image segmentation and management that, nowadays, is reaching a lot of interest in the fields of image processing, computer vision and specially machine learning.

The study about affective ludology seen in Cap. 2 presents an organic and progressive review to give a complete view of the topic spreading from an historical introduction with the definition of terms and theories until the formalization and deployment of the affective game levels.

It has been proposed and tested an evaluation methodology that handles the time-duration issue of the gaming activities and that shows strong results and improvements during the different experimentations; it is noticeable the novelty in exploiting the five pre-classified EEG values coming from the Emotiv EPOC headset, giving in that way scientific valence to them for experimental studies and, moreover, the machine learning methods exploited for that studies are robust, reliable and computational cheap but reaching anyway results over 90%.

Goodness of the design and development phases is notable both for the game level structures and for their activities, highlighting the possibility to induce and manipulate emotions and affective states: this study in fact produces the variations of affective data that were expected and encouraged during the design and the experimental setup; from that premises comes

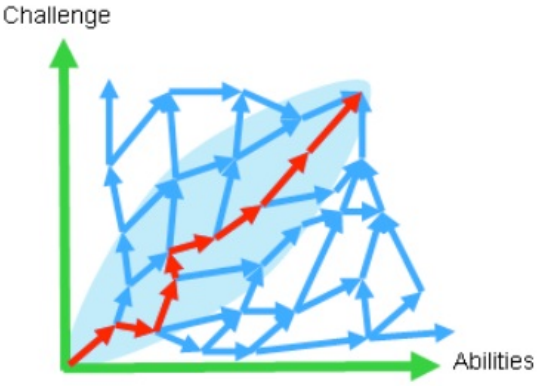


Figure 6.1: A Dynamic Difficulty Adjustment mechanisms to keep the Flow state.

that the proposed methodology it is easy to extend to real-world studies, for example by designing different time-limited versions of an interactive activity, with the aim to integrate classic evaluation methodologies with objective physiological data.

Despite the results obtained, several open problems remain allowing further works, starting with the improvement of the segmentation results obtained by the CNNs on medical images, followed by the raw affective data analysis to extract brainwave trends, followed by a comparison with that provided by the internal algorithms used here; moreover studies involving psychology and sociology experts should inquire about emotional differences between male and female subjects (and gamers) influencing economic studies of product-placement and market segmentation. From the screen captures taken during the gaming sessions might be automatically extracted gameplay dynamics like strategies and behavior patterns as in [92] so that the Artificial Intelligence of NPC and software agents can become more reactive toward the real people and environments with which they interface.

Forms of Dynamic Difficulty Adjustment (DDA) mechanisms [18] can benefit from this work, in fact with a real-time classification it could be possible to change an affective mood during a gaming session in order to always keep in the flow state the player experience (Fig. 6.1).

The design of a serious videogame seen in Cap. 5 of this work under-

lines the interesting aspect of the pedagogical use of digital libraries and interactive multimedia, with the aim to help high-specialized professionals like physicians in the delicate task of teaching by arousing involvement and engagement to transmit tacit knowledge hard to express with standard educational modes. The prototype presented has to be improved and tested with evaluation studies on the field by involving large amount of medical students to identify and compare progressions between the "standard" learning and the "gamified" one using qualitative and quantitative data; moreover the expansion of the presented gamification process to other medical specializations that involve annotation protocols may well deserve further insights as well as forms of cooperative and multiplayer features that could expand the learning experience improving the collaboration between experts.

The architecture for medical image management seen in Cap. 3 has the goal to achieve the entire pipeline that ranges from image gathering and annotation to the analysis and visualization of the associated metadata, with the aim to create and manage a proactive agent that supports dermatologists in their decisions and diagnoses.

Regarding the annotation tool, the evaluation evidences a good design due to the user-centered approach that supports its use for the data gathering and management: after observing common unskilled users, future evaluations will include the academic dermatologists and students during the daily activities in their working environment.

Raw and structured visual features coming from the manual annotations have been treated with machine learning techniques and image processing algorithms defining the patterns used as ground-truth for the automatic detection and prevention of skin melanomas; the evaluation metrics proposed result adequate to verify and measure the segmentation goodness, reaching positive performances while dealing with complex dermoscopic images but further evaluations are needed considering the variety and stratifications of skin lesion patterns. The breadth of this topic opens to future works that imply the improvement of approaches by exploiting Artificial Intelligence methods for advanced predictive performances and, moreover, by developing the specific modules dedicated to info-visualization and query search facilities.

Chapter 7

Publications

In this section we briefly report the research papers published during my PhD period.

"Layout Analysis and Content Classification in Digitized Books"
Corbelli Andrea, Baraldi Lorenzo, Balducci Fabrizio, Grana Costantino and Cucchiara Rita
Proceedings of IRCDL 2016 - 12th Italian Research Conference on Digital Libraries, Florence (Italy), 4-5 February 2016 in "Digital Libraries and Multimedia Archives", "Communications in Computer and Information Science" series, Springer, vol. 701 pp.153-165

Automatic layout analysis has proven to be extremely important in the process of digitization of large amounts of documents. In this paper we present a mixed approach to layout analysis, introducing a SVM-aided layout segmentation process and a classification process based on local and geometrical features. The final output of the automatic analysis algorithm is a complete and structured annotation in JSON format, containing the digitalized text as well as all the references to the illustrations of the input page, and which can be used by visualization interfaces as well as annotation interfaces. We evaluate our algorithm on a large dataset built upon the first volume of the "Enciclopedia Treccani".

"Classification of Affective Data to Evaluate the Level Design in a Role-playing Videogame"
Balducci Fabrizio, Grana Costantino and Cucchiara Rita
Proceedings of VS-Games 2015 - 7th International Conference on Virtual Worlds and Games for Serious Applications, Skövde (Sweden), 16-18 September 2015, IEEE Publications, pp.1-8

This paper presents a novel approach to evaluate game level design strategies, applied to role-playing games. Following a set of well defined guidelines, two game levels were designed for Neverwinter Nights 2 to manipulate particular emotions like boredom or flow, and tested by 13 subjects wearing a brain-computer interface helmet. A set of features was extracted from the affective data logs and used to classify different parts of the gaming sessions and to verify the correspondence of the original levels aims and the effective results on people emotions. The very interesting correlations observed, suggest that the technique is extensible to other similar evaluation tasks.

"Affective Level Design for a Role-Playing Videogame Evaluated by a Brain-Computer Interface and Machine Learning Methods"
Balducci Fabrizio, Grana Costantino and Cucchiara Rita
The Visual Computer - International Journal of Computer Graphics, Springer, April 2017, Vol. 33, Issue 4, pp.413-427

Game science has become a research field, which attracts industry attention due to a worldwide rich sell-market. To understand the player experience, concepts like flow or boredom mental states require formalization and empirical investigation, taking advantage of the objective data that psychophysiological methods like electroencephalography (EEG) can provide. This work studies the affective ludology and shows two different game levels for Neverwinter Nights 2 developed with the aim to manipulate emotions; two sets of affective design guidelines are presented, with a rigorous formalization that considers the characteristics of role-playing genre and its specific gameplay. An empirical investigation with a brain-computer interface headset has been conducted: by extracting numerical data features, machine learning techniques classify the different activities of the gaming sessions (tasks and events) to verify if their design differentiation coincides with the affective one. The observed results, also supported by subjective questionnaire data, confirm the goodness of the proposed guidelines, suggesting that this evaluation methodology could be extended to other evaluation tasks.

"Affective Classification of Gaming Activities Coming from RPG Gaming Sessions"
Balducci Fabrizio, Grana Costantino
Proceedings of EDUTAINMENT 2017 - 11th International conference on E-Learning and Games, Bournemouth (UK), 26-28 June 2017 in "E-Learning and Games", "Lecture Notes in Computer Science" series, Springer, vol. 10345 pp.93-100

Each human activity involves feelings and subjective emotions: different people will perform and sense the same task with different outcomes and experience; to understand this experience, concepts like Flow or Boredom must be investigated using objective data provided by methods like electroencephalography. This work carries on the analysis of EEG data coming from brain-computer interface and videogame the "Neverwinter Nights 2": we propose an experimental methodology comparing results coming from different off-the-shelf machine learning techniques, employed on the gaming activities to check if each affective state corresponds to the hypothesis fixed in their formal design guidelines.

"An Annotation Tool for a Digital Library System of Epidermal Data"
Balducci Fabrizio, Borghi Guido
Proceedings of IRCDL 2017 - 13th Italian Research Conference on Digital Libraries, Modena (Italy), 26-27 January 2017 in "Digital Libraries and Archives", "Communications in Computer and Information Science" series, Springer, vol.733 pp.173-186

Melanoma is one of the deadliest form of skin cancers so it becomes crucial the developing of automated systems that analyze and investigate epidermal images to early identify them also reducing unnecessary medical exams. A key element is the availability of user-friendly annotation tools that can be used by non-IT experts to produce well-annotated and high-quality medical data. In this work, we present an annotation tool to manually crate and annotate digital epidermal images, with the aim to extract meta-data (annotations, contour patterns and intersections, color information) stored and organized in an integrated digital library. This tool is obtained following rigid usability principles also based on physician interviews and opinions. A preliminary but functional evaluation phase has been conducted with non-medical subjects by using questionnaires, in order to check the general usability and the efficacy of the proposed tool.

"Pixel Classification Methods to Detect Skin Lesions on Dermoscopic Medical Images"
Balducci Fabrizio, Grana Costantino
Proceedings of ICIAP 2017 - 19th International Conference on Image Analysis and Processing, Catania (Italy), 11-15 September 2017 in "Image Analysis and Processing", "Lecture Notes in Computer Science" series, Springer, vol. 10485 pp.444-455

In recent years the interest of biomedical and computer vision communities in acquisition and analysis of epidermal images increased because melanoma is one of the deadliest form of skin cancer and its early identication could save lives reducing unnecessary medical treatments. User-friendly automatic tools can be very useful for physicians and dermatologists, in fact high-resolution images and their annotated data, combined with analysis pipelines and machine learning techniques, represent the base to develop intelligent and proactive diagnostic systems. In this work we present two skin lesion detection pipelines on dermoscopic medical images, by exploiting standard techniques combined with workarounds that improve results; moreover to highlight the performance we consider a set of metrics combined with pixel labeling and classication. A functional evaluation phase has been conducted with a sub-set of hard-to-treat images, in order to check which proposed detection pipeline reaches the best results.

"Annote: A Serious Game for Medical Students to Approach Lesion Skin Images of a Digital Library"
Balducci Fabrizio
Proceedings of IRCDL 2018 - 14th Italian Research Conference on Digital Libraries, Udine (Italy), 25-26 January 2018 in "Digital Libraries and Multimedia Archives" (CCIS series), Springer, vol. 806

Nowadays it is claimed that one method to learn how to execute a task is to present it as a gaming activity: in this way a teacher can oer a safe and controlled environment for learners also arousing excitement and engagement. In this work we present the design of the serious game 'Annote', to exploit a medical digital library with the aim to help dermatologists to teach students how to approach the examination of skin lesion images to prevent melanomas.

Bibliography

[1] Brain lobes. http://www.treccani.it/enciclopedia/lobi-cerebrali_%28Dizionario-di-Medicina%29. Accessed: 2017-11-15. 13

[2] Brain lobes (image). http://www.childneurologyfoundation.org/disorders/focal-and-multifocal-seizures. Accessed: 2017-11-15. 13

[3] Brain waves. http://en.wikipedia.org/wiki/Electroencephalography. Accessed: 2017-11-15. 15

[4] Deeplab v2 segmentation in keras. https://github.com/DavideA/deeplabv2-keras. Accessed: 2017-12-12. 92

[5] Emotiv epoc software development kit and user manual for release 1.0.0.5, par. 3.4.3. 18, 30

[6] Stanford cs class cs231n: Convolutional neural networks for visual recognition. http://cs231n.github.io/. Accessed: 2017-12-12. 84

[7] What are the detections based on? how were the algorithms created. https://emotiv.zendesk.com/hc/en-us/articles/204843839. Accessed: 2017-11-22. 18

[8] Dullrazor: A software approach to hair removal from images. *Computers in Biology and Medicine*, 27(6):533 – 543, 1997. 71

[9] F. Aarnoutse, L. Peursum, and F. Dalpiaz. The evolution of advergames development: A study in the netherlands. In *Games Media Entertainment (GEM), 2014 IEEE*, pages 1–8, Oct 2014. 5

[10] Christopher Ahlberg, Christopher Williamson, and Ben Shneiderman. Dynamic queries for information exploration: An implementation and evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '92, pages 619–626, New York, NY, USA, 1992. ACM. 60

[11] Abder-Rahman A Ali and Thomas M Deserno. A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. In *SPIE Medical Imaging*, pages 83181I–83181I. International Society for Optics and Photonics, 2012. 58

[12] John L Andreassi. *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2000. 14

[13] C Ardito, P Buono, MF Costabile, R Lanzilotti, and T Pederson. Mobile games to foster the learning of history at archaeological sites. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 81–86. IEEE Computer Society, 2007. 5, 93

[14] Per Backlund, Henrik Engström, Mikael Johannesson, and Mikael Lebram. Games for traffic education: An experimental study of a game-based driving simulator. *Simulation & gaming*, 41(2):145–169, 2010. 93

[15] Nicholas A. Badcock, Petroula Mousikou, Yatin Mahajan, Peter de Lissa, Johnson Thie, and Genevieve McArthur. Validation of the emotiv epoc eeg gaming system for measuring research quality auditory erps. 2013. 18

[16] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 88

[17] Samy Bakheet. An svm framework for malignant melanoma detection based on optimized hog features. *Computation*, 5(1):4, 2017. 59

[18] A. Baldwin, D. Johnson, P. Wyeth, and P. Sweetser. A framework of dynamic difficulty adjustment in competitive multiplayer video games.

In *Games Innovation Conference (IGIC), 2013 IEEE International*, pages 16–19, Sept 2013. 102

[19] Catarina Barata, Margarida Ruela, Mariana Francisco, Teresa Mendonça, and Jorge S Marques. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, 8(3):965–979, 2014. 59

[20] H. S. Baweja and T. Parhar. Leprosy lesion recognition using convolutional neural networks. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 141–145, July 2016. 59

[21] Jessica D. Bayliss and Kevin Bierre. Game design and development students: Who are they? In *Proceedings of the 3rd International Conference on Game Development in Computer Science Education*, GDCSE '08, pages 6–10. ACM, 2008. 5

[22] K. Berens and G. Howard. *The Rough Guide to Videogaming*. Miniguides Series. Rough Guides, 2002. 7

[23] E. J. Bergervoet, F. Sluis, E. M. A. G. Dijk, and A. Nijholt. Bombs, fish, and coral reefs. *The Visual Computer*, 29(2):99–110, 2012. 5

[24] Maresa Bertolo and Vanessa De Luca. Urban games to design the augmented city. 01 2012. 93

[25] Daire O Broin. Using a criteria-based user model for facilitating flow in serious games. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2011 Third International Conference on*, pages 63–69. IEEE, 2011. 23

[26] Emily Brown and Paul Cairns. A grounded investigation of game immersion. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1297–1300. ACM, 2004. 20

[27] Paolo Buono, Tiziana Cortese, Fabrizio Lionetti, Marco Minoia, and Adalberto Simeone. A simulation of a fire accident in second life. 93

[28] Paolo Burelli. Virtual cinematography in games: Investigating the impact on player experience. In *International Conference on the*

Foundations of Digital Games, Crete, Greece, May 14-17, 2013, pages 134–141, 2013. 23

[29] J. W. Burke, M. D. J. McNeill, D. K. Charles, P. J. Morrow, J. H. Crosbie, and S. M. McDonough. Optimising engagement for stroke rehabilitation using serious games. *The Visual Computer*, 25(12):1085–1099, October 2009. 17

[30] M. J. Callaghan, N. McShane, A. G. Eguluz, T. Teills, and P. Raspail. Practical application of the learning mechanics-game mechanics (lm-gm) framework for serious games analysis in engineering education. In *2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV)*, pages 391–395, Feb 2016. 95

[31] Gordon Calleja. *In-Game: From Immersion to Incorporation*. The MIT Press, 2011. 20

[32] Gordon Calleja. Narrative involvement in digital games. In *Conference proceedings from Foundations of Digital Games. Chania, Crete, Greece*, 2013. 10

[33] Alessandro Canossa, Anders Drachen, and Janus Rau Møller Sørensen. Arrrgghh!!!: Blending quantitative and qualitative methods to detect player frustration. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, pages 61–68. ACM, 2011. 12

[34] F. Carlà. *Space invaders: la vera storia dei videogames*. Castelvecchi, 1996. 7

[35] Valeria Carofiglio and Fabio Abbattista. A rough bci-based assessment of user's emotions for interface adaptation: Application to a 3d-virtual-environment exploration task. In *Proceedings of the First International Workshop on Intelligent User Interfaces: Artificial Intelligence meets Human Computer Interaction (AI*HCI 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 4, 2013.*, 2013. 17

[36] Marcus Carter, John Downs, Bjorn Nansen, Mitchell Harrop, and Martin Gibbs. Paradigms of games research in hci: A review of 10 years of research at chi. In *Proceedings of the First ACM SIGCHI*

Annual Symposium on Computer-human Interaction in Play, CHI PLAY '14, pages 27–36. ACM, 2014. 5

[37] M Emre Celebi, Quan Wen, Hitoshi Iyatomi, Kouhei Shimizu, Huiyu Zhou, and Gerald Schaefer. A state-of-the-art survey on lesion border detection in dermoscopy images. In *Dermoscopy Image Analysis*, pages 97–129. CRC Press, 2015. 58

[38] Alan Chalmers, Kurt Debattista, and Belma Ramic-Brkic. Towards high-fidelity multi-sensory virtual environments. *The Visual Computer*, 25(12):1101–1108, October 2009. 20

[39] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, MindTrek '08, pages 13–17. ACM, 2008. 12

[40] Jenova Chen. Flow in games (and everything else). *Commun. ACM*, 50(4):31–34, April 2007. 21, 93, 94

[41] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016. 88

[42] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 88

[43] Gifford K. Cheung, Thomas Zimmermann, and Nachiappan Nagappan. The first hour experience: How the initial play can engage (or lose) new players. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, CHI PLAY '14, pages 57–66. ACM, 2014. 6

[44] A. Choo and A. May. Virtual mindfulness meditation: Virtual reality and electroencephalography for health gamification. In *Games Media Entertainment (GEM), 2014 IEEE*, pages 1–3, Oct 2014. 19

[45] Nikolay Chumerin, Nikolay V Manyakov, Adrien Combaz, Arne Robben, Marijn van Vliet, and Marc M Van Hulle. Steady state visual evoked potential based computer gaming - the maze. In *Intelligent Technologies for Interactive Entertainment*, pages 28–37. Springer, 2012. 18

[46] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 118–126. Springer, 2015. 59

[47] Noel C. F. Codella, Quoc-Bao Nguyen, Sharath Pankanti, David Gutman, Brian Helba, Allan Halpern, and John R. Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *CoRR*, abs/1610.04662, 2016. 57, 58, 59, 92

[48] Oscar Colteli, Ximo Grandi, Ricardo Tosca, Pedro Latorre, José Salvador Sánchez, Luís Vicente Lizán, Francisco Ros-Bernal, and Conrrado Martínez-Cadenas. Designing serious games for learning support in medicine studies: A specific method to elicit and formalize requirements. In *Frontiers in Education Conference (FIE), 2014 IEEE*, pages 1–4. IEEE, 2014. 95

[49] Paul Coulton, Carlos Garcia Wylie, and Will Bamford. Brain interaction for mobile games. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 37–44. ACM, 2011. 17

[50] Lindley Craigh. Game taxonomies: A high level framework for game analysis and design. *Gamasutra*, April 2003. 5, 10

[51] Rui Craveirinha and Licínio Roque. Looking for the heart of interactive media: Reflections on video games' emotional expression. In *Proceedings of the 3rd International Conference on Fun and Games*, Fun and Games '10, pages 8–17. ACM, 2010. 14

[52] Mihaly Csikszentmihalyi. Does being human matter? on some interpretive problems of comparative ludology. *Behavioral and Brain Sciences*, 5:160–160, 3 1982. 11

[53] Mihaly Csikszentmihalyi. *Beyond boredom and anxiety*. Jossey-Bass, 2000. 21, 93

[54] Dario Deponti, Dario Maggiorini, and Claudio E Palazzi. Smartphone's physiatric serious game. In *IEEE Int. Conf. on Serious Games and Applications for Health*, 2011. 93

[55] TL Diepgen and V Mahler. The epidemiology of skin cancer. *British Journal of Dermatology*, 146(s61):1–6, 2002. 57

[56] John Dollard, Neal E Miller, Leonard W Doob, Orval Hobart Mowrer, and Robert R Sears. *Frustration and aggression*. Yale University Press, 1939. 19

[57] Claire Dormann and Robert Biddle. Understanding game design for affective learning. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, Future Play '08, pages 41–48. ACM, 2008. 12

[58] M Emre Celebi, Quan Wen, Sae Hwang, Hitoshi Iyatomi, and Gerald Schaefer. Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Research and Technology*, 19(1):e252–e258, 2013. 58

[59] Haidi Fan, Fengying Xie, Yang Li, Zhiguo Jiang, and Jie Liu. Automatic segmentation of dermoscopy images using saliency combined with otsu threshold. *Computers in Biology and Medicine*, pages –, 2017. 58

[60] Cynthia D Fisher. Boredom at work: A neglected concept. *Human Relations*, 46(3):395–417, 1993. 21

[61] Gonzalo Frasca. Simulation versus narrative: Introduction to ludology. In *The Video Game Theory Reader*, pages 221–236. Routledge, 2003. 10, 18

[62] James Paul Gee. Learning and games. *The ecology of games: Connecting youth, games, and learning*, 3:21–40, 2008. 11

[63] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. 39

[64] James J Gibson and Robert Shaw. Perceiving, acting, and knowing: Toward an ecological psychology. *The Theory of Affordances*, pages 67–82, 1977. 63

[65] Kiel Mark Gilleade, Alan Dix, and Jen Allanson. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In *In The 2005 International Conference on Changing Views: Worlds in Play*, 2005. 12

[66] Frasca Gonzalo. Ludologists love stories, too: notes from a debate that never took place. In *Proceedings of the 2003 DiGRA International Conference*, 2014. 11

[67] David Grimes, Desney S Tan, Scott E Hudson, Pradeep Shenoy, and Rajesh PN Rao. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 835–844. ACM, 2008. 17

[68] Christoph Groenegress, Bernhard Spanlang, and Mel Slater. The physiological mirror: a system for unconscious control of a virtual environment through physiological activity. *The Visual Computer*, 26(6-8):649–657, 2010. 12

[69] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 59, 91

[70] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does gamification work?–a literature review of empirical studies on gamification. In *Hawaii Int. Conf. on System Sciences*, 2014. 94

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 59

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 88

[73] Laura Herrewijn, Karolien Poels, and Gordon Calleja. The relationship between player involvement and immersion: An experimental investigation. In *FDG*, pages 364–367, 2013. 12

[74] Matthew Horsfall and Andreas Oikonomou. A study of how different game play aspects can affect the popularity of role-playing video games. In *CGAMES*, pages 63–69. IEEE Computer Society, 2011. 20

[75] Kai Huotari and Juho Hamari. Defining gamification-a service marketing perspective. *system*, 1(2):3–4, 2012. 93

[76] Eric Jacopin. Game ai planning analytics: The case of three first-person shooters. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2014. 12

[77] Bradley James, Barbara Fletcher, and Nia Wearn. Three corners of reward in computer games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games*, 2013. 23

[78] H Jasper. Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr Clin Neurophysiol*, 10:370–375, 1958. 14

[79] Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and defining the experience of immersion in games. *Int. J. Hum.-Comput. Stud.*, 66(9):641–661, September 2008. 20

[80] Jesper Juul. In search of lost time: On game goals and failure costs. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, pages 86–91. ACM, 2010. 23

[81] Amalia Kallergi and Fons J. Verbeek. Video games for collection exploration: Games for and out of data repositories. In *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '10, pages 143–146. ACM, 2010. 5

[82] Kyung-Kyu Kang, Jung-A Kim, and Dongho Kim. Development of a sensory gate-ball game system for the aged people. *The Visual Computer*, 25(12):1073–1083, 2009. 17

[83] Veli-Matti Karhulahti. Mechanic/aesthetic videogame genres: Adventure and adventure. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 71–74. ACM, 2011. 20

[84] Kaveh Khaleghi and Artur Lugmayr. Video game market segmentation based on user behavior. In *Proceeding of the 16th International Academic MindTrek Conference*, MindTrek '12, pages 283–286. ACM, 2012. 7

[85] Harold Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002. 57

[86] Hannu Korhonen, Janne Paavilainen, and Hannamari Saarenpää. Expert review method in game evaluations: Comparison of two playability heuristic sets. In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, MindTrek '09, pages 74–81. ACM, 2009. 12

[87] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 59

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, 2012. 83

[89] Beth Aileen Lameman, Magy Seif El-Nasr, Anders Drachen, Wendy Foster, Dinara Moura, and Bardia Aghabeigi. User studies: a strategy towards a successful industry-academic relationship. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology*, pages 134–142. ACM, 2010. 5

[90] R. Langer, M. Hancock, and S. D. Scott. Suspenseful design: Engaging emotionally with complex applications through compelling narratives. In *Games Media Entertainment (GEM), 2014 IEEE*, pages 1–8, Oct 2014. 10

[91] Petri Lankoski. Models for story consistency and interestingness in single-player rpgs. In *Proceedings of International Conference on*

Making Sense of Converging Media, AcademicMindTrek '13, pages 246:246–246:253. ACM, 2013. 19

[92] Geoffrey W Lee, Fabio Zambetta, Xiaodong Li, and Antonio G Paolini. Utilising reinforcement learning to develop strategies for driving auditory neural implants. *Journal of neural engineering*, 13(4):046027, 2016. 102

[93] Chris Lewis, Jim Whitehead, and Noah Wardrip-Fruin. What went wrong: A taxonomy of video game bugs. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, pages 108–115. ACM, 2010. 7

[94] Fotis Liarokapis, Athanasios Vourvopoulos, Alina Ene, and Panagiotis Petridis. Assessing brain-computer interfaces for controlling serious games. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on*, pages 1–4. IEEE, 2013. 17

[95] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, Nov 2015. 88

[96] Zhao Liu, Jiuai Sun, Melvyn Smith, Lyndon Smith, and Robert Warr. Unsupervised sub-segmentation for pigmented skin lesions. *Skin Research and Technology*, 18(1):77–87, 2012. 58

[97] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 59, 83

[98] Fabien Lotte. Brain-computer interfaces for 3d games: Hype or hope? In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, pages 325–327. ACM, 2011. 17

[99] Arnold M Lund. Measuring usability with the use questionnaire12. 2001. 67

[100] M. C. Machado, E. P. C. Fantini, and L. Chaimowicz. Player modeling: Towards a common taxonomy. In *Computer Games (CGAMES), 2011 16th International Conference on*, pages 50–57, July 2011. 7

[101] Dario Maggiorini, Christian Quadri, and Laura Anna Ripamonti. Opportunistic mobile games using public transportation systems: a deployability study. *Multimedia Systems*, 20(5):545–562, 2014. 93

[102] Ilias Maglogiannis and Charalampos N Doukas. Overview of advanced computer vision systems for skin lesions characterization. *IEEE transactions on information technology in biomedicine*, 13(5):721–733, 2009. 58, 84

[103] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg. Combining deep learning and hand-crafted features for skin lesion classification. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Dec 2016. 59, 87

[104] Regan L. Mandryk and Kori M. Inkpen. Physiological indicators for the evaluation of co-located collaborative play. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 102–111. ACM, 2004. 14

[105] Raphaël Marczak, Jasper van Vught, Gareth Schott, and Lennart E. Nacke. Feedback-based gameplay metrics: Measuring player experience via automatic visual analysis. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System*, IE '12, pages 6:1–6:10. ACM. 12

[106] Alan Mattiassi. Command systems and player-avatar interaction in successful fighting games in light of neuroscientific theories and models. 09 2017. 12

[107] F. Mäyrä. *An Introduction to Game Studies.* SAGE Publications, 2008. 97

[108] Monica McGill. Critical skills for game developers: An analysis of skills sought by industry. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, Future Play '08, pages 89–96. ACM, 2008. 5

[109] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5437–5440. IEEE, 2013. 59

[110] David Myers et al. Computer games genres. *Play & Culture*, 3(4):286–301, 1990. 7

[111] L.E. Nacke, S. Stellmach, and C.A. Lindley. Electroencephalographic assessment of player experience: A pilot study in affective ludology. *Simulation and Gaming*, 42(5):632–655, 2011. 20, 21

[112] Lennart Nacke. *Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment*. PhD thesis, Blekinge Institute of Technology, Karlskrona, Sweden, 2009. http://phd.acagamic.com, (ISBN) 978-91-7295-169-3. 11, 20

[113] Jeanne Nakamura and Mihaly Csikszentmihalyi. The concept of flow. *Handbook of positive psychology*, pages 89–105, 2002. 23, 93

[114] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian. Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1373–1376, Aug 2016. 86

[115] J. Adam Noah, Atsumichi Tachibana, and Shaw Bronner. Multi-core processing within the frontal lobe. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, pages 280–282. ACM, 2011. 12

[116] Pedro Nogueira, Rui Rodrigues, Eugnio Oliveira, and Lennart Nacke. Guided emotional state regulation: Understanding and shaping player's affective experiences in digital games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2013. 20

[117] Standard Electrode Position Nomenclature. American electroencephalographic society guidelines for. *Journal of clinical Neurophysiology*, 8(2):200–202, 1991. 14

[118] Donald A. Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43, May 1999. 63

[119] Michel Obbink, Hayrettin Gürkök, Danny Plass-Oude Bos, Gido Hakvoort, Mannes Poel, and Anton Nijholt. Social interaction in a cooperative brain-computer interface game. In *Intelligent Technologies for Interactive Entertainment*, pages 183–192. Springer, 2012. 33

[120] Janne Paavilainen. Critical review on video game evaluation heuristics: Social games perspective. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology*, Futureplay '10, pages 56–65. ACM, 2010. 12

[121] Jung-Yong Park and Jong-Hee Park. A graph-based representation of game scenarios; methodology for minimizing anomalies in computer game. *The Visual Computer*, 26(6-8):595–605, 2010. 20

[122] Francesco Peruch, Federica Bogo, Michele Bonazza, Vincenzo-Maria Cappelleri, and Enoch Peserico. Simpler, faster, more accurate melanocytic lesion segmentation through meds. *IEEE Transactions on Biomedical Engineering*, 61(2):557–565, 2014. 58

[123] Rosalind W. Picard. Affective computing. Technical Report 321, M.I.T Media Laboratory Perceptual Computing Section, 1995. 11

[124] David Pinelle, Nelson Wong, and Tadeusz Stach. Using genres to customize usability evaluations of video games. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, Future Play '08, pages 129–136. ACM, 2008. 7

[125] Anton Plotnikov, Natallia Stakheika, Carlotta Schatten, F Belotti, D Pranantha, R Berta, and A De Gloria. Measuring enjoyment in games through electroencephalogram (eeg) signal analysis. In *Proceedings of the 6th European Conference on Games-Based Learning (ECGBL 2012)*, pages 393–400, 2012. 12

[126] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(03):715–734, 2005. 16

[127] Robin Potanin. Forces in play: The business and culture of videogame production. In *Proceedings of the 3rd International Conference on Fun and Games*, Fun and Games '10, pages 135–143. ACM, 2010. 5

[128] Payam Aghaei Pour, Tauseef Gulrez, Omar AlZoubi, Gaetano Gargiulo, and Rafael A Calvo. Brain-computer interface: Next generation thought controlled distributed video game development platform. In *Computational Intelligence and Games, 2008. CIG'08. IEEE Symposium On*, pages 251–257. IEEE, 2008. 17

[129] W. Rasheed, T. B. Tang, and N. H. Bin Hamid. Default mode functional connectivity estimation and visualization framework for meg data. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1116–1119, April 2015. 14

[130] Mojdeh Rastgoo, Olivier Morel, Franck Marzani, and Rafael Garcia. Ensemble approach for differentiation of malignant melanoma. In *The International Conference on Quality Control by Artificial Vision 2015*, pages 953415–953415. International Society for Optics and Photonics, 2015. 59

[131] Robert W. A. Rawn and David R. Brodbeck. Examining the relationship between game type, player disposition and aggression. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, Future Play '08, pages 208–211. ACM, 2008. 12

[132] Stefan Rilling and Ulrich Wechselberger. A framework to meet didactical requirements for serious game design. *The Visual Computer*, 27(4):287–297, 2011. 12

[133] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 16

[134] A. Scavarelli and A. Arya. Cindr: A proposed framework for ethical systems in video games. In *Games Media Entertainment (GEM), 2014 IEEE*, pages 1–5, Oct 2014. 5, 7

[135] Gerald Schaefer, Bartosz Krawczyk, M Emre Celebi, and Hitoshi Iyatomi. An ensemble classification approach for melanoma diagnosis. *Memetic Computing*, 6(4):233–240, 2014. 59

[136] David Schwarz, Vivek Subramanian, Katie Zhuang, and Christine Adamczyk. Educational neurogaming: Eeg-controlled videogames as interactive teaching tools for introductory neuroscience. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2014. 18, 33

[137] Stefania Seidenari, Giovanni Pellacani, and Costantino Grana. Early detection of melanoma by image analysis. In *Bioengineering of the Skin: Skin Imaging & Analysis*, pages 305–312. CRC Press, 2006. 58

[138] Ben Shneiderman. 1.1 direct manipulation: a step beyond programming languages. *Sparks of innovation in human-computer interaction*, 17:1993, 1993. 64

[139] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010. 62

[140] S Singh, JH Stevenson, and D McGurty. An evaluation of polaroid photographic imaging for cutaneous-lesion referrals to an outpatient clinic: a pilot study. *British journal of plastic surgery*, 54(2):140–143, 2001. 58

[141] John Snowdon and Andreas Oikonomou. Games as a new medium for social criticism. In *16th International Conference on Computer Games, CGAMES 2011, Louisville, KY, USA, 27-30 July, 2011*, pages 101–106, 2011. 5

[142] Ingmar Stieger. Neverwinter nights extender v.4. `http://www.nwnx.org/`. 30

[143] William O. Tatum. Ellen r. grass lecture: Extraordinary eeg. *The Neurodiagnostic Journal*, 54(1):3–21, 2014. 15

[144] Li Ping Thong. Situated learning with role-playing games to improve transfer of learning in tertiary education classrooms. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2014 6th International Conference on*, pages 1–5. IEEE, 2014. 20

[145] Anders Tychsen, Michael Hitchens, and Thea Brolund. Motivations for play in computer role-playing games. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*, Future Play '08, pages 57–64. ACM, 2008. 20

[146] Heikki Tyni, Annakaisa Kultima, and Frans Mäyrä. Dimensions of hybrid in playful products. In *Proceedings of International Conference on Making Sense of Converging Media*, AcademicMindTrek '13, pages 237:237–237:244. ACM, 2013. 5

[147] Gary Ushaw, Janet Eyre, and Graham Morgan. A paradigm for the development of serious games for health as benefit delivery systems. In *Serious Games and Applications for Health (SeGAH), 2017 IEEE 5th International Conference on*, pages 1–8. IEEE, 2017. 95

[148] Vanus Vachiratamporn, Koichi Moriyama, Ken-ichi Fukui, and Masayuki Numao. An implementation of affective adaptation in survival horror games. In *2014 IEEE Conference on Computational Intelligence and Games, CIG 2014, Dortmund, Germany, August 26-29, 2014*, pages 1–8, 2014. 17, 33

[149] Juha-Matti Vanhatupa. Guidelines for personalizing the player experience in computer role-playing games. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, pages 46–52, New York, NY, USA, 2011. ACM. 20

[150] Athanasios Vourvopoulos and Fotis Liarokapis. Brain-controlled NXT robot: Tele-operating a robot through brain electrical activity. In *Third International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), Athens, May 4-6, 2011*, pages 140–143. IEEE, 2011. 18

[151] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse. Using wireless eeg signals to assess memory workload in the n-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, June 2016. 18

[152] Henrik Warpefelt and Björn Stråat. Anti-heuristics for maintaining immersion through believable non-player characters. In *International Conference on the Foundations of Digital Games, Crete, Greece, May 14-17, 2013*, pages 455–456, 2013. 23

[153] Huaxin Wei. Embedded narrative in game design. In *Proceedings of the International Academic Conference on the Future of Game Design and Technology*, Futureplay '10, pages 247–250. ACM, 2010. 10

[154] P. Wighton, T. K. Lee, H. Lui, D. I. McLean, and M. S. Atkins. Generalizing common tasks in automated skin lesion diagnosis. *IEEE Transactions on Information Technology in Biomedicine*, 15(4):622–629, July 2011. 58

[155] Daniel Wilcox-Netepczuk. Immersion and realism in video games - the confused moniker of video game engrossment. In *CGAMES*, pages 92–95. IEEE Computer Society, 2013. 20

[156] Paul Wilkinson. Affective educational games: Utilizing emotions in game-based learning. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on*, pages 1–8. IEEE, 2013. 11

[157] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition, 2011. 40

[158] Y. Yuan, M. Chao, and Y. C. Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, PP(99):1–1, 2017. 59

[159] José P. Zagal and Roger Altizer. Examining rpg elements: Systems of character progression. In *Proceedings of the 2014 Conference on the Foundations of Digital Games*, 2014. 23

[160] José P. Zagal, Amanda Ladd, and Terris Johnson. Characterizing and understanding game reviews. In *Proceedings of the 4th International Conference on Foundations of Digital Games*, FDG '09, pages 215–222. ACM, 2009. 5

[161] Nelson Zagalo and Ana Torres. Character emotion experience in virtual environments. *The Visual Computer*, 24(11):981–986, 2008. 10

[162] Ezzeddine Zagrouba and Walid Barhoumi. A prelimary approach for the automated recognition of malignant melanoma. *Image Analysis & Stereology*, 23(2):121–135, 2011. 70