Università Politecnica delle Marche
Scuola di Dottorato di Ricerca in Scienze dell'Ingegneria
Corso di Dottorato in Ingegneria Industriale

Tecniche di Deep Learning per analizzare e migliorare la Customer Experience in contesti digitali e fisici

Ph.D. Dissertation of:

Andrea Generosi

Supervisor:

Prof. Maura Mengoni

Ph.D. Course coordinator:

Prof. G. Di Nicola

XVIII edition - new series

Università Politecnica delle Marche
Dipartimento di Ingegneria Industriale e Scienze Matematiche
Via Brecce Bianche — 60131 - Ancona, Italy

Ringraziamenti

Vorrei ringraziare la mia famiglia e soprattutto la mia compagna, che mi ha supportato ed aiutato costantemente in questi anni di dottorato. Ringrazio inoltre tutti i miei colleghi di Emoj e quelli universitari, il mio tutor di dottorato, sempre disponibile per qualsiasi problema, e tutte le persone che hanno incrociato e influito, chi più chi meno, sul mio cammino in questo periodo, contribuendo a farmi crescere e a rendermi la persona che sono.

Abstract

La trasformazione digitale che oggi interessa la maggior parte dei settori industriali ha avuto un impatto significativo nell'intero ecosistema del Retail, dalla produzione fino alla vendita e post-vendita di prodotti e servizi. Il presente lavoro di tesi affronta tale cambiamento proponendo una metodologia di gestione del retail basata sul concetto di Customer Experience ed innovativi strumenti fondati su sistemi di intelligenza artificiale, che insieme ed integrati sono atti a potenziare e rendere più efficace tale trasformazione.
Un fattore importante del progressivo mutamento del Retail riguarda proprio l'introduzione di tecnologie basate sull'intelligenza artificiale, capaci di raccogliere ed interpretare la gran mole di dati generati dai diversi canali di vendita e contatto con il cliente, per migliorare la conoscenza dei consumatori, sempre più al centro dell'intero ecosistema, predirne comportamenti, attitudini e preferenze e attivare esperienze personalizzate capaci di connetterli con il brand, con la conseguenza di incrementare la fidelizzazione, le vendite ed il tasso di conversione.
L'esperienza del cliente di un prodotto, di un servizio o semplicemente dell'ambiente in cui conosce il brand è diventata sempre più centrale in ogni processo di progettazione, produzione, vendita, distribuzione e assistenza che interessa l'ecosistema del retail.
Lo studio su come progettare e gestire attraverso nuove tecnologie la Customer Experience attraverso delle azioni puntuali nei diversi punti di contatto tra cliente e brand (touchpoint) è oggi la chiave per molti retailers per raggiungere il successo su un mercato pieno di sfide competitive.
In tutti i touchpoint il cliente interagisce con il brand attraverso i sensi primari e sono diversi i modi in cui esso reagisce agli stimoli ricevuti, sia razionalmente che istintivamente ed emotivamente: è in quest'ultimo caso soprattutto che il successo nell'avere una buona CX è fondamentale per garantire la fidelizzazione con il brand. Data la complessità nel riuscire a monitorare tutti i touchpoint, spesso, la progettazione di una corretta CX viene trascurata se non del tutto omessa.
Ad oggi alcune delle metodologie più utilizzate per analizzare il livello di gradimento di un'esperienza qualsiasi da parte di un utente/cliente risultano molto macchinose ed esose in termini di tempo e risorse impiegate. In questo contesto nasce la ricerca su cui si focalizza la tesi di dottorato, ossia trovare una tecnologia capace di automatizzare la raccolta dati, interpretarli per conoscere la risposta del cliente ad una serie di stimoli multisensoriali e multimediali ricevuti e attuare una CX adattativa che abiliti una connessione empatica ed un coinvolgimento con il brand. Per ottenere questo scopo la ricerca in questione farà uso di strumenti e tecnologie pervasive e non invasive, al fine di ottenere una grande quantità di dati (Big Data) nella maniera più "autentica" possibile, così da non contaminare i risultati introducendo bias non desiderati: ad oggi questi strumenti che fanno parte sempre più della nostra vita quotidiana sono le telecamere, dalle webcam alle camere integrate negli smartphone. Per poter utilizzare e sfruttare a pieno queste tecnologie subentrano finalmente le discipline della Computer Vision e soprattutto del Deep Learning, che permettono di analizzare flussi video e predirne il contenuto ed il suo significato esattamente come se fosse un essere umano a visionarli.

Abstract

The digital transformation that today affects most industrial sectors has had a significant impact on the entire retail ecosystem, from the production to the sale and after-sale of products and services. This thesis work addresses this change by proposing a retail management methodology based on the concept of Customer Experience and innovative tools based on artificial intelligence systems, which together and integrated are able to enhance and make this transformation more effective.
An important factor in the progressive changes in Retail concerns the introduction of technologies based on artificial intelligence, capable of collecting and interpreting a large amount of data generated by the various sales and customer contact channels, in order to improve the knowledge of the consumers, who are increasingly at the centre of the entire ecosystem, predict their behaviour, attitudes and preferences and activate personalised experiences capable of connecting them with the brand, with the result of increasing loyalty, sales and conversion rates.
The customer's experience of a product, service or simply the environment in which they know the brand has become increasingly crucial in every process of design, production, sales, distribution and service that affects the retail ecosystem.
The study on how to design and manage through new technologies the Customer Experience through precise actions in the different points of contact between customer and brand (touchpoints) is today the key for many retailers to achieve success in a market full of competitive challenges.
In all touchpoints the customer interacts with the brand through the primary senses and there are different ways in which it reacts to the received stimuli, both rationally and emotionally: it is especially in this last case that the success in having a good CX is fundamental to ensure brand loyalty. Given the complexity of being able to monitor all the touchpoints, the design of a correct CX is often neglected if not completely omitted.
Today, some of the most widely used methodologies to analyze the acceptance level of any experience by a user/customer, are very cumbersome and costly in terms of time and spent resources. In this context, the research on which this PhD thesis is focused is born, that is to find a technology able to automate data collection, interpret them to know the customer's response to a series of multisensory and multimedia stimuli and implement an adaptive CX that enables an empathic connection and involvement with the brand. To achieve this goal, this research will make use of pervasive and not intrusive tools and technologies, in order to obtain a large amount of data (Big Data) in the most "authentic" possible way, so to not contaminate the results by introducing unwanted bias: today these tools that are increasingly part of our daily life are the cameras, from webcams to integrated smartphone cameras. In order to fully use and exploit these technologies, the disciplines of Computer Vision and especially Deep Learning will help, allowing to analyze video streams and predict their content and meaning, exactly as if a human being were watching them.

Sommario

Lista delle figure

Lista delle tabelle

Capitolo 1.

Introduzione

L'intelligenza artificiale e l'analisi dei dati (soprattutto dei Big Data) negli ultimi anni ha rivoluzionato una gran parte dei settori industriali: basti pensare a quanto sta avvenendo nel settore dell'automotive (assistenza alla guida, rilevamento di stanchezza solo per citarne alcune), dell'automazione industriale (manutenzione predittiva delle macchine), alla Business Intelligence per predire gli andamenti del mercato e ottimizzare la gestione del prodotto interna all'azienda, di qualsiasi tipologia di mercato essa si occupi.
Il retail di pari passo è un settore che negli ultimi anni sta subendo una progressiva trasformazione nella direzione della digitalizzazione, soprattutto per quanto riguarda il campo degli e-commerce, che negli Stati Uniti ad esempio, ha costituito il 9% di tutte le vendite retail della nazione (Goolsbee et al., 2018).
Eccetto per l'analisi dei Big Data e della Business Intelligence che, come citato precedentemente, sta investendo praticamente ogni settore dell'industria moderna, il binomio intelligenza artificiale e retail riguarda ad oggi una strada poco battuta se si esclude il Machine Learning adottato per fare del Data Science, ossia l'interpretazione dei dati tramite modelli statistici e matematici in grado di ricavare pattern che permettono di fare predizioni su una conoscenza pregressa, su dati che il più delle volte riguardano puramente l'aspetto commerciale dei punti vendita.
Ciò che ad oggi non è stato ancora concettualizzato, tranne qualche caso isolato, è che l'intelligenza artificiale può essere utilizzata nel retail per acquisire in maniera automatizzata una grossa quantità di dati sulle caratteristiche e i comportamenti nelle fasi d'acquisto della clientela, dunque per acquisire dati e non solo per analizzarli.
In questo contesto subentra uno degli aspetti chiave nel campo del retail, ossia la Customer Experience: l'esperienza che un cliente si ritrova a provare durante la fase d'acquisto, intesa come la totalità delle singole esperienze provate in tutti i momenti in cui questo entra in contatto con un brand. Riuscire a progettare ed implementare la migliore delle Customer Experience possibili per i clienti è fondamentale per qualsiasi brand e proprietario di un'attività retail, poiché è fondamentalmente da essa che si stabilirà il successo o meno della fase di acquisto e della fidelizzazione del cliente con il brand o l'esercizio commerciale.
Tutti i punti chiave in cui un cliente entra in contatto con un brand, e che di fatto definiscono la buona riuscita o meno della progettazione di una buona Customer Experience, sono chiamati Touchpoint: questi "luoghi" (sia fisici che virtuali) sono fondamentali e al contempo innumerevoli: la pubblicità, l'acquisto, il trasporto/consegna, il re-purchase e assistenza sono solo alcuni dei tanti touchpoint esistenti in un processo di vendita, tanto che tutti assieme finiscono per formare un vero e proprio ecosistema. In tutti questi punti chiave il cliente interagisce con il brand attraverso i sensi primari e sono diversi i modi in cui esso reagisce agli stimoli ricevuti, sia razionalmente che istintivamente ed emotivamente: è in quest'ultimo

caso soprattutto che il successo nell'avere una buona CX è fondamentale, poiché un cliente che prova un'emozione negativa di fronte ad un touchpoint, il più delle volte, è un cliente perso; viceversa provare un'emozione positiva può a volte garantire la fidelizzazione con il brand. Data la complessità nel riuscire a monitorare tutti i touchpoint, spesso la progettazione di una corretta CX viene trascurata se non del tutto omessa, in particolare quando parliamo di "retail fisico", ossia quello classico, dove monitorare molti dei touchpoint esistenti significherebbe utilizzare una o più persone adibite a quello specifico ruolo.
Ad oggi alcune delle metodologie più utilizzate per analizzare il livello di gradimento di un'esperienza qualsiasi, da parte di un utente/cliente, è quella di effettuare dei focus group o di proporre questionari tramite un'interazione diretta con gli intervistati stessi: queste tecniche sono molto macchinose ed esose in termini di tempo e risorse impiegate e applicarle ad un così vasto numero di situazioni (interazioni con i touchpoint) sarebbe altamente controproducente per un'azienda.
In questo contesto nasce la ricerca su cui si focalizzerà questa tesi di dottorato, che si prefigge lo scopo di dare una risposta a queste specifiche problematiche in maniera innovativa ed efficace/efficiente per le aziende che ne soffrono: automatizzare la raccolta dati che permetterà di analizzare e ottenere un feedback elaborato rispetto agli obiettivi aziendali sulla Customer Experience dei clienti in contesti retail. Per ottenere questo scopo la ricerca in questione farà uso di strumenti e tecnologie PERVASIVE e NON INVASIVE, al fine di ottenere una grande quantità di dati (Big Data) nella maniera più "autentica" possibile, così da non contaminare i risultati introducendo bias non desiderati: ad oggi questi strumenti che fanno parte sempre più della nostra vita quotidiana sono le telecamere, dalle webcam alle camere integrate negli smartphone. Per poter utilizzare e sfruttare a pieno queste tecnologie subentrano finalmente le discipline della Computer Vision e soprattutto del Deep Learning, che permettono di analizzare flussi video e predirne il contenuto ed il suo significato esattamente come se fosse un essere umano a visionarli.

1.1. Contesto e obiettivi della ricerca

I rivenditori hanno iniziato a fare sempre più attenzione alla progettazione di servizi esperienziali per stimolare le emozioni del cliente e creare esperienze uniche nei negozi (Zomerdijk et al., 2010). Questo ha determinato uno spostamento dell'attenzione dalla progettazione del servizio a quello dell'esperienza del cliente (Chen-Yu et al., 2001) e conseguentemente ciò ha fatto si che il retail sia considerato non più solamente un negozio dove i prodotti vengono esibiti per la vendita, ma anche uno spazio dove possono anche verificarsi degli eventi (Giraldi et al., 2016).
Tuttavia, fornire un intrattenimento ed organizzare eventi creativi e divertenti non è abbastanza per garantire una soddisfacente Customer Experience. Le aziende dovrebbero gestire tutti gli stimoli che vengono inviati ai clienti in base ad una strategia ben concepita e comprensiva che riguardi la progettazione della CX. In generale, gli stimoli che possono incidere sulla CX sono ovunque in un negozio (dai colori delle pareti alle luci, dagli odori ai suoni nel negozio, lo stile degli espositori, le uniformi dei commessi ecc.).
Le odierne reti distribuite di sensori e la disponibilità di connessione internet all'interno dei negozi, forniscono un'interessante opportunità per i negozianti e per tutti i soggetti interessati dalla CX per osservare cosa fanno i clienti, come interagiscono con lo spazio e con gli altri

clienti e per capire cosa provano, come percepiscono gli stimoli, qual è il loro stato emotivo e perchè cambia se avviene un particolare evento.
Tutti questi dati possono rappresentare una base per:

- Sviluppare degli stimoli all'interno dei negozi, in grado di focalizzare l'attenzione del cliente e metterlo in un "mood" positivo
- Implementare delle strategie di gestione della CX capaci di influenzare le probabilità di acquisto (Berry et al., 2002), la soddisfazione del cliente (Meyer et al., 2007) e la fedeltà del cliente (Khuong et al., 2015).
- Migliorare il design visivo dei prodotti e la disposizione degli interni del negozio
- Fornire delle indicazioni per la progettazione del prodotto/servizio offerto.

Le tecnologie emergenti per il monitoraggio del comportamento, dell'emozione e dell'attenzione dei clienti stanno portando enormi possibilità alla progettazione della CX di prodotti e servizi per il retail (Achar et al., 2016), evidenziando la necessità di nuovi strumenti in grado di raccogliere dati grezzi, organizzare e rappresentare tali informazioni sulla base delle finalità lavorative degli stakeholder e proporre azioni adeguate per rendere l'esperienza di acquisto più coinvolgente, il prodotto più attraente, i servizi in grado di rispondere maggiormente alle esigenze individuali.
In questo contesto si punta ad ottenere un obiettivo principale: studiare e sviluppare un sistema intelligente a supporto della definizione, progettazione e gestione delle strategie per la CX, in modo da adattare la Shopping Experience nei vari Touchpoint in base al comportamento e allo stato emotivo del cliente riconosciuto.
Un passo significativo per raggiungere i suddetti obiettivi è la definizione di CX.

1.1.1. La Customer Experience

Meyer e Schwager definiscono la Customer Experience come "la reazione interiore e soggettiva del cliente di fronte ad un qualsiasi contatto diretto o indiretto con un'impresa".
La Customer Experience rappresenta dunque l'ultimo stadio evolutivo degli approcci di marketing e management ed è da intendersi come lo studio del modo con cui il cliente percepisce l'interazione con un determinato brand a livello sia conscio che inconscio. Contribuiscono a creare tale esperienza non solo le interazioni col prodotto stesso, ma tutta una serie di valutazioni che il cliente fa ogni qualvolta relaziona i propri bisogni e necessità con l'offerta: contatti con il sito web, l'immagine del punto vendita, l'operatore del call center durante l'assistenza, la cordialità e la professionalità del commesso del negozio, le brochure promozionali e così via.
Creare valore per il cliente significherà allora renderlo protagonista di una memorabile esperienza di acquisto emozionandolo tramite la stimolazione di tutti i suoi sensi.
La soddisfazione del cliente e della sua esperienza con l'azienda in ogni punto di contatto è fondamentale per costruire una loyalty robusta: migliore è l'esperienza, maggiore sarà il legame con l'azienda. Per avere clienti soddisfatti, fedeli e soprattutto profittevoli bisogna però iniziare a chiedersi e a capire se la loro esperienza con l'azienda è positiva e soprattutto se sono clienti "felici".
La motivazione del perché molte aziende mettono la Customer Experience alla base delle loro scelte strategiche risiede nel fatto che oggi si trovano a che fare sostanzialmente con un tipo di customer di nuova tipologia: sono ormai abituati ad usare strumenti mobile, sempre

alla ricerca di informazioni utili e nel minor tempo possibile, pretendenti ed impazienti ma allo stesso tempo dipendenti da dinamiche decisionali fondate sulla fiducia e percezione del valore (non del prezzo!) nella scelta del marchio.

1.1.2 I tre livelli di Customer Experience

Secondo H. Manning e K. Bodine (Manning et al., 2012), la CX si può rappresentare come una piramide a tre livelli: il primo considera quanto l'interazione risulta soddisfacente per il cliente rispetto le sue esigenze, il secondo riguarda il grado di complessità dell'interazione, ossia lo sforzo che il cliente deve compiere e quindi determinare quanto è stato facile interagire, mentre l'ultimo livello prende in esame la piacevolezza complessiva dell'interazione, cioè la misura in cui questa viene percepita come gradevole e rassicurante.
I clienti, quindi, giudicano la loro esperienza valutando le interazioni con l'azienda sulla base di tre criteri: soddisfazione delle esigenze, facilità e piacevolezza.

Figura 1: I tre livelli di giudizio del cliente nell'interazione con un'azienda

La capacità di soddisfare le esigenze di base dei clienti, collocata alla base della piramide, rappresenta una condizione imprescindibile, necessaria (ma non più sufficiente come poteva esserlo tempo fa) per la sopravvivenza stessa del business.

Al livello successivo si considera quanto risulti semplice per il cliente, interagire con l'azienda; ad esempio quanto sia facile l'acquisto di un prodotto, l'uso di un servizio, l'accesso all'assistenza tecnica, l'ottenimento di informazioni, ecc.
Il livello posto alla cima della piramide è quello in cui risulta massimo l'aspetto emozionale del cliente nella sua interazione con l'azienda, su tutti i livelli: nella definizione di coinvolgimento emotivo intervengono aspetti relazionali e psicologici come ad esempio il comportamento degli addetti con cui il cliente si interfaccia – con cortesia, rapidità, professionalità, capacità di risolvere un problema oppure con sgarbo e incompetenza.
Ogni volta che interagiscono con un prodotto, un servizio, una persona, un sistema automatico, un luogo fisico o virtuale, i clienti valutano in modo positivo o negativo la propria esperienza sulla base di: quanto l'interazione li ha aiutati a raggiungere i propri obiettivi, quanto tempo hanno dovuto dedicare ad essa e quanto l'hanno trovata piacevole.
Il consumatore considera ormai un requisito minimo la presenza di una qualità intrinseca nell'acquisto e si è drasticamente abbassata, sino ad annullarsi, la soglia di disponibilità ad accettare errori, ritardi e malfunzionamenti.
Per contro il consumatore pretende sempre una maggiore efficacia nella risoluzione dei problemi, trasparenza, rapidità, possibilità di scelta e personalizzazione.
La piramide a tre livelli della Customer Experience fornisce una chiara visualizzazione di come la soddisfazione del cliente, il livello base della piramide, rappresenti solamente uno degli elementi necessari per un'esperienza positiva.

1.2 L'innovazione proposta

In questo contesto, una delle più grandi criticità della Customer Experience riguarda il riuscire a monitorare tutti i punti dell'ecosistema formato da tutti i Touchpoint esistenti: lo scopo del sistema che verrà proposto è proprio quello di fornire una soluzione che introduca un'AUTOMATISMO tramite soluzioni tecnologiche allo stato dell'arte e soprattutto innovative dal punto di vista del software utilizzato.
Per ***monitorare*** si intende il riuscire a **catturare** ed **interpretare** due tipologie di reazioni del cliente:

- Reazioni emotive
- Reazioni comportamentali

Come affermato precedentemente per un sistema del genere è necessario che esso abbia tre principali caratteristiche:

- Pervasività
- Non invasività
- Capacità di raccogliere una gran quantità di dati in maniera automatica

Ad oggi, la tecnologia hardware che permette di raggiungere questi obiettivi per quanto riguarda la fase di acquisizione sono le telecamere, utilizzabili praticamene ovunque e, tranne in casi evidenti, non invasive rispetto a strumentazioni come braccialetti, caschetti per ECG e altre tipologie di sensori in grado di acquisire dati biometrici ma che richiedono all'utente di indossare qualcosa durante le fasi di acquisizione dati. Per quanto riguarda il terzo punto,

a livello software per poter analizzare le immagini acquisite dalle telecamere in tempo reale oggi giorno esistono le branche della Computer Vision, o Visione Artificiale, ma soprattutto il Deep Learning, figlia del Machine Learning e dunque dell'Intelligenza Artificiale, che permette di interpretare in maniera automatica il contenuto di immagini, audio ecc. sulla base delle esperienze acquisite in fase di addestramento.
Utilizzando queste tecnologie l'obiettivo è quello di analizzare, progettare ed implementare una piattaforma tecnologica in grado di risolvere le criticità ivi analizzate facendo affidamento ad una struttura totalmente modulare ed applicabile in un gran numero di contesti differenti e utilizzando un approccio al mercato "Omnichannel".

Capitolo 2.

Research Background e Stato dell'Arte

Come già accennato, l'esperienza del cliente può essere definita come la risposta della persona (interna e soggettiva) a tutte le interazioni (dirette o indirette) con un'azienda. Tale risposta è di natura olistica ed è determinata dalle risposte cognitive, affettive, emotive e sociali dei clienti agli stimoli percepiti durante l'interazione (Pine et al., 1999). In particolare, la CX è influenzata da tutti i prodotti e servizi o stimoli con cui i clienti entrano in contatto durante il loro viaggio (Grewal et al., 2009). Tutti questi elementi, solitamente definiti punti di contatto (Touchpoint), costituiscono sicuramente gli ambienti (o contesti) all'interno dei quali si svolge l'esperienza di acquisto. Se un'azienda è in grado di identificare correttamente i Touchpoint che influenzano maggiormente la Shopping Experience e di capire quali stimoli devono essere forniti per garantire la miglior CX, a seconda della natura del punto di contatto in cui avviene l'interazione, quell'azienda sarà effettivamente in grado di influenzare il cliente a scegliere o riacquistare i prodotti in modo tale da generare maggiori profitti per l'azienda. In questo contesto, per "prodotti" si intendono tutti i beni tangibili forniti dai rivenditori, come il negozio, le luci, gli espositori, il sito di e-commerce, ecc. In particolare, molti studi presentano modelli concettuali volti a supportare le strategie manageriali (Teixeira et al., 2012; Maguire et al., 2001; Stauss et al., 1997). Pochi studi considerano l'importanza dell'esperienza totale del cliente nella progettazione dei servizi (Bailenson et al., 2008), mentre mancano ancora metodologie orientate alla CX in grado di supportare in modo sistematico l'introduzione dei requisiti di CX nella progettazione di prodotti e servizi. A supporto della progettazione della CX, il ben noto approccio UCD (User-centered design), sembra essere adatto alle discipline multidisciplinari e ai metodi e agli strumenti che propone per assicurare che i prodotti soddisfino le aspettative degli utenti. Per coinvolgere il cliente nella progettazione di prodotti/servizi, si possono utilizzare tecnologie virtuali, l'UCD, tuttavia, si concentra sul punto di contatto specifico dell'uso del prodotto, non considerando l'interazione totale tra il cliente e l'azienda. Inoltre, si focalizza principalmente sulle esigenze di particolari clienti, ossia di coloro che utilizzeranno effettivamente il prodotto, in modo da non considerare sia le esigenze di tutti gli stakeholder, che il prodotto e i servizi devono soddisfare per garantire l'appagamento del cliente in ogni fase di CX, sia tutti i requisiti che il prodotto deve soddisfare per essere coerente con tutti i prodotti e i servizi correlati, secondo una strategia CX definita.

Per supportare l'identificazione e l'analisi delle caratteristiche dei punti di contatto che più determinano l'esperienza di acquisto, la maggior parte degli studi sfrutta la tecnica della mappatura dei viaggi. Molti studi propongono metodi per analizzare la risposta del cliente e misurare la qualità CX lungo il percorso del cliente sulla base di interviste e sulla costruzione della curva emozionale.

Sebbene diversi studi evidenzino l'importanza delle emozioni nel determinare l'esperienza del cliente, a nostra conoscenza, nessuno studio ha sfruttato gli strumenti di riconoscimento delle emozioni per monitorare l'esperienza dei clienti nel loro viaggio.

2.1. L'ecosistema Retail

Quello del retail è un mercato in cui una vasta gamma di prodotti e commodities sono messi a disposizione dei clienti. Di conseguenza, i consumatori sono continuamente esposti a nuove offerte e stimoli. Conseguentemente, il marketing e le sue leve devono necessariamente essere improntati sui meccanismi di decisione che si innescano nella mente dei consumatori, stimolando la loro parte emotiva ed esperienziale. Difatti, il 95% delle attività mentali risiedono nella mente inconscia, all'interno dei ricordi, pensieri, emozioni e altri processi cognitivi di cui i consumatori non sono consapevoli (Kagan, 2002).
Pertanto, anche i meccanismi decisionali di acquisto dipendono da questi importantissimi fattori.
Di fatto, negli ultimi anni si è passati da un focus orientato alla progettazione di prodotti e servizi al design Customer Shopping Experience, ad un focus indirizzato alla definizione di tutti gli indizi che le persone rilevano nei loro processi di shopping che lo rendono unico.
Fattori come la Digital Transformation e Omnicanalità stanno cambiando completamente il mondo del retail, tanto da cominciar a parlare di Retail 4.0.
È sotto gli occhi di tutti che i consumatori usano costantemente tecnologia fissa e mobile, e per questo motivo, sono influenzati non solo dai media tradizionali, ma anche e soprattutto dai social media. Partendo da questo concetto, il Retail 4.0 utilizza soluzioni ibride e iper-convergenti. In altre parole, il negozio cambia pelle mediante l'utilizzo di tecnologie sempre più intelligenti ed interattive. Negli ultimi anni si è assistito ad una progressiva trasformazione digitale di questo settore e non c'è più un modello di business improntato su una strategia Multi-Channel, ma oggi si parla sempre più di omnicanalità del mercato. Infatti, le aziende non considerano più la rete dei punti vendita (web, mobile e punti vendita) come dei canali separati, ma li interpretano con una visione olistica, senza alcuna distinzione tra canali. Conseguentemente, c'è una gestione unitaria dei canali di distribuzione, in modo tale da allinearsi ai clienti che costantemente utilizzano mezzi analogici e digitali.
Dunque, la filosofia che sta dietro tutto questo è la seguente: se il consumatore è diventato omnicanale, anche la distribuzione deve necessariamente esserlo, conciliando tecnologia e strategia. Un metodo per raggiungere tale traguardo è personalizzare i processi di vendita fino al punto da farli diventare dei servizi su misura, adottando sistemi di tracciabilità e rintracciabilità delle informazioni.
È il consumatore stesso che definisce il ruolo da attribuire a ciascun canale, le aspettative e le modalità con cui si relaziona con le aziende. Pertanto, è evidente che raccogliere, analizzare e interpretare messaggi e segnali dei consumatori è di cruciale importanza per disegnare un modello di business omnichannel.

2.2. La Journey Map: Touchpoint, strategie e metodi di rappresentazione

Le indagini effettuate recentemente dimostrano come la CX abbia una valenza *cross* channel: è il risultato di come il cliente percepisce il suo rapporto complessivo con l'azienda attraverso tutti i canali di contatto.
Per semplificare lo studio di questo complesso insieme di interazioni è stato ideato lo strumento della "mappa del cliente", la "*customer journey map"*: è un documento che illustra visivamente le tappe percorse dal cliente-tipo attraverso i diversi canali, intesi come momenti di contatto con l'azienda nel corso del tempo.
L'esperienza del cliente viene scomposta in singole interazioni, per rendere più facili da individuare le sue esigenze ed emozioni in relazione ai vari *touchpoint*, indicando quelli che hanno su di lui un impatto positivo e quelli che invece rappresentano una criticità, perché magari comportano insoddisfazioni e disagi. Ed è proprio su questi ultimi che l'azienda deve focalizzarsi per ottenere un miglioramento della CX.
Per rendere totale la comprensione della CX è però necessario allargare la mappa fino a includere tutti gli elementi del sistema aziendale che hanno influenza sul cliente, anche quelli che a lui non risultano visibili.
Ad esempio, nel caso in cui un cliente voglia segnalare un errore del conto telefonico ed ottenere un riaccredito, oltre ai touchpoint diretti come il conto cartaceo, il sito web, il risponditore automatico del servizio clienti, l'operatore del call center o il social network dove esprimere la sua frustrazione (i quali rappresentano strumenti con i quali il consumatore si relaziona direttamente), hanno un impatto importante anche gli elementi per lui non visibili o indiretti come l'addetto alla fatturazione, il progettista del conto, il reparto marketing, il gruppo di monitoraggio del Social Media, e così via.

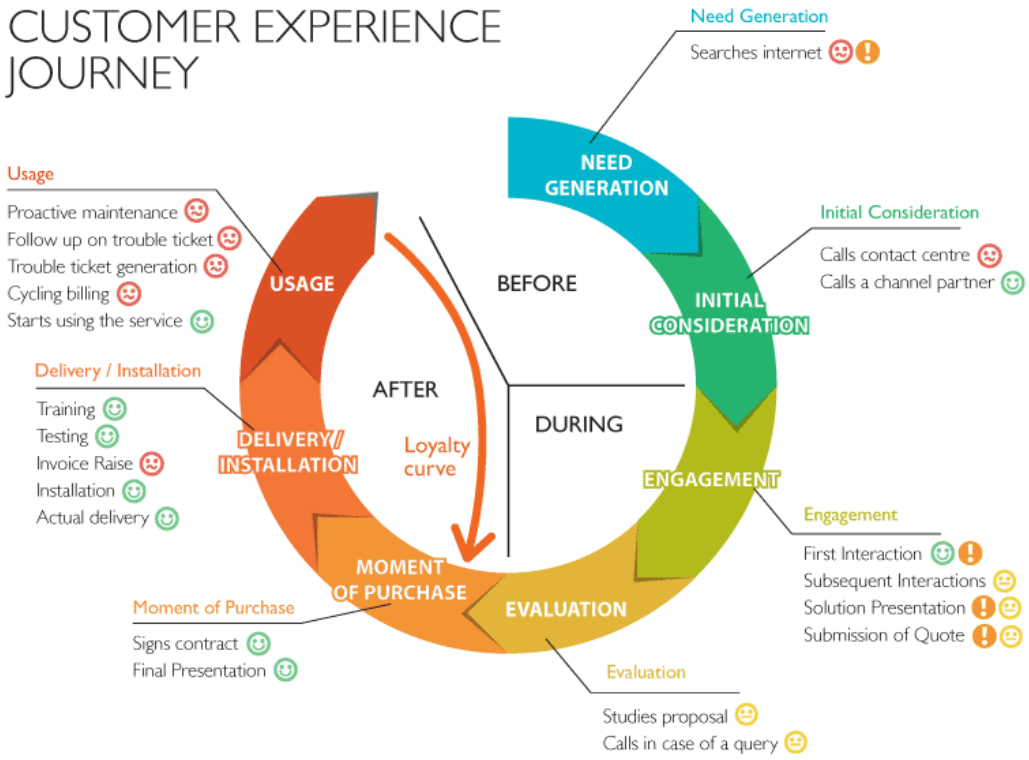

Figura 2: Un modello di Customer Journey

L'illustrazione ci mostra perfettamente quali sono i *Touchpoint* attraverso i quali il cliente riesce a dare la propria valutazione in termini di soddisfazione rispetto a ciascun momento riguardante la sua esperienza d'acquisto.
Per cogliere ciò che accade nelle varie tappe del viaggio del cliente si svolgono indagini quantitative come sondaggi, o qualitative come la raccolta di feedback spontanei dei clienti, raccogliendo i dati monitorandoli ad esempio attraverso i Social Media in maniera tale che l'azienda abbia la possibilità di costruire una visione a 360 gradi della Customer Experience del cliente e di mettere in atto le iniziative necessarie per risolvere le criticità.
Questo significa che il *brand* canalizzerà i propri sforzi nel diversificarsi dai propri concorrenti e nel creare un *engagement* particolare con il proprio target. Le aziende così si aprono, diventano più trasparenti, si "umanizzano" con l'obiettivo di sviluppare un legame di tipo *emotivo* con i clienti.
Dall'altro lato, quando un consumatore giudica positiva la propria C*ustomer Experience*, è più propenso a dare fiducia ai prodotti di quel determinato marchio e sostenerlo presso le proprie cerchie di contatti attraverso, ad esempio, il passaparola, creando quindi ulteriore pubblicità a favore dell'azienda.

2.3. Dal design del prodotto al design dell'esperienza

Un punto fondametale da sviluppare ai fini di questa tesi, è il concetto di User Experience come nuovo paradigma per creare e modellare un'esperienza sempre più unica e personale per ogni cliente. La *User Experience*, in base alla definizione più accettata, è intesa come l'insieme delle "percezioni e risposte della persona risultanti dall'uso e/o dall'aspettativa di utilizzo di un prodotto, sistema o servizio" (ISO 9241-201: 2010). Di conseguenza, la UX mira a valutare la qualità dell'interazione persona-prodotto in un contesto specifico: il momento e il luogo in cui la persona si identifica come utente o utilizza direttamente il prodotto. Essa è determinata da tutte le emozioni, le convinzioni, le preferenze, le percezioni degli utenti, le risposte fisiche e psicologiche e i comportamenti che si verificano prima, durante e dopo l'uso. Di conseguenza essa in generale può venire influenzata da diversi fattori: dall'immagine del brand, dalle funzionalità e dalle prestazioni e dal comportamento del prodotto, dalle capacità del prodotto di comunicare il suo funzionamento, dalle caratteristiche fisiche e cognitive nonché dal background dell'utente (ad es. precedenti esperienze, attitudini, abilità e personalità, ecc.) e dal contesto di utilizzo.
Secondo Hassenzahl (Hassenzahl, 2008), due tipologie di attributi del prodotto determinano maggiormente la UX percepita, e quindi la qualità del prodotto percepita relativa all'uso del prodotto: gli attributi pragmatici e gli attributi edonici. In generale, la qualità pragmatica si riferisce alla capacità percepita del prodotto di supportare il raggiungimento dei "do-goal", quindi è essenzialmente connessa alla percezione dell'utente circa l'utilità e l'usabilità del prodotto in relazione a potenziali attività esterne all'utente. La qualità edonica è legata alla capacità del prodotto di supportare l'utente nel raggiungimento di obiettivi essenzialmente personali o psicologici (es. essere speciale, essere competente, ecc.). Tali attributi contribuiscono a determinare la cosiddetta *Product Image*, ovvero l'opinione che l'utente si crea riguardo ad un prodotto, sistema o servizio.
Come è ben noto, la *Product Image* viene determinata mediante un processo di elaborazione cognitiva dei suoi aspetti funzionali (es. funzionalità, estetica, immagine mentale) e delle sue caratteristiche psicologiche (es. facilità d'uso, sentimenti con il prodotto, emozioni) (Echtner e Ritchie, 1991). Tali caratteristiche tuttavia possono in generale essere percepite in modo differente in base alla natura, più o meno olistica, del processo cognitivo adottato dall'utente e al contesto in cui avviene l'esperienza.
Lo stesso processo che porta alla determinazione delle Product Image, si verifica ogni qualvolta un cliente vive un'esperienza interagendo con un ambiente retail, quali un negozio o un e-commerce, e porta alla determinazione di ciò che possiamo chiamare *Store Image*.
In generale, la determinazione della Product Image e della Store Image non avviene in modo indipendente. Infatti, tutti gli stimoli che trascendono le caratteristiche puramente intrinseche del prodotto (es. prezzo, marca del prodotto, nome del negozio, paese di origine) influenzano fortemente la percezione dell'utente sul prodotto e influenzano il comportamento del consumatore (Swagler, 1979; Davis, 1985; Davis, 1987; Baugh and Davis, 1989).

2.4. Il retail e l'omnicanalità

Gli analisti di McKinsey hanno individuato vari trend collegati ai nuovi modelli del retail 4.0.
Tra questi è utile sottolineare la *Shopping Experience Omnicanale*.
I clienti del futuro saranno i cosiddetti *millenials*, noti per non legare il valore del bene al suo valore intrinseco, ma all'esperienza che gli permette di vivere. La parola d'ordine quindi sarà coinvolgere questo tipo di target e diversificare l'esperienza a seconda del momento, del desiderio e della necessità.
Inoltre, tecnologia e business saranno direttamente proporzionali. Nello specifico, la tecnologia permetterà di incrementare la produttività dei brand e quindi le aziende vincenti saranno quelle che si mostreranno più ricettive all'adozione di tecnologie e processi innovativi.
A questo proposito, Oracle nel suo ebook *"How to build digital, connected and adaptive Customer Experience"*, individua i cinque approcci tecnologici utili per sviluppare e consolidare strategie di CX management efficaci:

- **Iper-personalizzazione**: è la chiave per far sentire il cliente realmente coinvolto nelle strategie del brand. Grazie ai Big Data, agli analytics e agli insight è possibile realizzare prodotti e servizi altamente personalizzati.
- **Chatbot**: l'utilizzo del machine learning e dei bot permette di migliorare l'assistenza pre e post-vendita, fornendo al cliente risposte veloci e adeguate.
- **Integrazione organica dei punti di contatto**: non si parla più di omnicanalità ma di "touchpoint organici". I diversi punti di contatto con il cliente si ricombinano per formarne uno, indistinto, fonte d'interazione tra brand e persona.
- **Internet of Things**: la possibilità di acquisire dati in tempo reale su come i clienti utilizzano prodotti complessi
- **Intelligenza "adattiva"**: machine learning (apprendimento automatico) e intelligenza artificiale (AI), permettono un apprendimento rapido in grado di migliorare sensibilmente la customer experience.

2.5. Il cliente e l'importanza della componente psicologica ed emotiva

La componente psicologica è un elemento fondamentale da tenere in considerazione.
In generale rendere più facile per gli utenti completare le loro attività e raggiungere i loro obiettivi in modo efficiente ed efficace durante le loro fasi d'acquisto permetterà agli utenti di acquistare il prodotto o il servizio ad esso legato con più convinzione. Nel mondo odierno una buona usabilità non è facoltativa, è essenziale: è in questo contesto che entrano in gioco le emozioni legate, appunto, alle interazioni con l'utente.
Le risposte emotive possono essere distinte in due tipologie:

- Negative. Il designer della CX deve concentrarsi sulla riduzione di queste risposte e cercare di eliminare ciò che le crea, non solo perché vanno ad influenzare l'acquisto ma anche e soprattutto perché influiscono sull'immagine e la credibilità del brand

- Positive. Per generare questo tipo di risposta si deve definire, nel modo più accurato possibile, il target a cui rivolgersi. I modi per arrivare a questi riscontri sono vari, per esempio eliminare ciò che c'è di negativo all'interno dell'esperienza di acquisto, progettare la CX in modo efficace, creare esperienze emotive per associazione – progettare in modo che le esperienze vissute in prossimità dei Touchpoint si possano associare a qualsiasi cosa generi emozioni positive.

Nel suo libro, The Design of Everyday Things, Don Norman analizza tre livelli di risposta che l'utente ha nell'interazione con un sistema; applicando questo sistema si può tentare di avere una comprensione più approfondita delle percezioni ed emozioni degli utenti.
Il primo livello è quello che lui definisce **viscerale*:*** esso è responsabile dell'aspetto inconscio ed automatico dell'emozione umana e riguarda principalmente l'aspetto estetico del prodotto e le prime impressioni su di esso. Lo scopo che si vuole ottenere è quello di attirare l'attenzione del cliente, in questo caso gli aspetti superficiali di un prodotto possono aiutarlo a distinguersi dalla concorrenza. Se si parla di sviluppare un'interfaccia bisogna dare una chiara idea di ciò che si offre in modo immediato. In questo caso il branding gioca un ruolo fondamentale poiché costituisce l'insieme di valori e convinzioni che rendono un prodotto o più in generale un'azienda - diversa dalla concorrenza, incoraggiando i clienti a connettersi emotivamente con esso. In conclusione si può dire che le impressioni che un utente ha nel primo impatto sono inconsce e, poiché la bellezza è soggettiva bisogna ricercare in modo molto approfondito cosa significhi per il target che si vuole colpire.
Il secondo livello è quello **comportamentale**: esso si riferisce a ciò che si è abituati a conoscere come usabilità intesa nel modo più generico del termine. Lo studio della User Experience si basa principalmente su questo livello perché se un utente non è in grado di usare qualcosa in modo semplice ed efficace, nient'altro conta. L'impatto che un utente ha con un prodotto (che sia esso "fisico" o "digitale") può essere studiato attraverso un test con l'utente stesso; in questo caso si studia la sensazione di soddisfazione nell'aver raggiunto o meno un determinato risultato. La cattiva usabilità influenza il modo in cui ci si sente nei confronti di un prodotto e può far trascurare il suo fascino estetico (viscerale) e di conseguenza può influenzare negativamente la decisione del cliente.
Il terzo ed ultimo livello è quello **riflessivo**, che in termini di design emotivo è il livello più alto e tiene conto dei pensieri coscienti di un utente e del suo potere decisionale; è l'unico livello che coinvolge una forma consapevole di elaborazione, ma nonostante tutto viene fortemente influenzato dagli altri due. Si può dunque dire che migliore è la risposta emotiva dell'utente, maggiore è la possibilità che l'utente crei una connessione con il brand e di conseguenza prenda una decisione positiva nei confronti del prodotto che è intenzionato ad acquistare.
Definendo dunque il target dall'inizio si potranno condurre ricerche adeguate che permetteranno di attingere a dati che porteranno ad uno studio per ottenere una buona usabilità del sito e conseguentemente una buona risposta emotiva degli utenti.

Ciò che è possibile monitorare in maniera automatizzata con una piattaforma tecnologica, ossia con il sistema oggetto di questa tesi, possono essere tutte e tre gli aspetti analizzati da Norman, in particolar modo quello viscerale e quindi di conseguenza **affettivo**.
È bene sottolineare però che non tutto è così facilmente schematizzabile secondo questo modello, nel senso che le risposte dell'individuo di fronte all'oggetto sono molto complesse ed i tre livelli tendono a sovrapporsi ed influenzarsi reciprocamente: in questo senso il collegamento che verrà analizzato successivamente tra i livelli di risposta di Norman e come tali risposte possono venir monitorate dalla piattaforma tecnologica, potrebbe non risultare sufficiente ad un esperto psicologo o di User Experience: tuttavia tra gli scopi di questo sistema c'è innanzitutto la volontà di fornire una risposta automatizzata al raccoglimento di dati per fornire, solo a posteriori e qualora necessario, un'analisi a livello psicologico/cognitivo dei livelli di interazione tra utente e prodotto/servizio/brand. Di fatto, queste tematiche investono campi, come quello appunto della psicologia cognitiva, al di fuori dagli obiettivi specifici affrontati in questa tesi.
Oltre a queste tipologie di dati, ovviamente assumono grande importanza le caratteristiche intrinseche degli utenti/clienti: in particolar modo il loro sesso ed età.
Nell'ambito del marketing e analisi di mercato, è necessario considerare uomini e donne non come un "unico gruppo omogeneo" ma come "due gruppi distinti". Questo perché le strategie di marketing devono andare incontro ai bisogni, agli interessi e alle preferenze delle persone e per fare ciò è indispensabile diversificare i vari target per poter effettuare scelte ad hoc e molto mirate alla persona. Tale strategia si definisce anche come "Gender Marketing" e si basa sul fatto che le differenze tra sessi sono fondamentali in quanto le donne hanno comportamenti di "acquisto" più strategici mentre gli uomini sono propensi all'acquisto impulsivo; le donne tendono a considerare le opinioni altrui per prendere una decisione mentre gli uomini considerano la decisione altrui come una guida per "formare" le proprie opinioni. Anche la differenziazione rispetto alle fasce di età è molto importante nel Marketing strategico: fino ad ora si era considerato un unico target denominato "Millenials" che comprendeva i nati tra la fine degli anni '70 e il 2000 ma ci si è resi conto, soprattutto dopo la crisi del 2001, che era necessario suddividere tale categoria in sotto categorie in quanto gli interessi e le esperienze di vita erano troppo differenti. Importante notare che la generazione successiva a quella dei "Millenials" definita come "Generazione I" (generazione delle reti) risulta quella dei ragazzi, profondi conoscitori ed utilizzatori dei sistemi informatici, ai quali è necessario rivolgersi con strategie di mercato e mezzi di comunicazione del tutto differenti da quelli delle altre categorie.

2.5.1 L'acquisto e le emozioni

Per lungo tempo il marketing ha considerato la sfera delle emozioni troppo sfuggente per essere studiata utilmente al fine di comprendere meglio i comportamenti dei consumatori. Sono state ritenute troppo imparentate con concetti metafisici per poter essere ricondotte al modello naturalistico delle leggi universali e al rigore dei metodi e delle procedure di controllo della scienza e della ragione.

Le aziende sono guidate dai numeri, dalle statistiche, dai risultati misurabili di determinate azioni e non dagli umori e dalla creatività. O meglio, la creatività arriva quasi sempre dopo l'analisi accurata delle possibilità.
In tale prospettiva la componente emozionale delle decisioni è stata considerata come un fattore di fondo, da valutare se proprio necessario con l'esperienza e con strumenti di ricerca di tipo convenzionale, quali le indagini quantitative su grandi numeri di rispondenti oppure le più sofisticate ricerche qualitative, più contenute nelle dimensioni ma anche più ricche di indicazioni.
Guardando, ad esempio, una bella confezione di profumo di alta gamma, il piacere inizia già durante la visione della forma del pack e prosegue man mano che si entra nei dettagli del prodotto in esso contenuto fino alla prova.
I colori, il profumo e il coinvolgimento esperienziale sono quanto serve al prodotto per fare la differenza e fare breccia nella sfera emozionale del cliente.
Le ricerche condotte appena dopo la prova del profumo in questione consentono di cogliere "a caldo" i residui delle emozioni prodotte dal rapporto con il prodotto.
Quando poi le risposte fornite da più persone rispetto alla stessa esperienza d'uso dello stesso profumo convergono verso risposte simili si può ragionevolmente parlare di misurazioni attendibili e metodologie affidabili.
Il dualismo tra ragione ed emozione, che sembra generare due diversi modelli d'interpretazione della realtà, è stato oggetto di molti studi di marketing, basti pensare al modello PAD (Mehrabian, 1996) che individua tre diverse componenti emozionali: pleasure (la piacevolezza dell'atmosfera di un punto vendita), arousal (il grado di coinvolgimento vissuto dal visitatore), dominance (il controllo del visitatore sull'ambiente), oppure gli studi condotti dal professor Paul Ekman, docente di psicologia all'Università della California, sull'espressione delle emozioni e la corrispondente attività fisiologica sul viso. Attualmente, grazie al neuromarketing, questo dualismo sembra essere superato e non vi è opposizione fra questi due livelli dell'esperienza ma di una interazione piena tra componente emotiva e quella razionale.
Dall'elaborato di Francesco Gallucci (economista e saggista, tra i maggiori esperti di neuromarketing e marketing emozionale) Marketing emozionale, si evince chiaramente la natura non semplice del termine in quanto la parola è stata per tempo studiata nel significato, trovando però da sempre molte difficoltà nel definirla: egli spiega che "è una parola che è entrata solo da pochi decenni nell'uso corrente del nostro linguaggio e deve, probabilmente, trovare ancora una certa stabilità nel proprio significato".
In effetti effettuando ricerche sul termine, si ottengono risultati alquanto vari nel significato, ma che ho perlomeno tentato di riassumere in questi termini: un'emozione è uno stato mentale e psicologico associato ad un'ampia varietà di sentimenti, pensieri e comportamenti interni (fisici) o esterni (sociali); è definita come una reazione affettiva intensa determinata da uno stimolo ambientale o interno, che può essere naturale o appreso; un processo cognitivo nel quale la percezione di un insieme di stimoli consente una valutazione cognitiva che permette alle persone di etichettare determinati sentimenti.
In ogni caso le emozioni si attivano come reazione a specifici stimoli o situazioni e si manifestano come sistemi coordinati che comprendono:

- La valutazione degli stimoli
- I mutamenti fisiologici

- Le risposte comportamentali (per esempio attacco o fuga)
- Le risonanze affettive
- Le risonanze cognitive

Un altro aspetto riguarda la natura fisiologica delle emozioni, molto importante per lo sviluppo di applicazioni di neuromarketing basate sulla misurazione delle variazioni degli stati del corpo e del cervello in relazione ad uno stimolo, che può produrre alterazioni del battito cardiaco, della pressione sanguigna e del ritmo della respirazione. Quindi, mutando punto di vista, possiamo definire l'emozione come la catena di eventi che si innescano tra uno stimolo scatenante (input) e l'esecuzione del comportamento elaborato come risposta (output).

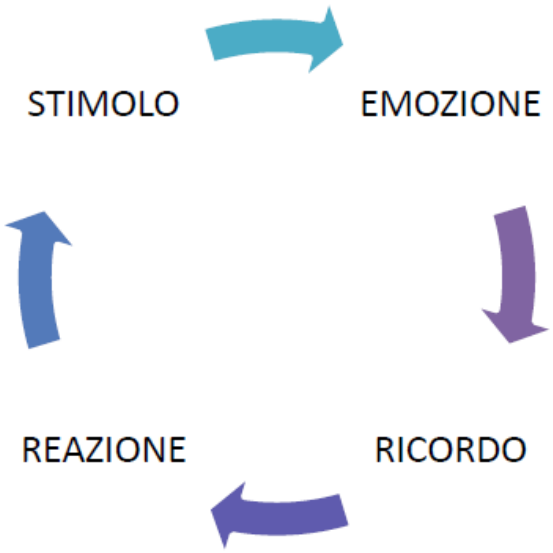

Figura 3:Il ciclo retroattivo di stimoli-reazioni basato sulle emozioni

Gli stimoli emozionali possono essere un evento, una scena, un'espressione facciale o una campagna pubblicitaria, e comportano nel nostro organismo una reazione fisiologica quale ad esempio l'accelerazione del battito cardiaco, sudorazione improvvisa, arrossimento, cambiamento del ritmo respiratorio e la tensione muscolare (Jarrold, 2004).
Da secoli l'uomo con la sua psicologia si è occupato della classificazione delle emozioni e molti sono gli studi che, in accordo con la scienza moderna, hanno reso possibile una classificazione utilizzando quattro diversi principi: principi logici, principi psicologici, principi sociologici e principi biologici.
Aristotele, uno dei primi studiosi che tentò di inserire le emozioni all'interno di un quadro sistematico, ricorrendo dunque ad un criterio logico, le considerava "passività", in quanto passioni e non azioni e le metteva in relazione con il sistema cognitivo in quanto "potenzialmente modificabili dalla persuasione". Tale definizione logica è giunta fino ai

giorni nostri, passando per diverse filosofie di pensiero fino ad arrivare ad essere letta come un tentativo di ristabilire un rapporto con il mondo dopo un'improvvisa destrutturazione, dopo un avvenimento che in qualche maniera abbia turbato il "nostro stato emotivo standard", sia in modo positivo che negativo.
Con il criterio psicologico, le emozioni sono la conseguenza dell'adattamento o disadattamento alla situazione.
In primo luogo è possibile individuare una serie di emozioni e catalogarle in base ai processi adattativi del comportamento, elencandone otto e definendole primarie: gioia, tristezza (o dispiacere), accettazione (approvazione), rabbia, paura, disgusto, attesa (o aspettativa) e sorpresa.
Successivamente possiamo definire le emozioni complesse combinando le primarie o articolando attributi fondamentali di queste (per esempio l'amore).
Questa tendenza a combinare le emozioni base tra loro è dunque in sintesi l'approccio attuale del criterio psicologico di classificazione delle emozioni.

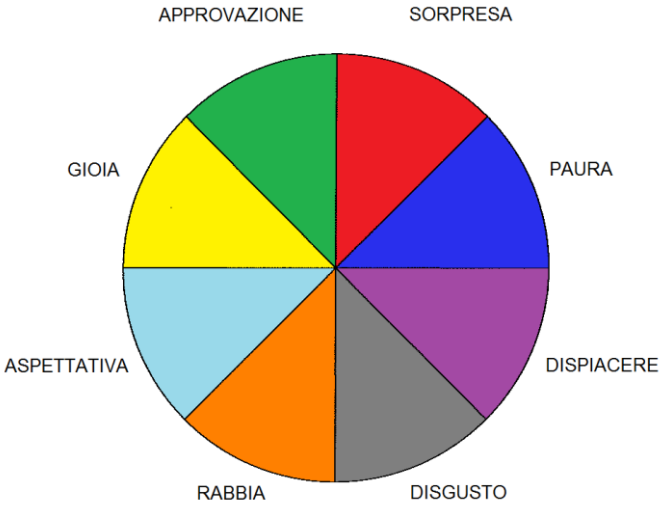

Figura 4:Lo spettro delle emozioni primarie

Utilizzando i criteri sociologici, le emozioni sono considerate in relazione alle ideologie, alle convinzioni dei gruppi di appartenenza e alla loro convalida, alla cultura, stato sociale, all'organizzazione gerarchica dell'individuo. Avvalendosi di tale punto di vista si individuano emozioni di tipo altruistico come quelle familiari, sessuali o prettamente sociali; emozioni di tipo egoistico, legate all'affermazione e alla difesa di sé; in ultima istanza le emozioni definite superiori, ovvero quelle che trascendono la sfera dell'Io e si rivolgono espressamente al più ampio contesto sociale-umano.
L'ultimo ambito di classificazione che andiamo ad analizzare è quello *biologico*.

Questo è l'approccio che sicuramente offre un grado di oggettività superiore a tutti gli ambiti fino ad ora considerati. Tuttavia, in egual modo, preclude la possibilità di una descrizione del "senso" dell'emozione: ad esempio, una stessa modificazione fisiologica può sottostare a diversi stati emotivi, differenziandosi unicamente per l'intensità con cui si manifesta. Nel caso della *collera* e della *gioia*, queste emozioni di natura decisamente diversa, si manifestano con le stesse reazioni fisiologiche (accelerazione del ritmo cardiaco e respiratorio, tonicità muscolare aumentata, variazione della pressione arteriosa, ecc..) che per una e per l'altra emozione avranno un diverso grado di intensità.
In ogni caso, più avanti nel corso della trattazione, vedremo come in qualche caso nell'ambito della ricerca si sono potuti fare passi avanti nella correlazione, quantomeno empirica ed abbastanza vicino alla realtà, tra stati psicofisici ed emozioni.
Robert Plutchik, psicologo che elaborò una delle prime classificazioni organiche delle emozioni, fu uno dei precursori sostenitori della teoria delle *otto emozioni primarie*, innate ed universali, la cui combinazione può dare origine ad emozioni secondarie o complesse.
Tale classificazione è quella che in effetto ha avuto più successo, anche perché non si basa solo sull'elenco delle emozioni fondamentali ma su un vero e proprio modello che rappresenta bene le osservazioni della realtà e quindi ha resistito a molte verifiche sul piano empirico.
Le emozioni possono essere positive o negative, costruttive o distruttive, gradevoli o sgradevoli a seconda di ciò che le causa e delle situazioni in cui la persona si trova.
Oggi si tende a parlare di emozioni solo in senso positivo. Quella che dovrebbe essere una componente normale della vita affettiva di chiunque viene rappresentata sempre più spesso come un avvenimento eccezionale, possibile solo nei momenti di evasione, oppure associata ad un'esperienza memorabile. In ogni caso in congiunzione di situazioni speciali.

2.6. Il riconoscimento automatico delle emozioni: le ricerche di Paul Ekman

Tra i vari aspetti della vita umana, le emozioni giocano indubbiamente un ruolo centrale: esse danno un significato particolare alle esperienze intra ed extra personali. Negli ultimi due decenni sono stati condotti molti studi sia sulle origini neurologiche che sulle funzioni sociali delle emozioni.
La comprensione delle emozioni si traduce spesso nella possibilità di migliorare le interazioni uomo-macchina (HCI): esistono diverse tecnologie per riconoscere automaticamente le emozioni umane, che comprendono l'analisi dell'espressione facciale, l'elaborazione acustica del parlato e l'interpretazione delle risposte biologiche; il Facial Action Coding System (FACS) originariamente sviluppato dall'anatomista svedese Carl-Herman Hjortsjo, poi adottato da Paul Ekman, W.V Friesen e Joseph C.Hager nel 1970, è una di queste. Si tratta di una guida tecnica dettagliata che spiega come categorizzare le espressioni facciali in base ai movimenti muscolari che le producono. La contrazione di ogni muscolo facciale cambia l'aspetto del viso, questo è uno standard comune per categorizzare sistematicamente l'espressione fisica delle emozioni.
Quando si sperimenta un'emozione, la rilevazione dell'emozione può essere ottenuta rilevando le espressioni facciali ad essa correlate, da ognuna di esse si estrae un insieme di

“Action Units” facciali, che identificano, univocamente, un movimento indipendente di un tratto del volto. I sistemi di riconoscimento delle espressioni facciali basati sul riconoscimento delle emozioni implementano questo approccio catturando immagini delle espressioni facciali dell'utente e dei movimenti della testa. Questi sistemi rilevano dunque i movimenti muscolari percepiti come cambiamenti nella posizione degli occhi, del naso e della bocca, tramite movimenti di punti in un sistema di coordinate. Quindi, analizzando tali cambiamenti, è possibile determinare l'occorrenza di un’Action Unit (Affect Measurement: A Roadmap Through Approaches, Technologies,and Data Analysis, 2017).

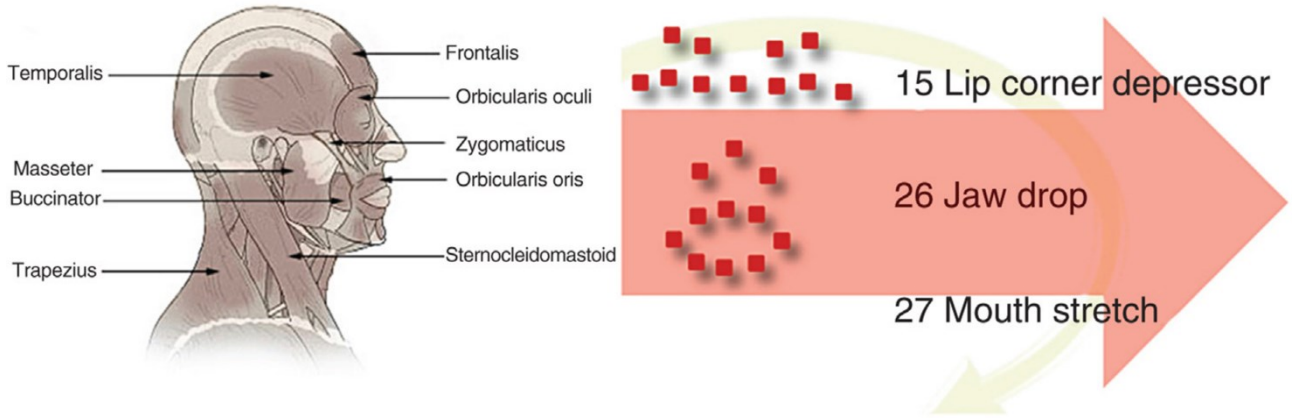

Figura 5: La definizione delle Action Units dai muscoli facciali

Quindi riepilogando, i movimenti dei singoli muscoli facciali sono codificati come Action Units, ossia quelle che vengono considerati come le unità “atomiche”, ossia non ulteriormente scomponibili per rappresentare un movimento facciale. Ad esempio, la depressione angolare delle labbra è codificata come Action Unit 15, l’abbassamento della mascella come Action Unit 26 e l'allungamento della bocca come Action Unit 27.
Il lavoro dello psicologo Paul Ekman viene considerato tutt’oggi come il nucleo del riconoscimento delle emozioni (Emotion Recognition) basato sul riconoscimento delle espressioni facciali: ormai più di 40 anni fa egli teorizzò la presenza di un insieme discreto di emozioni fisiologicamente distinte: rabbia, disgusto, paura, felicità, tristezza e sorpresa e le loro corrispondenti espressioni facciali *universali* corrispondenti, universali (da qui l’attribuzione del nome “Emozioni Universali di Ekman”), poiché esse sono espresse allo stesso modo in tutto il mondo, al di là dell’etnia o dalla provenienza geografica di un essere umano. Rilevando queste espressioni facciali si può capire lo stato emotivo di una persona senza sapere nulla del proprio background, questo ha un enorme potenziale in un grande numero di settori industriali e commerciali, compreso ovviamente quello del retail.

2.6.1 Valence ed Engagement

La valenza (valence) dell'emozione differenzia semplicemente tra sentimenti piacevoli (positivi) e sgradevoli (negativi). Avere una visione della dimensione della valenza è spesso utile, in quanto possono comparire stati affettivi non basici, senza dimenticare che anche le emozioni di base spesso si fondono insieme. Questo è il motivo per cui la valenza è una dimensione comunemente indagata, specialmente nelle aree di ricerca psicologica.
Chiaramente, una partizione discreta delle emozioni è un'utile semplificazione con cui lavorare; tuttavia, la percezione reale dei sentimenti è molto più complessa e continua, e questo si riflette in modo particolare sulla difficoltà che molte persone hanno nel descrivere e valutare le proprie emozioni. Con questo in mente, le dimensioni di valenza sono utilizzate da uno dei cosiddetti "modelli circonplex" per le emozioni, in particolare con quello in (Russel, 1980), in cui combinazioni lineari di valenza e livelli di eccitazione (engagement) forniscono uno spazio affettivo bidimensionale.

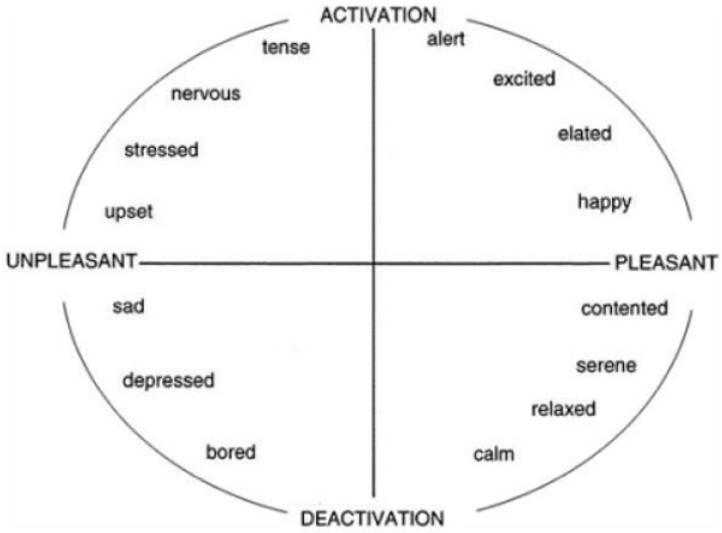

Figura 6: Il modello circonplex di Russel

La valenza è riportata nell'asse orizzontale (da totalmente sgradevole a pienamente piacevole), mentre l'eccitazione è riportata in quello verticale.
Il coinvolgimento (engagement) dell'utente racconta invece quanto l'utente sia affascinato da un'esperienza, e si rivela così essere un altro strumento prezioso nei contesti di marketing. L'engagement non può essere definito in modo univoco, per cui si possono applicare diverse metodologie per stimare quanto l'utente è attratto da un prodotto. Per i nostri scopi, consideriamo l'engagement per lo più legato al contatto visivo dell'utente con ciò che gli sta

di fronte. In base a ciò, l'engagement può essere misurato attraverso l'elaborazione di frame video e la combinazione di informazioni sullo sguardo e sullo stato emotivo.

2.7. Tecnologie per il riconoscimento automatico dalle immagini: il Deep Learning

Come già accennato, molti studi, propongono metodi per l'analisi della risposta del cliente e la misurazione della qualità CX lungo il percorso del cliente, che si basano sulla costruzione della curva emozionale.
È facile immaginare quanto possa essere rivoluzionario in questo settore un sistema in grado di rilevare e monitorare in tempo reale e in modo totalmente automatico le caratteristiche statiche e dinamiche del cliente, come l'età, il sesso e l'etnia, nonché le emozioni e i comportamenti.
Al giorno d'oggi, la crescente disponibilità di sensori e dispositivi intelligenti connessi a Internet, e alimentati dalla tecnologia pervasiva dei Sistemi Cyber-Fisici e dell'Internet delle Cose, creano una crescita esponenziale dei dati disponibili. Ad esempio, i rivenditori possono raccogliere informazioni demografiche sui clienti, come l'età, il sesso e l'etnia. Allo stesso modo, possono contare il numero di clienti, misurare il tempo che trascorrono nel negozio, rilevare i loro modelli di movimento, misurare il tempo che passano in aree diverse e monitorare le code in tempo reale. Informazioni preziose possono essere ottenute correlando queste informazioni con i dati demografici dei clienti per guidare le decisioni relative a posizionamento del prodotto, prezzo, ottimizzazione dell'assortimento, design promozionale, cross-selling, ottimizzazione del layout e del personale. Tuttavia, in base alle attuali conoscenze, nessuno studio ha sfruttato gli strumenti di riconoscimento delle emozioni per monitorare l'esperienza del cliente durante il suo Journey. Oggi diversi metodi e tecnologie consentono il riconoscimento delle emozioni umane, che si differenziano per il livello di invadenza. Ovviamente, l'uso di strumenti invasivi (ad es. ECG o EEG, sensori biometrici) può influenzare il comportamento dei soggetti e in particolare può adulterare la loro spontaneità e, di conseguenza, le emozioni vissute da essi, introducendo bias che finiscono inevitabilmente per compromettere i risultati attesi. La maggior parte di tali tecniche, metodi e strumenti fa riferimento a tre aree di ricerca: analisi delle emozioni facciali, analisi del riconoscimento vocale e analisi delle emozioni biofeedback.
L'analisi delle emozioni facciali mira a riconoscere i modelli dalle espressioni facciali e collegarli alle emozioni. Spesso utilizza algoritmi di Deep Learning, in particolare basati su Convolutional Neural Networks (CNN), un modello matematico di Deep Learning che prende in input diversi tipi di immagini e fa previsioni sulla base del modello addestrato. Questo tipo di reti neurali, rispetto a quelle classiche, aggiunge diversi livelli, nella prima parte dell'intera rete, che applicano un'operazione di convoluzione alle immagini in ingresso e passano il risultato ai livelli successivi. Il più delle volte, gli ultimi strati sono composti da normali reti neurali.

2.8. Lo stato dell’arte per il riconoscimento automatico di emozioni, sesso, età e sguardo

2.8.1 Il riconoscimento di sesso ed età

Età e genere sono altre due caratteristiche importanti tra quelli estraibili dai tratti del volto oltre all'emozione, questi attributi demografici potrebbero anche essere deterministici quando si cerca di comprendere le diverse preferenze di genere e fasce di età. Le attuali metodologie allo stato dell'arte sono in grado di prevedere il genere con una precisione di circa 87% e di stimare l'età con un tasso di errore di circa ± 5 anni (Levi et al., 2015)

2.8.2 Il riconoscimento dello sguardo

Essere in grado di tracciare il movimento degli occhi è una sfida molto importante nel campo dell'HCI e della Computer Vision. Quando si tratta di progettazione di prodotti (fisici o digitali, come anche una piattaforma web), è facilmente comprensibile cosa significhi sapere dove sta guardando un cliente/utente.
Approcci diversi possono essere utilizzati per cogliere le coordinate del punto su cui l'utente si sta concentrando. La maggior parte di essi utilizza tecnologie invasive, come gli occhiali prodotti da Tobii e di solito richiedono una fase di calibrazione iniziale. Al giorno d'oggi, è difficile trovare sistemi commerciali che sfruttano tecnologie non invasive per scopi di tracciamento oculare: ad esempio, iMotion, leader nel settore delle soluzioni software integrate per supportare la ricerca nel comportamento umano, ha una partnership solo con Tobii.
A tal proposito, la ricerca degli ultimi anni mira a raggiungere risultati abbastanza accurati con l'impiego di tecnologie non invasive e standard (soprattutto webcam).
Le tecniche di Gaze Tracking possono essere suddivise globalmente in due categorie principali: model-based e appearance-based (Hansen et al., 2009).
I primi approcci si basano fortemente su immagini di alta qualità per determinare con precisione le caratteristiche dell'occhio, che a loro volta richiedono spesso attrezzature di laboratorio (ad esempio, sorgenti di luce IR, occhiali speciali / lenti a contatto, macchine fotografiche multiple, ecc.) Al contrario, gli approcci appearance-based prendono i contenuti dell'immagine direttamente in input e cercano implicitamente di estrarre alcune caratteristiche rilevanti, stabilendo così una mappatura delle coordinate dello schermo. L'applicabilità di quest'ultimo tipo di metodo è notevolmente ampia, anche se la potenziale gestione delle immagini a bassa risoluzione rende la loro accuratezza generalmente inferiore (Zhang, 2015).
Recentemente, il Deep Learning e le Reti Neurali Convoluzionali (CNN) hanno guadagnato interesse per la stima dello sguardo, tanto che negli ultimi anni sono emerse proposte per molti dataset e architetture di rete.
Nel 2016, Krafka et al. ha pubblicato un articolo con una solida proposta di CNN per il rilevamento dello sguardo, insieme ad un grande dataset di volti raccolti in crowdsourcing; molti ricercatori vi si riferiscono da allora.

2.8.3 Il riconoscimento delle emozioni dalle espressioni facciali

Oggi, il riconoscimento delle emozioni è uno dei temi più impegnativi per numerosi settori applicativi: dall'automotive dove viene utilizzato per la guida autonoma, all'industria per supportare i processi decisionali, fino alla robotica per la realizzazione di interazioni empatiche. Attualmente, il modo migliore per riconoscere le emozioni umane è quello di elaborare i video e le immagini catturate da una videocamera senza introdurre distorsioni dovute all'uso di tecnologie intrusive come alcune tecnologie indossabili (ad esempio caschi, bracciali) o sensori distribuiti. Esistono diverse ricerche per raggiungere questo obiettivo in modo automatizzato e la maggior parte utilizza Reti Neurali Convoluzionali.
Numerosi studi sono stati condotti sull'analisi del riconoscimento dell'espressione facciale per la sua importanza pratica nella robotica sociale, nella medicina, nella sorveglianza della stanchezza dei conducenti di autoveicoli e in molti altri sistemi di interazione uomo-macchina. Dal 2013, competition di Emotion Recognition come FER2013 (Goodfellow et al., 2013) ed Emotion Recognition in the Wild (EmotiW) (Dhall et al., 2015, 2016, 2017) hanno raccolto dati di addestramento (training data) di qualità relativamente buona da scenari reali impegnativi, che implicitamente promuovono la transizione di FER da ambienti controllati in laboratorio a quelli "in the wild", ossia in ambienti di tutti i giorni in cui le condizioni ambientali come luce, distanze di acquisizione dall'obiettivo ecc. non sono controllabili.
Grazie all'incredibile incremento delle capacità di elaborazione dei chip anno dopo anno (ad esempio, nel caso delle GPU) e di architetture di rete sempre meglio progettate, studi in vari campi hanno iniziato ad adottare metodi di Deep Learming, che hanno raggiunto l'accuratezza di riconoscimento allo stato dell'arte e superato ampiamente i risultati precedenti (Krizhevsky et al., 2012, Simonyan et al., 2014, Szegedy et al., 2015, He et al., 2016). Allo stesso modo, grazie a dati di training sempre più efficaci nel riconoscimento delle espressioni facciali, sono state implementate sempre più tecniche di Deep Learning in grado di gestire le complicate situazioni in cui il riconoscimento delle emozioni avviene "in the wild" (Li et al. 2018).
Tra gli strumenti principali per l'analisi emotiva visiva, c'è Affdex di Affectiva, che consente di fornire la tendenza emotiva di un soggetto, attraverso il riconoscimento delle emozioni primarie di Ekman & Keltner (Ekman et al., 1970), e i Microsoft Cognitive Services basati sul Piattaforma di Azure. Entrambi, oltre a quel tipo di analisi, sono anche in grado di riconoscere età, genere ed etnia e sono basati su reti neurali convoluzionali e/o ricorrenti.

2.8.4 L'analisi dei biofeedback

Per avere la visione più completa possibile sullo stato emotivo di un soggetto, possono venir raccolte informazioni aggiuntive tramite i biosensori. Mentre è evidente che in determinati contesti è possibile nascondere le reali emozioni provate assumendo espressioni facciali "forzate", la reale condizione affettiva si riflette necessariamente su quegli aspetti biologici controllati dal Sistema Nervoso Autonomico (o Autonomic Nervous System, ANS), che è involontario (Hamdi et al. 2012).

I segnali di biofeedback includono la variabilità della frequenza cardiaca (HRV), attività elettrodermica (EDA), fotopletimogrammi (PPG), temperatura della pelle (ST) e tracce elettroencefalografiche (EEG).
Set di stimoli costituiti da immagini che suscitano emozioni specifiche sono cresciuti attraverso test di ricerca, dando vita all'International Affective Picture System (IAPS) (Lang et al., 2007). Esperimenti basati su IAPS hanno dimostrato che da segnali biologici si possono estrarre una serie di caratteristiche che permettono di discriminare tra insiemi discreti di emozioni (Bradley et al., 2001).
Gli ultimi anni di ricerca hanno portato alla frammentazione dei sensori (ad esempio, attraverso l'introduzione di sistemi indossabili e controlli basati su smartphone) raggiungendo buone precisioni (dall'80% al 90%) nel rilevare gli stati emotivi in generale (Sherratt et al., 2018). Tuttavia, l'analisi emotiva sui segnali biologici è ancora agli inizi, sia per questioni tecnologiche che per la mancanza di una comprensione completa dei fenomeni fisiologici; ad oggi, la classificazione delle sei emozioni di Ekman attraverso l'analisi biofeedback raggiunge appena il 50% di precisione (Maria et al., 2019). Oltre a questo, come già accennato l'utilizzo di sensori indossabili introduce inevitabilmente dei bias che finirebbero per far cadere uno dei punti forti del sistema in esame in questa tesi: la non intrusività.

Capitolo 3.

L’approccio proposto: una piattaforma omnicanale per l’analisi automatizzata della Customer Experience

Per rispondere ai problemi affrontati relativi al settore del retail e al monitoraggio dell’esperienza del cliente (CX) in prossimità dei touchpoint, viene proposta una piattaforma tecnologica in grado di fornire delle soluzioni in grado di automatizzare la raccolta dati relativi alle caratteristiche e ai comportamenti dei clienti/utenti.
Come verrà spiegato più nel dettaglio successivamente, il sistema proposto si basa innanzitutto sul concetto di omnicanalità, vuole dunque cercare di fornire una soluzione sia ad ambienti in cui il cliente interagisce con prodotti fisici (negozi, centri commerciali ecc.), sia in ambienti in cui il cliente/utente interagisce con prodotti digitali (applicazioni mobile e piattaforme web, soprattutto e-commerce). Oltre a questa caratteristica, il sistema in esame deve possedere due fondamentali caratteristiche riguardo la tecnologia adottata per la raccolta dati:

- Non invasività
- Pervasività

Questo sistema adotta due diverse strategie per catturare emozioni e caratteristiche del cliente e rendere l'esperienza adattiva e reattiva (Fig. 7). Questa piattaforma agisce direttamente sull’ambiente di acquisto, circondando i clienti per migliorare la loro esperienza in maniera reattiva, e fornisce un sistema di supporto alle decisioni (DSS) in grado di coadiuvare la progettazione della CX e la pianificazione della stessa a breve e a lungo termine. Tale piattaforma consente l'acquisizione di vari tipi di informazioni relative ai comportamenti dei clienti, tra le espressioni facciali o la voce, in maniera non intrusiva. Il sistema sfrutta algoritmi di intelligenza artificiale (Machine Learning) basati sull'inferenza induttiva e in grado di fornire “decisioni” sulla base di regole logiche derivate da un quadro di conoscenze composto da tre moduli principali: la Service Ontology (SO) e la Product Ontology (PO), che insieme determinano l'ontologia dell'offerta, e la Customer Ontology (CO). Questi moduli rappresentano dunque l’implementazione di un’ontologia, ossia la rappresentazione formale di domini di interesse (in questo specifico caso di Servizi, Prodotti e Clienti) al fine di dare una interpretazione semantica ai dati che sono stati raccolti e associati proprio ai domini sopra evidenziati. I SO e PO permettono dunque al retailer di mappare, strutturare e gestire le proprie conoscenze relative al prodotto/servizio con un modello semantico che farà da base per aiutare a progettare il servizio. Per definire le ontologie SO e PO e le relative interrelazioni, sono state prese a riferimento le “Good Relations” per ontologie web proposte da Hepp nel 2008. Tuttavia, per considerare le peculiari caratteristiche ambientali di prodotti e servizi forniti dai retailer nei negozi, e le caratteristiche dei diversi soggetti interessati, come

il personale, sono state introdotte altre due classi: Ambiente e Stakeholder. La classe Ambiente rappresenta il contesto in cui si svolge l'esperienza del cliente. La tassonomia dell'Ambiente mira ad includere tutti i concetti necessari per definire le caratteristiche degli elementi che caratterizzano l'ambiente di shopping e possono avere un ruolo nella determinazione della CX, compresi arredi per interni, espositori di merci, vetrine, scaffali, illuminazione, musica, aromi, grafica e contenuti delle piattaforme di e-shop, ecc. La classe degli Stakeholder è composta da tre principali tassonomie: Tassonomia Demografica, Tassonomia Psicografica e Tassonomia Tecnografica. Le tassonomie Demografica e Psicografica includono i concetti relative alle informazioni personali e alla personalità degli stakeholder considerati, come ad esempio il personale impiegato nel negozio o gli assistenti negli acquisti virtuali, mentre la tassonomia Tecnografica è composta da tutti i concetti relativi alla loro competenza professionale. La CO fornisce una struttura per tutti i dati semantici relativi alle caratteristiche e agli obiettivi (o bisogni) degli utenti/clienti: essa si basa su due categorie principali di informazioni: gli obiettivi del cliente e le caratteristiche dei clienti. Le caratteristiche dei clienti comprendono quattro domini informativi: demografico, psicografico, di salute e comportamentale. Il demografico include tutti i concetti che permettono la definizione delle caratteristiche personali e socioeconomiche dei clienti, compresa l'età, il sesso, l'istruzione, lo stato civile, le dimensioni della famiglia, occupazione, ecc. Il dominio psicografico comprende tutti i concetti relativi alle diverse caratteristiche dei clienti relative a tipologia e tratti della personalità, atteggiamenti personali e stile di vita. Gli attributi relativi alla salute mirano a definire lo spettro delle capacità e delle condizioni di salute dell'individuo. Questa parte della CO è definita a partire dagli otto domini di funzionalità del corpo codificate dall'International Classification of Function, Disability and Health (ICF) (WHO, 2001). Il comportamento del cliente include tutti i concetti e le relazioni necessarie per identificare il "modo di agire" del cliente stesso (statistiche sugli acquisti, preferenze, ecc.) ed il suo stato emotivo. Infine, gli obiettivi del cliente identificano le possibili ragioni per cui il cliente vorrebbe comprare un prodotto. Questa parte della CO viene definita a partire dai nove domini di attività codificate dall'ICF. Tali domini includono le nove categorie di bisogni umani definiti da Max-Neef nel 1991. Esiste una stretta correlazione tra l'ontologia del cliente e le ontologie di prodotti e servizi; si presume infatti che un prodotto esista per soddisfare alcuni dei fondamentali bisogni umani, o di fatto non sarebbe stato prodotto (Ullah et al. 2016). In questa prospettiva, il concetto del "possedere degli oggetti" è correlato al possesso di un prodotto o di una particolare caratteristica del prodotto stesso che soddisfa uno o più bisogni fondamentali. Di conseguenza, le conoscenze necessarie per gestire il sistema di "reazioni" alle emozioni viene definito attraverso le relazioni tra le entità di tali ontologie, in base ai risultati di studi nella psicologia e nel marketing. Il cuore del sistema è lo Smart Engine (SE), che è caratterizzato da due moduli distinti: il Machine Learning Engine for Real-Time Actions (MLERTA) e il Machine Learning Engine for Decision Support Systems (MLEDSS). Il SE prende le sue decisioni sulla base degli algoritmi di Machine Learning che implementano regole logiche di inferenza, come ad esempio il CART, una delle implementazioni per gli Alberi Decisionali. Un modulo adibito alla gestione della Knowledge Base si occupa di mappare le informazioni in arrivo dallo Smart Engine tramite apposito linguaggio, come OWL, necessario per descrivere ed implementare le ontologie SO, PO e CO, ed aggiornarli ad orari prestabiliti. Sulla base dei dati gestiti dalla Knowledge Base, lo Smart Engine genera delle regole logiche sotto forma di ramificazioni "if-then-else", in base alle relazioni che collegano le varie entità delle

ontologie e le salva in un database, il modulo Rules Storage. Dato il duplice scopo della piattaforma, sono state definite due tipologie di regole: Action Rules e DSS Rules. Le Action Rules servono a gestire il comportamento reattivo del sistema, ossia le “reazioni”, attraverso la modifica delle caratteristiche dei servizi offerti (ad esempio, il numero delle casse aperte) e dell’ambientazione (ad esempio il colore dell'illuminazione) nel negozio, in base al livello di soddisfazione sperimentato dai clienti. In generale, le norme DSS potrebbero includere le cosiddette "business rules" (Hai et al., 2000), che consentono la gestione del comportamento dell'impresa secondo obiettivi e vincoli adeguati. Per definire le regole per il DSS, la tecnica delle Ontology-driven Business Rules potrebbero essere utilizzate al fine anche di definire un corretto modello d’impresa (Gailly et al., 2013). Per definire invece le Action Rules, sono stati considerati degli approcci omogenei o ibridi: entrambi sono utilizzati per risolvere problemi relativi alla rappresentazione della conoscenza, ma gli approcci ibridi si basano su modelli implementati attraverso i linguaggi di programmazione Answer Set (programmazione logica di tipo dichiarativo) come AnsProlog (Eiter et al., 2006), mentre quelli omogenei sono invece solitamente definiti utilizzando linguaggi come SWRL, che permette di definire delle regole logiche direttamente all'interno della knowledge base, utilizzando i concetti definiti nel linguaggio OWL. Ogni volta che lo Smart Engine riceve i dati in input provenienti dalla piattaforma di riconoscimento delle emozioni, relativi ad uno specifico touchpoint, esso prende una decisione sulla base della corrispondente Action Rule ed attiva le corrette reactions (le “reazioni”) per servizi e prodotti, come ad esempio fornire offerte personalizzate ad uno specifico cliente, cambiare musica nel negozio, ecc. Allo stesso tempo, tutti i dati ricevuti in input sono memorizzati in un database così da consentire le successive elaborazioni statistiche. Sulla base del risultato delle analisi statistiche, lo strumento di DSS fornisce al manager alcuni suggerimenti sulle possibili azioni da intraprendere per migliorare la pianificazione di una corretta strategia di CX, definendone obiettivi e vincoli.

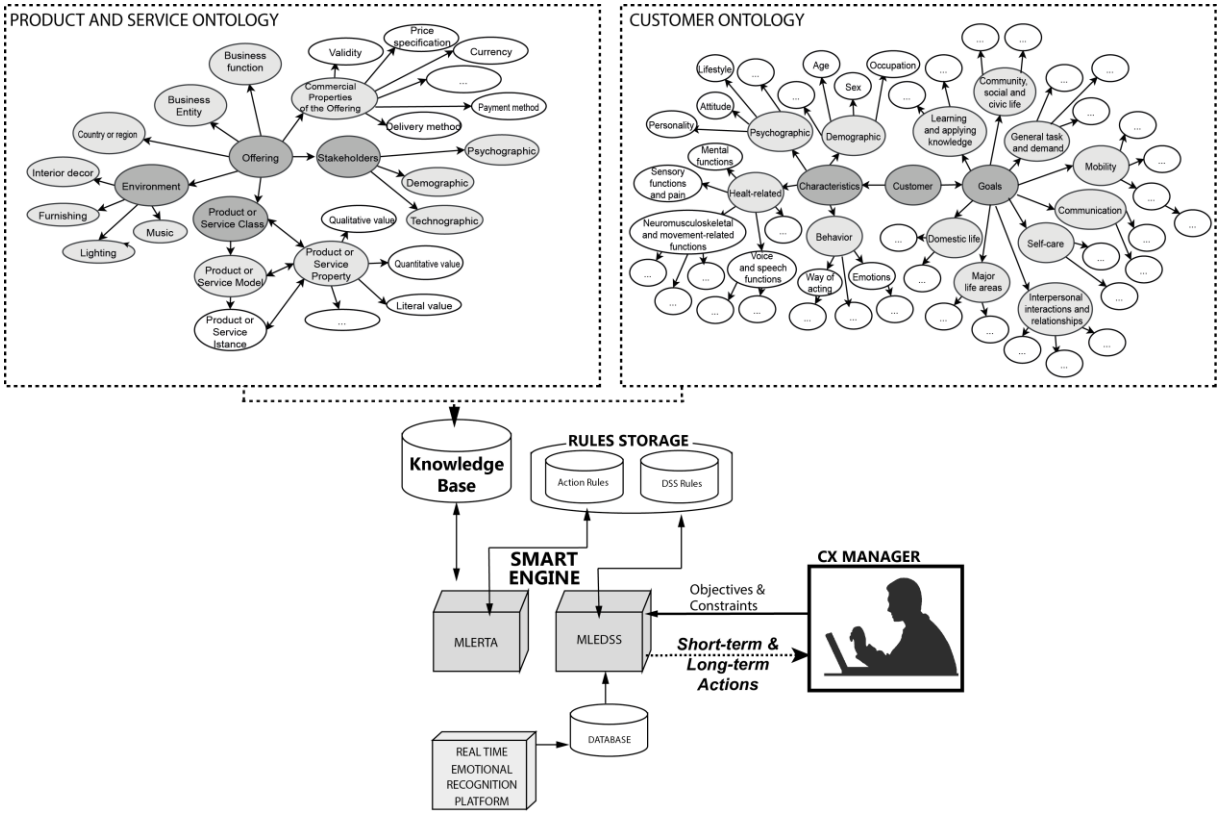

Figura 7: Lo Shopping Experience System Manager

Per identificare i clienti e rilevare il loro comportamento nel negozio, è possibile utilizzare i video forniti dalle telecamere di videosorveglianza o da altre IP camere o webcam installate ad hoc, così da consentire il tracciamento della posizione della persona, il riconoscimento del volto e il rilevamento delle espressioni facciali attraverso appositi strumenti SW basati su algoritmi di Deep Learning. In questo modo sarebbe possibile contrassegnare le persone con degli identificativi univoci e registrare i loro spostamenti all'interno del negozio, determinare facilmente il tempo che trascorrono nelle varie aree del negozio e il modo in cui interagiscono con i venditori. Come già accennato, tali sistemi possono essere implementati utilizzando le più diffuse tecnologie di acquisizione video (webcam, telecamere IP, CCTV, ecc.). Il sistema non terrà traccia in alcun modo dei dati biometrici dei clienti (quindi le loro foto), per cui non sarà in grado di riconoscere le persone che visitano nuovamente il negozio. Come succitato tutti i dati di output raccolti ed elaborati in tempo reale, vengono registrati in un database per permettere delle elaborazioni statistiche al fine di migliorare la Shopping e Customer Experience del cliente/utente, fornendo al CX Manager delle indicazioni per migliorare pianificazione e strategie.

Nello specifico, tali dati possono servire per addestrare un sistema basato su una Bayesian Belief Network (BBN), ossia una rete basata sui modelli Bayesiani, uno dei più semplici e al contempo maggiormente utilizzati in statistica e basato sull'inferenza.

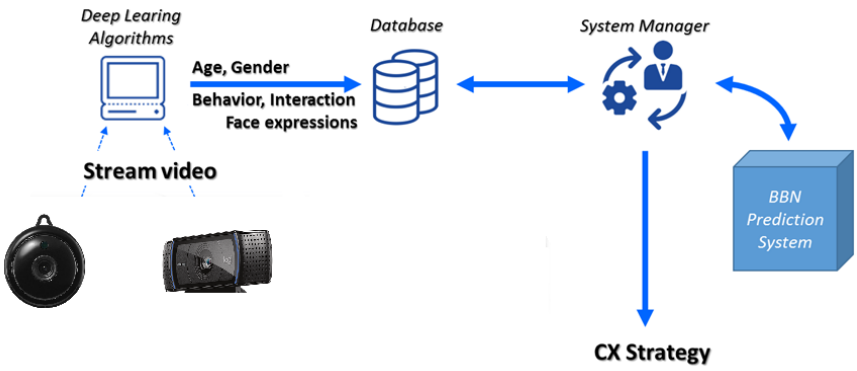

Figura 8: L'architettura del sistema con il BBN Prediction System

3.1 Gli scenari fisici e digitali

Date le caratteristiche di alto livello di questo sistema, è ora possibile definire alcuni scenari di utilizzo per le due macro aree in cui le applicazioni della piattaforma si collocano: lo scenario fisico e quello digitale.

3.1.1 Lo scenario fisico

Questo è il contesto in cui il cliente si interfaccia con prodotti fisici e tangibili, come possono essere i classici beni acquistabili sugli scaffali di un negozio, e l'interazione con i servizi offerti dal retailer riguarda il contatto umano con persone fisiche. In questo scenario si hanno le maggiori criticità riscontrabili nel cercare di fornire una buona Customer Experience: il contatto umano gioca un ruolo fondamentale sia nel bene che nel male, applicare correttamente le giuste azioni da intraprendere nel servire il cliente possono già fornire una buona parte del successo nel processo di vendita, così come possono inevitabilmente portare al fallimento e alla totale rovina del rapporto tra il cliente e un intero brand, di cui spesso anche un semplice commesso finisce per esserne il rappresentante. Un altro fattore critico negli scenari fisici, riguarda lo strumento d'acquisizione dati: la telecamera. Questo lavoro di tesi non vuole entrare nel merito di questioni legate alla privacy o alla psicologia (o quantomeno senza scendere troppo nei dettagli) ma è innegabile che l'utilizzo di telecamere per analizzare aspetti commerciali può portare ad una percezione negativa da parte del cliente, fino al punto di vanificare totalmente qualsiasi beneficio possa portare la piattaforma nel cercare di legarlo al brand, creando una connessione empatica con esso. Risulta dunque essenziale in questo contesto, da parte del retailer, applicare in maniera intelligente e

rispettoso del cliente qualsiasi applicazione di questa piattaforma perché essa abbia successo nello scopo che si prefigge, in particolar modo in questa tipologia di scenari.

3.1.2 Lo scenario digitale

Analogamente a quanto avviene per il fisico, lo stesso sistema può essere applicato anche agli scenari digitali, ossia in tutte quelle casistiche legate al retail in cui il cliente interagisce con prodotti e servizi appunto digitali e in cui il canale di connessione tra il cliente e il brand non è più il contatto diretto con un essere umano ma uno schermo collegato ad un elaboratore. In questi contesti è possibile avere maggior margine di manovra rispetto al fisico, senza dubbio un commesso maleducato può incidere maggiormente nella psiche di un cliente rispetto ad una interfaccia utente non particolarmente brillante: tuttavia è comunque essenziale che anche in queste casistiche la Customer e (in questo caso specialmente) la User Experience siano progettate ed applicate in maniera ottimale. Gli scenari digitali riguardano dunque tutti quei casi in cui l'utente interagisce con un totem di digital signage, il laptop di casa, il proprio smartphone… e soprattutto con webcam e camere integrate. Il vantaggio in questo caso è che l'invasività è estremamente ridotta rispetto ai contesti fisici, questo perché l'hardware utilizzato per l'acquisizione è di fatto lo stesso con cui l'utente interagisce tutti i giorni, rendendo l'impatto con una tecnologia "estranea" adottata appositamente per analizzarlo totalmente diverso. Casi tipici in cui la tecnologia di analisi può venire applicata riguarda ad esempio l'utilizzo di webcam per analizzare l'interazione del cliente con qualsiasi elemento di una piattaforma di e-commerce, le proprie espressioni facciali, le proprie caratteristiche, il modo con cui si approccia nello scorrere una particolare pagina della piattaforma, quasi tutto può essere monitorabile utilizzando i giusti tool software e una semplice webcam/camera integrata.

3.2 Le tipologie di dati ricavabili dal sistema

Analizzati i contesti digitali e fisici, quale tipologie di dati sarebbe dunque possibile ricavare in maniera automatica in queste due situazioni? Ognuno delle due tipologie di canali di vendita hanno pregi e difetti e tra le diverse caratteristiche vi sono senza dubbio le diverse tipologie di dati acquisibili, seppure ci sia una specifica tipologia acquisibile in ogni contesto esistente e sulla quale si focalizzerà il lavoro svolto per questa tesi, ossia i dati analizzabili dal volto del cliente tramite le telecamere e il Deep Learning.
Come espresso precedentemente, esistono tre tipologie di risposte alle interazioni del cliente con il brand, e sono viscerale, comportamentale e riflessiva. Considerando l'aspetto viscerale come l'aspetto emotivo provato ed esprimibile dal cliente, è stato scelto di affrontare l'analisi di queste due tipologie di risposte analizzate da Norman tramite l'acquisizione dei dati mostrati nella tabella 1.

	RISPOSTA VISCERALE	RISPOSTA COMPORTAMENTALE	RISPOSTA RIFLESSIVA
CANALE FISICO	- Rilevamento % di gioia, paura, disgusto, rabbia, sorpresa, tristezza	- Posizione dell'utente - Movimenti del corpo - Aree di osservazione	- Rilevamento % di gioia, paura, disgusto, rabbia, sorpresa, tristezza - % di coinvolgimento per ogni area di osservazione
CANALE DIGITALE	- Rilevamento % di gioia, paura, disgusto, rabbia, sorpresa, tristezza	- Azioni compiute su applicazione (tap, scroll, cambio schermata, ecc.) - Registrazione istante temporale per ogni azione compiuta su applicazione - Tempo trascorso da utente su ogni singola schermata - Aree di osservazione	- Rilevamento % di gioia, paura, disgusto, rabbia, sorpresa, tristezza - % di coinvolgimento per ogni area di osservazione

Tabella 1: Dati acquisibili in relazione al canale retail e alla risposta di Norman

Per quanto riguarda le colonne relative alla risposta viscerale e quella riflessiva, la differenza sostanziale è nel timing di acquisizione: i rilevamenti per acquisire la risposta viscerale dell'utente devono avvenire nei primi istanti di interazione con l'entità oggetto dell'analisi, solo in questo modo i dati saranno significativi per determinare questo specifico aspetto. I rilevamenti relativi alla risposta riflessiva, invece, riguardano gli istanti successivi alle acquisizioni avvenute per analizzare i primi due aspetti: solo una volta che l'utente avrà metabolizzato quanto avvenuto a seguito del primo impatto "viscerale" e alle interazioni successive con l'entità, sarà possibile estrapolare le risposte relative alla gradevolezza o meno dell'interazione "riflessiva".
Nello specifico, per quanto riguarda il canale fisico: i dati mostrati nella colonna delle risposte viscerali riguardano sostanzialmente le sei emozioni universali di Ekman analizzate nel capitolo 2, mentre quelli nella colonna successiva riguardano l'analisi della posizione dell'utente (utile ad esempio per tracciare il proprio Customer Journey all'interno del negozio), dei movimenti del proprio corpo e della propria attenzione rivolta a determinati prodotti piuttosto che ad altri.

Per quanto riguarda invece il canale digitale, si è scelto ancora una volta di analizzare le sei emozioni di Ekman e la relazione di queste con le aree in cui lo sguardo si è focalizzato maggiormente (aree di osservazione) mentre osservava uno schermo. Nella colonna successiva è possibile notare come sia stato scelto questa volta di acquisire dati di natura differente, ma sempre correlati con il comportamento dell'utente durante le proprie interazioni con il brand, quindi movimenti del mouse, click, aree della schermata osservate maggiormente, tempi di interazione ecc.

La cosa fin comune di questi dati richiedono sempre la stessa tecnologia per l'acquisizione: le telecamere. Oltre alla risposta del cliente, è possibile con le telecamere raccogliere anche altri dati che riguardano le caratteristiche del cliente stesso, nello specifico il **sesso** e **l'età**.

Capitolo 4.

La piattaforma tecnologica: soluzioni progettuali e implementative

Una volta fissata la metodologia si è proceduto a definire la struttura della tecnologia a livello progettuale e successivamente implementativo. Nel corso delle attività si sono succedute diverse versioni di questa piattaforma, che mano a mano si è arricchita di nuovi elementi, in ottica incrementale ed iterativa. Tra le caratteristiche principali e punto di forza di questa piattaforma vi è sempre stata la totale modularità, ossia la possibilità di utilizzare, a seconda dell'occorrenza, un modulo piuttosto che un altro, anche in base al canale (fisico o digitale) o ai vincoli tecnici ed ambientali.
Nella prima versione della piattaforma sono state valutate diverse tipologie di tool di terze parti e descritti più nel dettaglio nei prossimi paragrafi: in questo scenario la piattaforma in questione fungeva da integratore di approcci e tecnologie utilizzate solitamente stand-alone. Nel corso del tempo sono state sviluppati internamente alcuni dei software presi in considerazione per la prima versione, come per esempio il motore di riconoscimento delle emozioni dalle espressioni facciali, o quello per il sesso e l'età, abbandonando dunque la maggior parte dei tool citati nel prossimo paragrafo, come Affdex e VPGLIB.

4.1 La prima versione della piattaforma: il canale fisico

Questo strumento aveva lo scopo principalmente di analizzare le emozioni dei clienti nei contesti fisici in modo non intrusivo. Esso era composto da quattro moduli principali (Figura 9). Il primo modulo identifica una persona ogni volta che questa viene rilevata in prossimità di un Touchpoint. Gli altri tre moduli, che permettono di acquisire e analizzare le informazioni emotive, sono i moduli di riconoscimento delle espressioni facciali, di riconoscimento vocale e di analisi dei biofeedback.
Il modulo di riconoscimento delle espressioni facciali si avvale di IP Camere FullHD dotate di tecnologia PTZ con autofocus. Ogni telecamera, installata in corrispondenza di ogni touchpoint, invia continuamente uno stream video al server centrale, che elabora ogni fotogramma del video stesso e restituisce la misura delle emozioni del cliente. Questo modulo aveva incorporato l'SDK open-source di Affectiva, Affdex, che fornisce in output un valore percentuale associato all'intensità delle principali emozioni di Ekman: gioia, tristezza, rabbia, paura, disprezzo, disgusto e sorpresa. Inoltre, questo modulo fornisce misure per l'Engagment, la misura di quanto il soggetto è "ingaggiato" e quindi emotivamente coinvolto, e il Valence, che dà una misura di positività o negatività dell'esperienza. Questo motore si basa su modelli di Deep Learning, ed in particolare sulle Convolutional Neural Networks (CNN) e Recurrent Neural Networks (RNN) con layer personalizzati, per riconoscere le espressioni facciali e le relative emozioni principali di Ekman. Nello specifico, scopo di

questo modulo è quello di riconoscere le Action Units del FACS di Ekman descritto nel capitolo 2 e ricondurre queste alle emozioni universali.
Il modulo di riconoscimento vocale si riferisce a tracce di registrazioni del parlato raccolte durante sondaggi organizzati o registrati da microfoni installati in ogni touchpoint. Anche in questo caso il software sarà integrato con un motore di riconoscimento delle emozioni già esistente. Diversi strumenti possono essere utilizzati, a seconda dell'approccio di analisi del parlato che si vuole adottare. Ad esempio, è possibile utilizzare le API di Google per convertire la voce umana in testo scritto e il software Watson di IBM per l'analisi emozionale del testo. Altrimenti, è possibile adottare lo strumento di PeakProfiling (ex AudioProfiling) per estrarre direttamente le emozioni dalle feature della voce: volume, articolazione, tempo, ritmo, melodia e timbro.
Il terzo modulo permette l'analisi dei dati biometrici attraverso l'acquisizione del battito cardiaco e/o della frequenza respiratoria. Tali bio-informazioni sono solitamente monitorate utilizzando sensori intrusivi come l'ECG, tuttavia, VPGLIB (ex Qpulsecapture) viene utilizzato per monitorare tali parametri in modo non intrusivo. L'applicazione è una libreria estensione delle OpenCV, che utilizza l'elaborazione digitale delle immagini per estrarre la frequenza cardiaca e fornisce una stima della frequenza respiratoria dal video del volto umano. In questo modo è possibile acquisire la frequenza respiratoria e il battito cardiaco della persona attraverso la stessa telecamera utilizzata per il riconoscimento facciale, con un errore assoluto inferiore a 5 bpm nella maggior parte dei casi (VPGLIB, 2017).
Il passo successivo è quello di ricondurre queste misure ai valori percentuali emozionali di gioia, rabbia, ecc. È stato dimostrato che esiste una correlazione tra queste misurazioni e lo stato emotivo (Quintana et al., 2012), per questo scopo erano state studiate le API di SensauraTech (oggi non più esistente) al fine di ricondurre appunto i biosegnali tracciati alle relative emozioni. Per garantire l'identificazione del cliente ad ogni touchpoint, il modulo di identificazione implementa un motore di riconoscimento facciale che utilizza un database di immagini precedentemente memorizzate. Tali immagini possono essere raccolte durante la registrazione del cliente richiesta per la carta fedeltà; in questo modo, i clienti vengono identificati ogni volta che passano davanti ad una telecamera. A questo scopo possono essere utilizzate diverse API, come la Face Recognition API di Lambda Labs.
Una web GUI (Fig. 10) visualizza i dati forniti in uscita dalla piattaforma di riconoscimento delle emozioni: tale interfaccia fornisce una rappresentazione della curva emotiva dei clienti in funzione del valore temporale e del valore di valence, misurati su una scala da
-100 a 100; inoltre, visualizza le percentuali relative alle emozioni primarie.
I dati in tempo reale relativi a ciascun touchpoint possono essere tracciati per uno o più clienti. In quest'ultimo caso vengono forniti i valori medi aggregati.

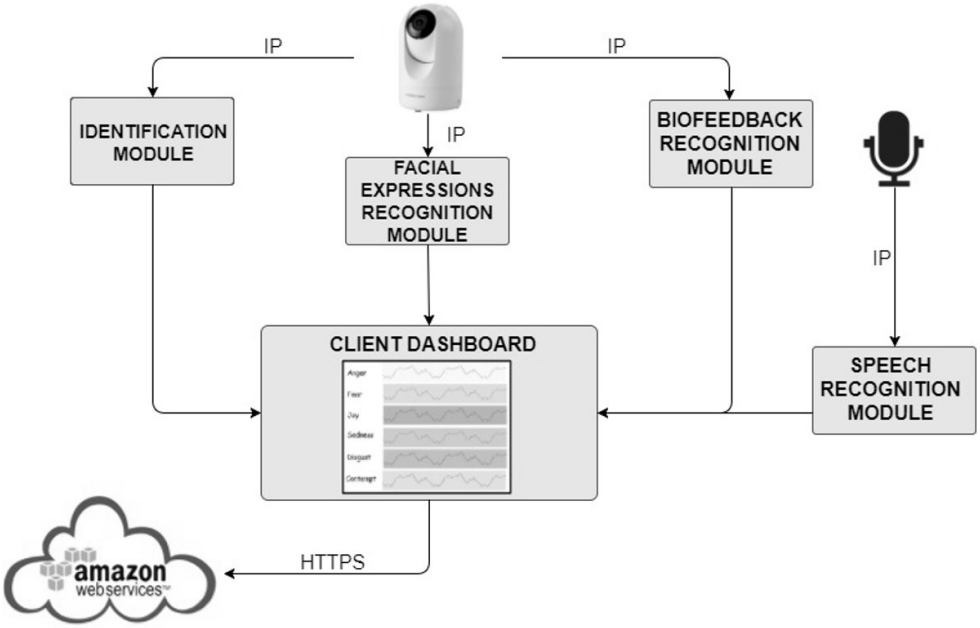

Figura 9: L'architettura della prima versione della piattaforma

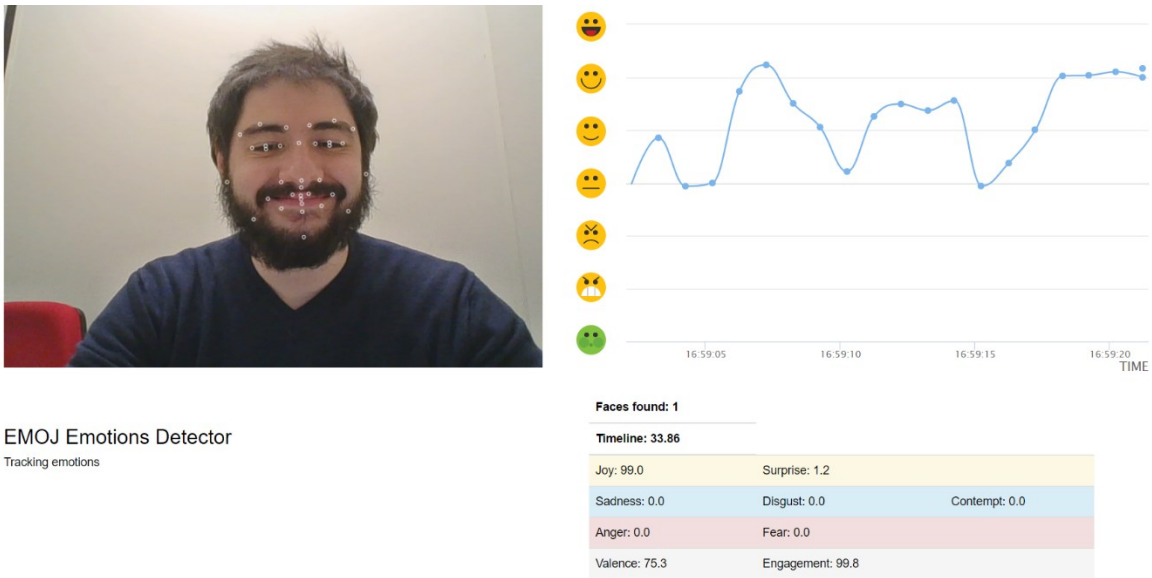

Figura 10: La prima versione dell'interfaccia grafica con valence e percentuali emotive

4.2 La seconda versione della piattaforma

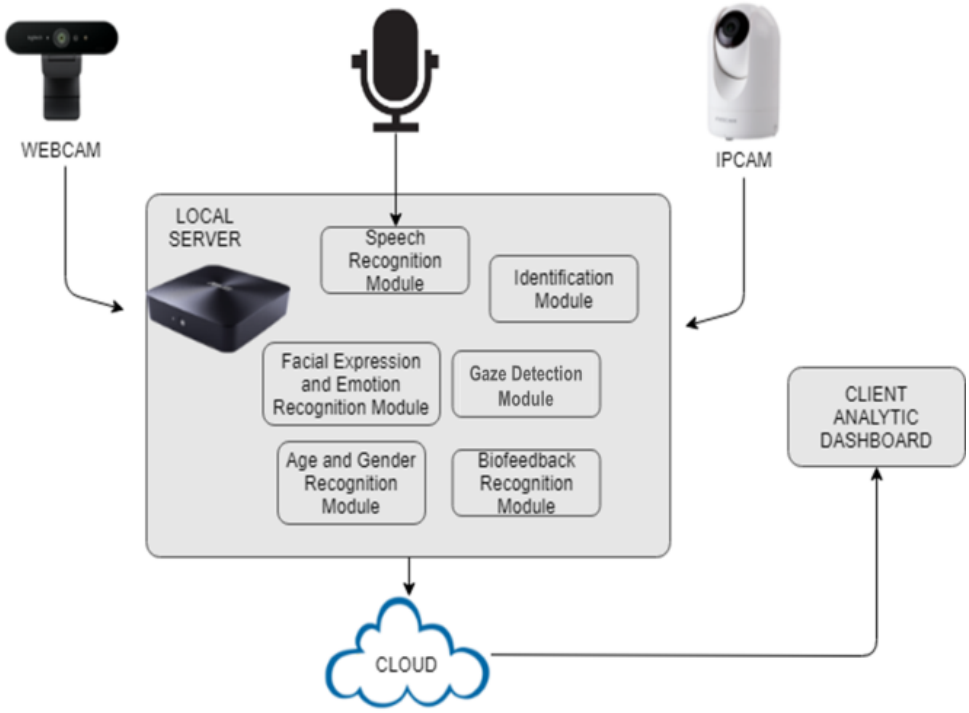

Figura 11: L'architettura della seconda versione della piattaforma

Rispetto alla prima versione sono stati aggiunti due nuovi moduli oltre ai precedenti: il modulo di riconoscimento di sesso ed età e il modulo di riconoscimento dello sguardo; si è inoltre tenuto conto di alcuni vincoli ambientali in fase di installazione, come il caso in cui ci sia la possibilità di installare telecamere ad altezza uomo, per cui di solito può essere utile adottare delle webcam. Le telecamere, installate in corrispondenza di ogni touchpoint, inviano continuamente flussi video al server centrale, che elabora ogni fotogramma video e restituisce la misura delle emozioni, dell'età, del sesso e delle coordinate dello sguardo del cliente. Il modulo Facial Expressions and Emotions Recognition incorpora una Convolutional Neural Network sviluppata autonomamente e che, prendendo in input tutti i diversi frame che compongono il flusso video, fornisce in output un valore percentuale associato all'intensità delle principali emozioni di Ekman (Gioia, Tristezza, Rabbia, Paura, Paura, Disgusto e Sorpresa).
La Rete Neurale Convoluzionale, che verrà approfondita nei successivi paragrafi, è stata addestrata utilizzando due diversi dataset crowdsource: il FER+ (Barsoum et al., 2016) ed EmotioNet (Fabian-Benitez-Quiroz, 2016). La prima usa le emozioni di Ekman per etichettare tutte le circa 36000 foto che compongono il dataset, mentre la seconda ha circa un milione di immagini scaricate ed etichettate con le action unit del FACS. Anche il modulo di

riconoscimento di sesso ed età è basato su una CNN, addestrata con il dataset IMDB-Wiki, composto da più di 500000 immagini scansionate ed etichettate dal sito IMDB.
L'altro modulo basato sul Deep Learning è il modulo di Gaze Detection. Tale modulo implementa nuovamente una CNN, addestrata con dataset opensources e altri dataset ottenuti utilizzando piattaforme di crowdsourcing, come Microworkers e Amazon Mechanical Turk. Questo modulo prende in ingresso un flusso video ottenuto dalla telecamera, lo divide in diversi fotogrammi e analizza ogni fotogramma in modo da rilevare il volto e la posizione della pupilla attraverso la rete neurale e abbinando le coordinate risultanti con un'istantanea di ciò che l'utente stava realmente guardando durante le riprese video.
Questo tipo di strumento può essere utile per avere una stima di quali tipologie di prodotto attirano effettivamente lo sguardo del cliente rispetto agli altri. Tale modulo fornisce risultati in output attraverso una serie di heatmaps disegnate utilizzando come base le istantanee costruite a partire dalle foto dell'ambiente nel negozio, oppure utilizzando schemi che rappresentano nel dettaglio i touchpoint.

4.3 La progettazione della piattaforma per il canale digitale e l'SDK per smartphone

Per supportare pienamente la valutazione della Customer (e più in generale anche della User) Experience delle applicazioni mobile in the wild, ossia in ambienti non controllati come in laboratorio, è stato sviluppato un SDK per dispositivi mobile iOS in grado di fornire i dati demografici degli utenti (cioè età e sesso), i dati sulle prestazioni (tempo per navigare in una schermata) e i dati di utilizzo (scrolling e taping). Inoltre, esso sfrutta un sistema di riconoscimento dello sguardo (gaze tracking) e di riconoscimento delle emozioni per consentire la raccolta di informazioni comportamentali.
Il fatto che l'SDK sia stato implementato solamente per iOS è dovuto al motore utilizzato per il riconoscimento dello sguardo e che verrà approfondito successivamente: la rete neurale su cui esso si basa richiede in input, tra gli altri dati, anche il riferimento spaziale della posizione della telecamera, che viene utilizzato come origine degli assi nel sistema di coordinate preso a riferimento dai dati in output della rete. Questa informazione, per automatizzare il sistema, può essere reperita attraverso datasheet che nel caso dei dispositivi Android sono di difficile reperibilità. Un'altra motivazione riguarda la notevole eterogeneità nel panorama di dispositivi esistenti per Android rispetto agli iPhone, cosa che ha reso altrettanto complicato riuscire ad automatizzare un sistema in grado di reperire le informazioni necessarie per la raggiunta di una buona accuratezza da parte della rete per il riconoscimento dello sguardo, come la dimensione degli schermi.

4.3.1 Le tipologie di architettura adottabili

La comunicazione tra i dispositivi mobile e il server remoto avviene in maniera sicura tramite chiamata POST con certificato SSL, quindi in HTTPS: ciò garantisce una maggior sicurezza data la sensibilità del materiale inviato.
Si propongono 3 diverse soluzioni architetturali per poter incrementare il grado di sicurezza nel trattamento del dato.

- Distribuita: l'elaborazione delle foto avviene lato client, ossia direttamente sul dispositivo mobile – quest'approccio evita di esportare verso un server remoto la foto dell'utente (ad essere inviata al server non è più la foto ma solo il dato già elaborato) al prezzo di una quantità maggiore di dati da scaricare dopo che l'applicazione è stata installata.

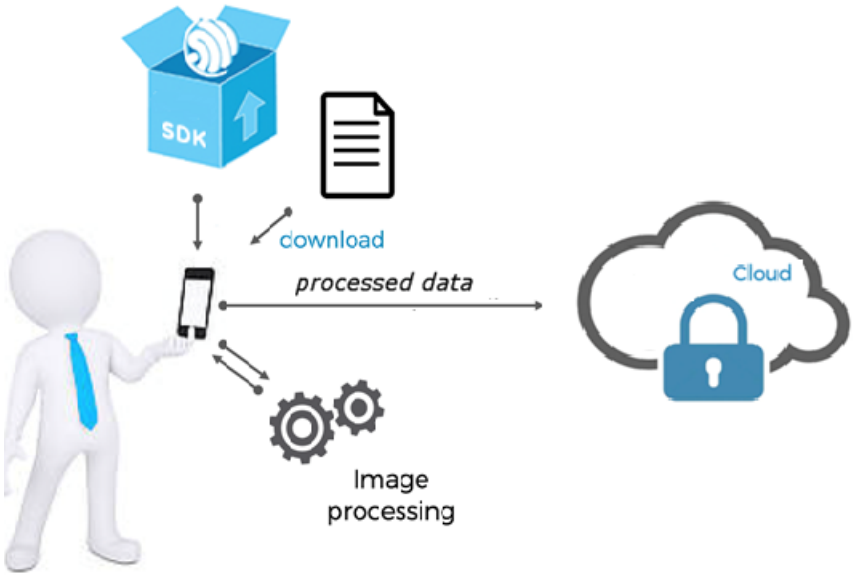

Figura 12: Uno schema dell'architettura distribuita per mobile

- Centralizzata: il dispositivo mobile invia delle foto ad un server esterno che quindi si occupa di tutta l'elaborazione – in questo caso il traffico dati in uscita aumenta (sebbene di poco) ma l'intera applicazione risulta molto più leggera.

Figura 13: Uno schema dell'architettura centralizzata per mobile

- Ibrida: il dispositivo opera una preprocessing sulla foto finchè ad essere inviato all'esterno non è più la foto stessa ma una serie di dati pre-processati che verranno inviati al server per l'elaborazione finale; ciò garantisce un buon alleggerimento del traffico in uscita rispetto all'approccio centralizzato, con una riduzione delle dimensioni finali dell'app.

Figura 14: Uno schema dell'architettura ibrida per mobile

4.3.2 L'implementazione dell'architettura centralizzata

Questo sistema si avvale di un'architettura centralizzata che, come mostrato in Figura 15, ha due attori principali: l'SDK iOS lato client e la piattaforma di Deep Learning (DL) lato server. L'SDK mobile è un framework per iOS che espone alcune API per monitorare tutte le interazioni degli utenti durante l'utilizzo delle applicazioni mobile. Tra queste caratteristiche, c'è la possibilità di attivare la fotocamera che scatta diverse foto con una certa frequenza impostabile.

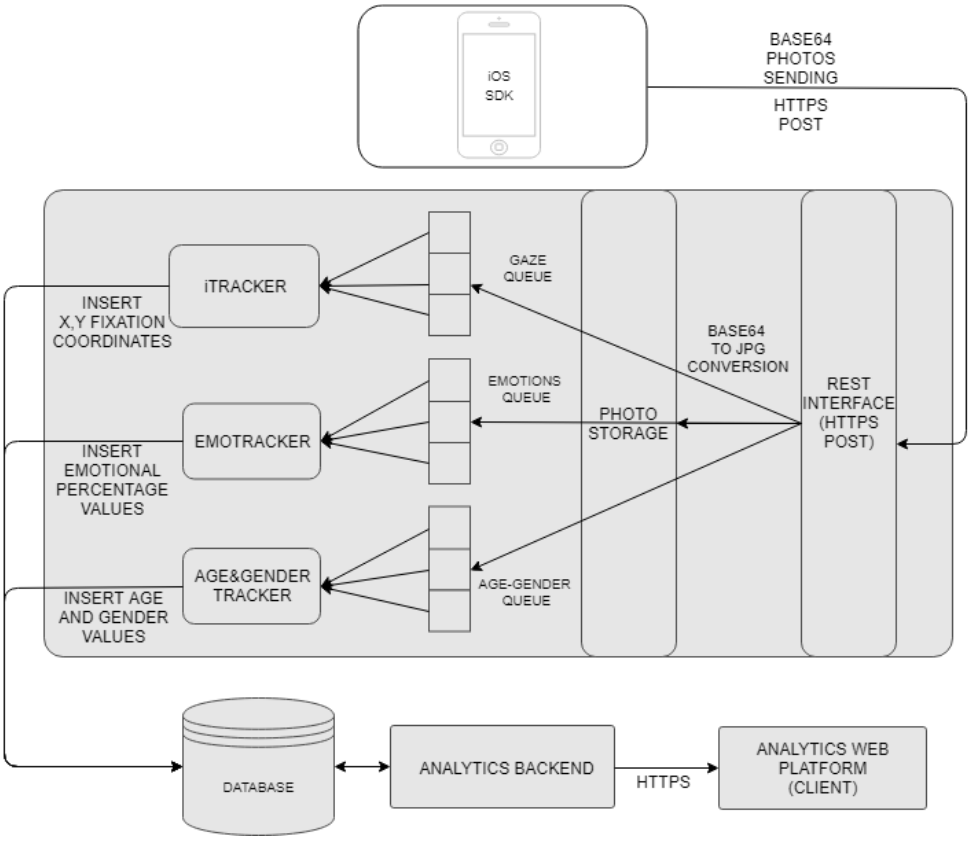

Figura 15: L'architettura a servizi della piattaforma

Queste foto sono codificate in Base64 e inviate ad un server con protocollo HTTPS. Il server centrale che supporta tutta l'architettura della piattaforma, gestisce le chiamate in arrivo dal framework iOS attraverso un'interfaccia REST sviluppata in Python, che attende le chiamate POST HTTPS indirizzate all'endpoint esposto. Una volta ricevuta la chiamata, il contenuto viene analizzato e decodificato per ottenere tutti i dati, incluso il JPEG originale successivamente memorizzato nella memoria fisica. Successivamente, il nome del file JPEG, che identifica univocamente la foto, viene memorizzato in tre diverse code Redis in modo che, attraverso il percorso della directory in cui si trovano fisicamente i file, è possibile per ogni Tracker Module della piattaforma DL ottenere la posizione delle foto ogni volta che queste arrivano. Queste code sono così utilizzate dai tre Tracker Modules per ottenere rispettivamente la stima delle coordinate x-y dello sguardo dell'utente, del suo stato emotivo, del sesso e dell'età. Tutti i DL Tracker Modules sono basati su Convolutional Neural Networks (CNN) implementate in Python. Ogni volta che l'elaborazione di una foto è terminata per tutte le CNN, i dati risultanti saranno memorizzati in un database e la foto stessa sarà definitivamente cancellata dal server. Tutti i dati memorizzati saranno disponibili attraverso una piattaforma web di analisi.

4.3.2.1 iTracker

Questa piattaforma implementa una CNN per il tracciamento dello sguardo presentata nell'articolo "Eye Tracking for Everyone" (Krafka et al., 2016), chiamata iTracker, che è stata addestrata con un dataset su larga scala etichettato appositamente per il riconoscimento dello sguardo su dispotivi iOS (tramite il tool GazeCapture) e che contiene dati provenienti da 1474 soggetti (~2.5 milioni di frames). Tale dataset è stato raccolto attraverso il crowdsourcing, in modo che la grande eterogeneità di dati permetta di migliorare la robustezza del modello.
La CNN prende in input le immagini dell'occhio sinistro, dell'occhio destro e del volto rilevato e ritagliato dall'immagine originale, e una maschera binaria (griglia facciale) utilizzata per indicare la posizione e le dimensioni del volto all'interno del fotogramma, e fornisce in output le coordinate x-y prendendo come origine degli assi la posizione della fotocamera (in centimetri); queste coordinate rappresentano proprio la predizione di dove si è posato lo sguardo dell'utente sullo schermo relativamente al frame ricevuto in input.

4.3.2.2 EmoTracker

Questo modulo sfrutta la stessa CNN presentata nel pragrafo 4.2, basata su una versione riveduta della VGG13. Si compone di 10 layers convoluzionali distanziati tra loro da layers di max pooling e dropout ed è stata addestrata da zero sulla base del dataset FER+ ed EmotioNet. La CNN prende in ingresso un'immagine in scala di grigi allineata e scalata a 64x64 pixel e restituisce una probabilità in percentuale per ogni emozione di Ekman.

4.3.2.3 AgenderTracker

La stima dell'età e del sesso viene effettuata da un'altra CNN addestrata da zero sulla base del dataset IMDB-WIKI adottatando l'architettura Wide Residual Network (WideResNet) e aggiungendo due layers di classificazione: uno con 101 unità per la stima dell'età e un altro con 2 unità per la classificazione di genere. I fattori di approfondimento e di ampliamento della rete sono fissati rispettivamente a 16 e 8. Il modello è alimentato con immagini del volto allineate e scalate a 256x256 pixel, e fornisce in output le probabilità in percentuale di età e di genere.

4.3.2.4 L'SDK

Il lato client di questa piattaforma è un framework progettato per gli smartphone Apple. Questo framework, una volta incorporato nelle applicazioni iOS, permette di inviare al server tutte le attività svolte dall'utente con l'applicazione. In primo luogo, utilizzando la fotocamera frontale dello smartphone, scatta foto in silenzio con una frequenza di 0,5 Hz (impostabile) e le invia al server remoto. In secondo luogo, ogni volta che un utente tocca lo schermo, memorizza le coordinate x-y del punto intercettato (in pixel). Vengono prese in considerazione anche le informazioni sulle attività di scorrimento della schermata (scrolling), timestamp di acquisizione del dato e di avvio della schermata corrente, screenshot della schermata osservata ed ogni volta che l'utente fa scorrere la barra di scrolling, viene ricavato

l’offset sull’asse y dall'angolo in alto a sinistra dello schermo. Tutti questi valori vengono inviati al server ogni volta che viene eseguita una delle suddette azioni.

4.3.2.5 Piattaforma di Analytics

È stata in seguito sviluppata una piattaforma web (Figura 16) per visualizzare i dati registrati nel database sull'utilizzo dell'applicazione. Le informazioni riguardanti i tap (click con il dito sullo schermo) e lo sguardo, ossia le coordinate x-y, sono utilizzate per generare mappe termiche (heatmap).

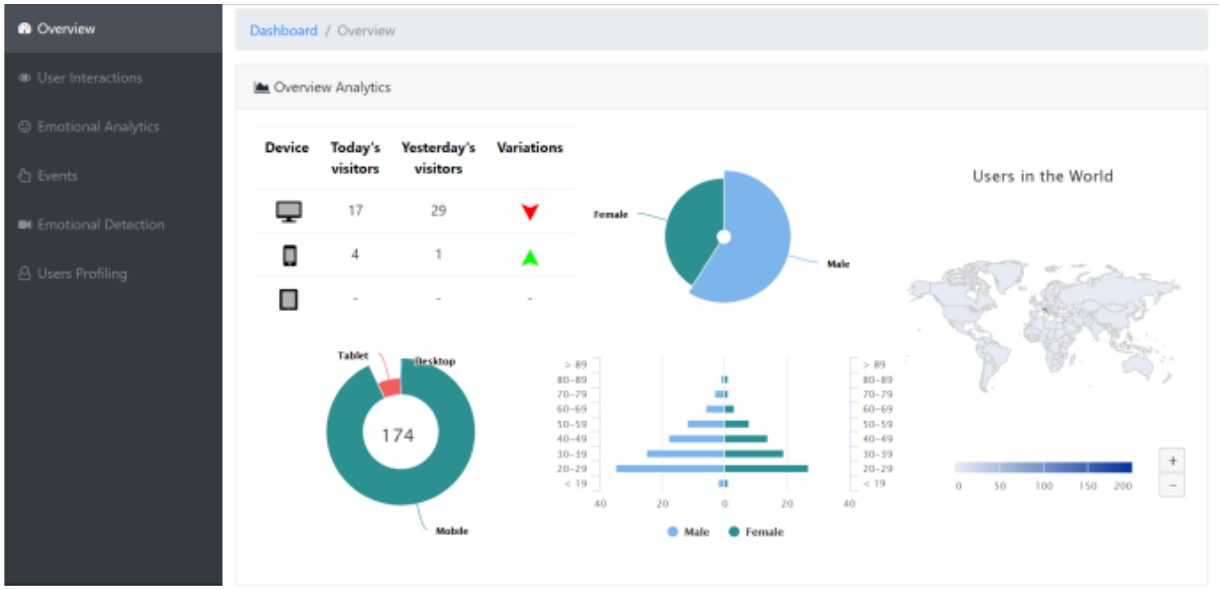

Figura 16: La sezione Overview della piattaforma di Analytics

Sono stati successivamente effettuati alcuni test preliminari su iPhone 4, iPhone 6s e iPhone 7 Plus per determinare il consumo di CPU e batteria e l'allocazione della RAM necessaria, variando la luminosità dello schermo e la frequenza di frame shot. I risultati sono riportati nella tabella seguente.

Shots frequency	Screen brightness	iPhone 4			iPhone 6s			iPhone 7 Plus		
		CPU (%)	RAM (MB)	Battery (%)	CPU (%)	RAM (MB)	Battery (%)	CPU (%)	RAM (MB)	Battery (%)
0.5 Hz	50%	25%	10	11%	7%	21	7%	14%	30	4%
	100%			16%			15%			13%
1 Hz	50%	28%	6	12%	9%	15	9%	19%	38	5%
	100%			17%			15%			14%
3 Hz	50%	56%	9	12%	15%	21	10%	24%	90	7%
	100%			19%			18%			15%

Tabella 2: Le prestazioni e i consumi dell'SDK per iOS

I futuri miglioramenti del sistema riguardano principalmente le versioni SDK per framework Java/Android e multipiattaforma, e modelli CNN più leggeri da incorporare direttamente all'interno dell'SDK, così da gestire il processamento in locale e ridurre il traffico di rete richiesto per l'invio delle foto al server (architettura distribuita).

4.3.3 La piattaforma per il web

Concettualmente l'architettura presentata nel paragrafo precedente è esattamente replicabile anche per siti e piattaforme web: l'unica sostanziale differenza è lato client, in cui l'SDK per smartphone è sostituito da un SDK in Javascript o da un plugin per CMS dedicati allo sviluppo di e-commerce, come Prestashop e Magento. Esattamente come per l'SDK per iPhone, questo componente è stato sviluppato per essere integrato in piattaforme web preesistenti e fornisce agli sviluppatori delle API per l'implementazione di funzionalità in grado di ricavare le coordinate dei click sullo schermo, le percentuali di scrolling, i timestamp relativi agli istanti di acquisizione e di permanenza nelle pagine, le informazioni strutturali relative agli elementi delle pagine stesse, gli screenshot delle pagine osservate e le foto scattate da webcam e convertite in base64. Ancora una volta, queste informazioni vengono inviate con una frequenza prefissata allo stesso server precedentemente analizzato per l'analisi delle foto ricevute.

4.4 La versione finale della piattaforma: riepilogo delle caratteristiche principali e ultime soluzioni tecnologiche integrate

Lo scopo di questa piattaforma è quello di studiare le caratteristiche intrinseche di uno o più utenti ed i loro comportamenti durante l'interazione con uno spazio fisico e/o digitale mediante l'utilizzo primariamente di telecamere, e solo qualora sia necessario e senza mai minare il fondamento della non invasività su cui si basa il sistema proposto in questa tesi, di altri sensori di varia tipologia.
Tutte le informazioni, i dati analizzati ed i risultati ottenuti, vengono utilizzati per attuare soluzioni ed azioni che coinvolgono l'utente o vengono archiviati (ad eccezione delle foto dell'utente). Lo spazio ove vengono effettuate queste rilevazioni può essere uno "spazio fisico" per attività di retail come un negozio, un grande magazzino, un'area, uno showroom, una fiera, uno spazio domestico o di utilizzo privato, ma anche uno spazio fisico per attività socio-culturali come un museo, una mostra, una apposita area turistica, un hotel od altro. All'interno di queste aree, frequentate da persone fisiche, sono presenti delle telecamere o dei sensori posizionati ove si ritiene più opportuno, in modo tale da poter monitorare e seguire le persone che si trovano nei pressi delle aree espositive. Differentemente, lo "spazio" può essere anche digitale creato cioè mediante l'utilizzo di piattaforme web, applicazioni mobile quali tablet, cellulari, computer portatili o applicazioni di digital signage attualmente molto utilizzate come forma di comunicazione in prossimità di punti vendita, spazi pubblici aperti o chiusi, anche privati o domestici, in cui i contenuti vengono mostrati agli utenti attraverso schermi digitali. In questo caso è possibile utilizzare, per la raccolta dei dati, le camere già integrate nei dispositivi o webcam esterne. In qualsiasi situazione di "spazio di analisi", lo scopo primario è quello di studiare le caratteristiche intrinseche degli utenti e i loro comportamenti estrapolando espressioni facciali, posizioni del corpo, emissione di calore, movimenti ricorrenti ed altri aspetti fisici, caratteriali o comportamentali che, sulla base delle ricerche dello psicologo statunitense Paul Ekman, portano ad estrapolare ed individuare le emozioni di "Gioia", "Sorpresa", "Disgusto", "Rabbia", "Tristezza", "Paura" e "Neutralità".
Una delle innovazioni introdotte da questa piattaforma è la modularità, quindi la possibilità di utilizzare a seconda del contesto le tre o più diverse modalità di analisi che, lavorando contemporaneamente, permettono di raggiungere la perfetta analisi del comportamento e delle emozioni.
La procedura di acquisizione e processamento delle immagini viene esplicata nel diagramma di flusso rappresentato in figura 17.

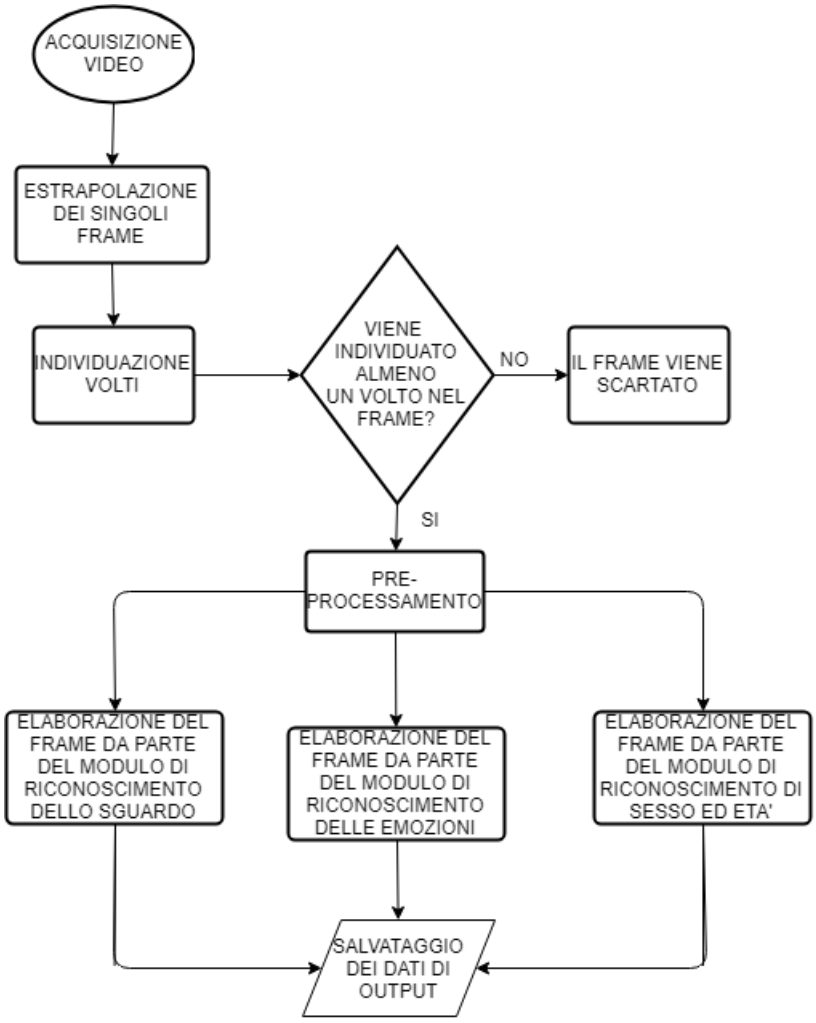

Figura 17: Il diagramma di flusso del procedimento di elaborazione delle immagini

Ciò che accomuna i tre moduli primari è una prima fase di individuazione dei volti all'interno delle immagini tramite l'utilizzo di camere, video-camere, webcam, IP camere, camere RGBD o telecamere di video sorveglianza o altre tipologie di sensori idonei alla cattura delle immagini. Il principio di funzionamento è idoneo sia per lo scenario digitale che quello fisico e consiste nella prima fase di acquisizione da parte delle telecamere dei tracciati video che vengono inviati successivamente al server principale, fotogramma per fotogramma, nel caso di spazi digitali, o direttamente al computer collegato alla/e telecamera/e (nel caso di spazi fisici), che si occupa della trasformazione del video in singole immagini e del pre-processamento dei singoli frames, raddrizzamento del volto e crop mediante ingrandimento

sul dettaglio del volto stesso; il procedimento in pratica è un riconoscimento (idoneo ad essere poi analizzato dai software) del volto denominato tecnicamente "Face Detection".
Viene riportata in figura 17 la rappresentazione della architettura Hardware – Software del sistema relativo ad uno spazio fisico idoneo alla gestione dei casi prospettati.
Il pre-processamento succitato e implementabile nelle applicazioni fisiche, permette di snellire ed aumentare la velocità di elaborazione dati al fine di rendere il sistema più reattivo ed immediato. Una volta inviati i dati al server centrale il software sviluppato permette di monitorare tutte le interazioni dell'utente nello spazio fisico mediante i dati provenienti dai sensori.
Ogni immagine o dato rilevato viene poi suddiviso ed inviato ai Tracker Modules che compiono ciascuno una funzione di elaborazione differente. Ogni modulo è specializzato in una tipologia di analisi e predizione differente: il modulo di Riconoscimento Facciale è il modulo che gestisce i dati ricevuti dalle camere, convertiti in digitale, e che tramite due tipologie di modelli specializzati nel riconoscimento, uno dei volti lontani e l'altro dei volti vicini, ha il compito di riconoscere tutti i volti umani presenti in una delle immagini che compongono il flusso video ricevuto in ingresso. Questi modelli restituiscono le coordinate di un riquadro all'interno dell'immagine che, secondo la loro predizione, conterrebbe un volto umano.
La seconda fase di questo modulo, che è dunque anche adibito al pre-processamento dell'immagine, è quello di ritagliare tutti i volti dall'immagine originale in base alle coordinate prima fornite dall'applicazione del modello succitato, ridimensionarli e convertirli nello stesso formato con cui sono state addestrate le reti neurali convoluzionali. Il secondo modulo di Tracciamento delle Emozioni consiste in una Rete Neurale Convoluzionale che riceve in ingresso l'immagine di un volto riconosciuto ed elaborato precedentemente e che ritorna in uscita la percentuale di probabilità che quel volto appartenga alle categorie di neutralità, felicità, sorpresa, tristezza, rabbia, disgusto e paura, assieme all'intensità dell'emozione provata dalla persona inquadrata. Questo modulo ha anche il compito di calcolare, tramite una media pesata dei valori associati alle singole emozioni rilevate, la valence.
Il modulo di Tracciamento Sesso ed Età consiste in una Rete Neurale Convoluzionale che, sempre a partire dall'immagine di un volto ritagliato dal modulo di Riconoscimento Facciale, predice il sesso e l'età puntuale di una persona.
Il modulo di Tracciamento dello sguardo consiste in una CNN che predice le coordinate rispetto al piano osservato da una persona (lo schermo di un pc o di uno smartphone per il digitale o uno scaffale per il fisico) e il livello d'attenzione rispetto a ciò che sta osservando.
Gli altri moduli riguardano il rilevamento della posizione di una persona all'interno dello spazio fisico tramite tecnologie come RFID o iBeacon, e dei suoi movimenti tramite camere e il tool opensource OpenPose (Cao et al., 2018) e l'analisi delle emissioni di calore al fine di migliorare la predizione dello stato emozionale tramite termocamere. Una volta estrapolati, mediante l'utilizzo dei diversi moduli, i dati vengono archiviati in un database, mentre le immagini e quanto rilevato dai sensori viene definitivamente cancellato dalle memorie dei computer. Questi dati possono venire utilizzati successivamente per venire adoperate da una piattaforma di "Web Analytics", ossia una piattaforma di Business Intelligence che mostrerà tutta una serie di statistiche relative alla Customer e User Experience di un cliente/utente durante la propria esperienza. Questa piattaforma a livello architetturale è composta da una parte di Backend gestita da un server remoto su Cloud, e da una parte client rappresentata

dall'elaboratore che permetterà la visualizzazione dell'interfaccia utente della piattaforma di "Web Analytics".

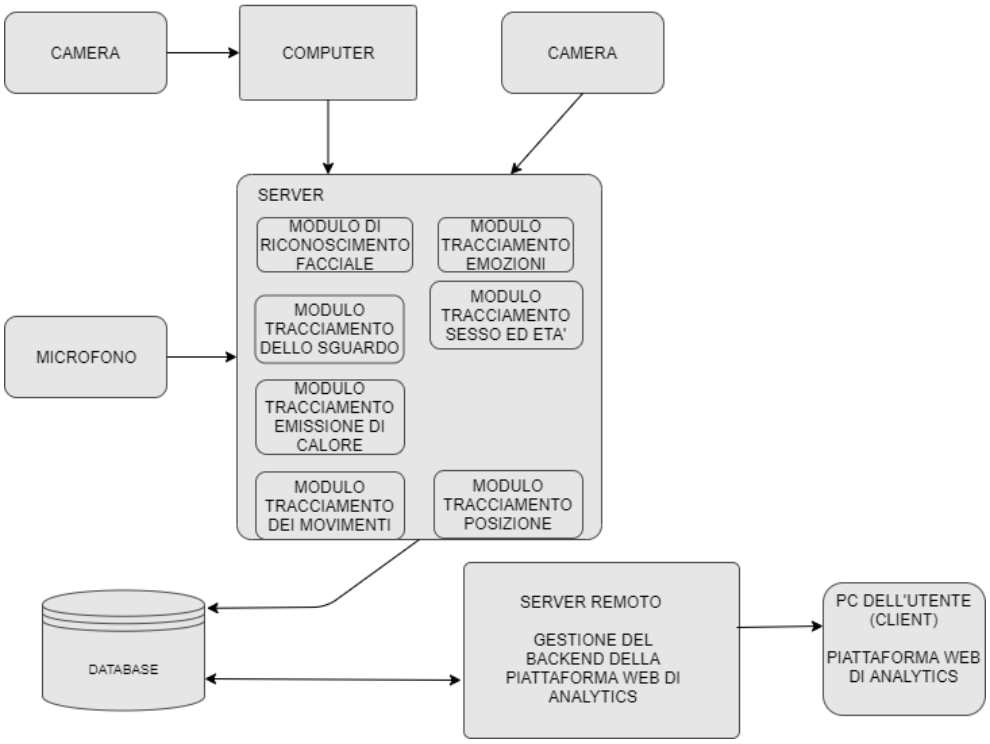

Figura 18: L'architettura finale della piattaforma per il settore Fisico

Analogamente l'architettura per gli spazi digitale, mostrata in figura 19, implementa gli stessi moduli di tracciamento sopra descritti, a parte quelli specificatamente progettati per un utente "fisico", come quelli per il tracciamento della posizione e dei movimenti. In questo caso la struttura è la stessa descritta nel paragrafo 4.3.2 ed implementa una struttura a servizi.

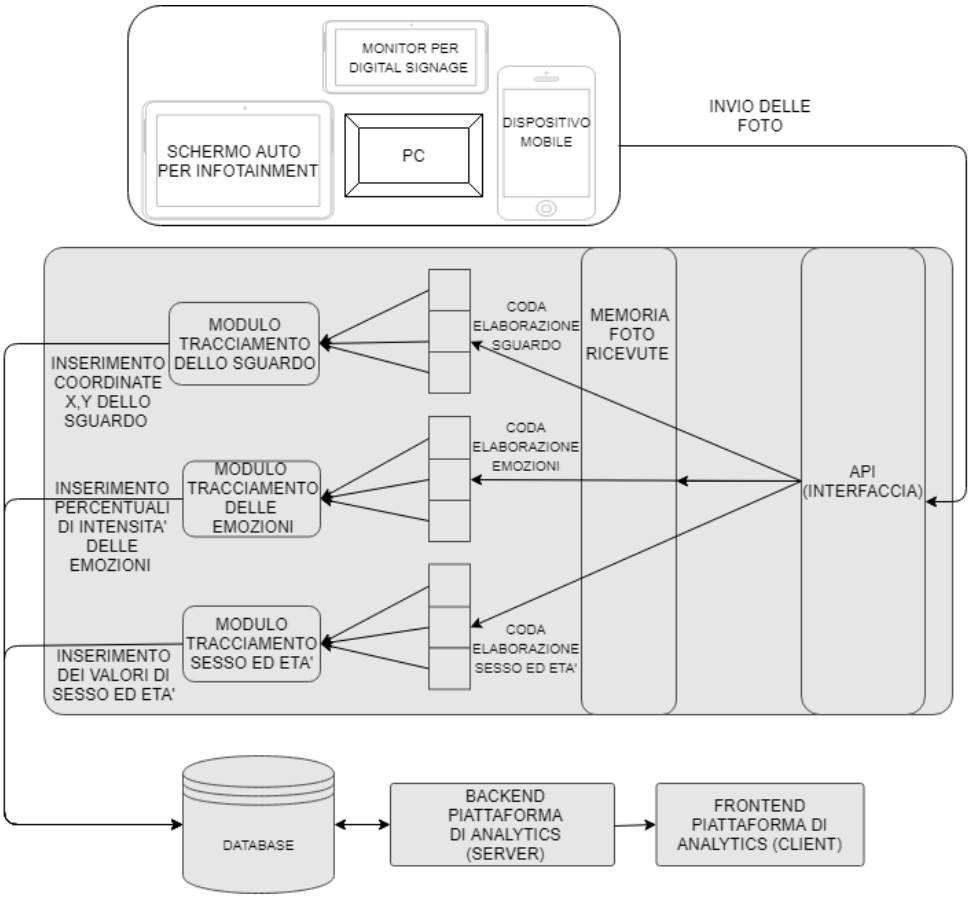

Figura 19: L'architettura finale della piattaforma per il settore Digitale

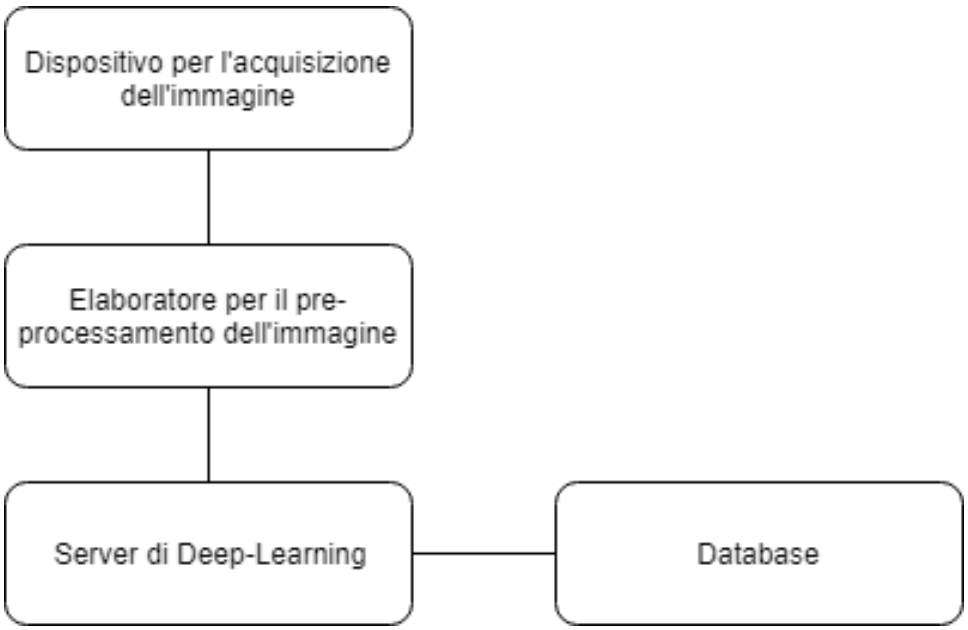

Figura 20: Il flusso dati

4.5 Le Reti Neurali Convoluzionali implementate

4.5.1 Il modulo di Face Detection

Il riconoscimento del volto (Face Detection) costituisce il vero collo di bottiglia di qualsiasi sistema di Deep Learning che si prefigge lo scopo di analizzare volti umani: come è già stato analizzato precedentemente infatti, ciò che le tre principali CNN ricevono in ingresso sono immagini molto leggere e composte da poche centinaia di pixel, che contengono solamente il volto della persona rilevata. Ciò che pesa realmente in fase di elaborazione, in termini di tempi di risposta del sistema e frequenza di elaborazione dei frame ricevuti in ingresso, è il riconoscimento del volto/volti all'interno dell'intero frame ricevuto dalla telecamera, che in determinati casi, soprattutto quando è necessario rilevare volti lontani anche circa 10 metri, possiedono una risoluzione molto elevata, come quella dei sistemi di acquisizione in 4K.
Nella piattaforma esaminata sono presenti due diverse implementazioni di Face Detection, che possono essere configurate per essere utilizzate in condizioni diverse.
Il primo modello di Face Detection permette di ottenere il riconoscimento di un volto posizionato frontalmente (entro un'angolazione massima di +-25 gradi lateralmente e verticalmente) rispetto alla telecamera ed è basato sull'istogramma del gradiente orientato (Histogram of oriented Gradient, HoG) e Support Vector Machine (SVM). Il toolkit software Dlib fornisce un modello di rilevazione frontale preallestito, ed addestrato su una dimensione minima dell'immagine del volto estratta di 80x80 pixel, il che risulta problematico quando si tratta di rilevare volti da una distanza maggiore di tre metri. Quando la dimensione del frame

che inquadra il volto è inferiore ad 80x80 pixel, quindi solitamente nel caso di volti lontani dalla fotocamera, viene utilizzato il modello Singleshot Multibox Detector (SSD), in grado di rilevare volti su varie scale. SSD non è sostituibile da HoG poiché il primo rileva ottimamente anche volti che non risultano totalmente frontali, tuttavia, questo causa problemi nell'accuratezza risultante dal rilevamento delle emozioni (per cui sono necessari in input immagini in cui sono ben visibili tutte le caratteristiche necessarie all'analisi delle espressioni facciali). Oltre a questo, i test affrontati mostrano che la velocità di rilevamento di HoG scende significativamente quando la risoluzione della telecamera è molto alta (4K o FHD), mentre ciò non influisce sulle prestazioni di SSD. Dato dunque che solo le facce frontali danno le migliori prestazioni in termini di accuratezza nel riconoscimento delle emozioni, HoG può essere utilizzato come filtro dopo il rilevamento del volto da parte di SSD all'interno del frame, così da scartare tutti i volti non frontali e al contempo mantenere prestazioni computazionali accettabili dato che ad HoG non verrà passato l'intero frame, ma solo la porzione di volto fornita in output da SSD.

4.5.2 La rete per il riconoscimento delle emozioni

Solitamente le ricerche associate al Deep Learning per il riconoscimeno delle emozioni si riferiscono a modelli addestrati utilizzando dataset costruiti in ambienti controllati, dove è possibile ottenere i migliori punteggi di accuratezza, o, addestrati con dati ottenuti dai crawler sul web, con bassa precisione ma che riflettono contesti reali. Ad oggi, non sono ancora stati valutati approcci che permettono di ottenere una buona accuratezza utilizzando i dati ottenuti "in the wild".
In questo contesto, è stato realizzato un approccio ibrido che cerca di ottenere una buona accuratezza per riconoscere le emozioni umane in molti contesti possibili: lo scopo principale del software implementato sarà quello di riconoscere le emozioni umane al di fuori del contesto in cui si trova l'utente, quindi in un'attività di retail utilizzando telecamere di sicurezza, oltre che davanti a uno smartphone.
Come già accennato, la rete in questione è stata implementata in Python basandosi sui framework Keras e Tensorflow, fondendo tre diversi dataset pubblici e provando e testando le più note architetture per CNN, come VGGG13 e VGG16.
Tutti i modelli di riconoscimento delle espressioni facciali hanno raggiunto precisioni diverse a seconda dei dataset su cui sono stati addestrati. In "Deep Facial Expression Recognition: A Survey", vengono elencate le diverse precisioni raggiunte dai più noti modelli: è stato osservato che tutti i modelli ad alta precisione vengono addestrati sui dataset generati in laboratorio come MMI e CK+. Tuttavia, i modelli addestrati con i dataset raccolti "in the wild", quindi in ambienti non controllati (di solito immagini di facce reperite in maniera automatizzata dal web tramite crawling) hanno precisioni inferiori, questo perché la maggior parte delle foto raccolte dal web hanno etichette imprecise.
Nel lavoro presentato, sono stati condotti degli esperimenti con un dataset costruito dal CK+ generato dal laboratorio e per questo di dimensioni ridotte, dalla versione ri-taggata del FER (il FER+), con una precisione di labeling superiore al 90% ma contenente solo circa 35000 immagini e da AffectNet, con oltre un milione di immagini di volti scansionati dal web e 450.000 immagini categorizzate da esperti umani. L'ipotesi è che combinando il dataset

altamente accurato generato in laboratorio con i dataset "in the wild", si può ottenere un modello con accuracy migliori per i benchmark in the wild.
Per addestrare CNN esistono diversi tipi di architetture a cui fare riferimento, il già citato framework per Python, Keras, fornisce un layer di frontend al framework Tensorflow sottostante, mettendo a disposizione degli sviluppatori diverse interfacce che semplificano l'implementazione di queste architetture; tra queste possiamo trovare: VGGG13, VGG16, VGG19 ed Inception. Per questo progetto è stato implementato e testato uno script di training per ognuna delle architetture sopra citate, in modo da vedere quale di esse restituisce i migliori risultati nel riconoscimento delle emozioni.
Come accennato, i dataset usati in fase di addestramento svolgono un ruolo cruciale nell'apprendimento supervisionato, i modelli di rete neurale dipendono in larga misura da essi. Per questo studio sono state esaminate tutte le immagini e implementato uno script per eliminare tutte le foto senza volti o con più di una faccia.
Un nuovo dataset è stato così costruito unendo le immagini filtrate di AffectNet, CK+ e FER+ utilizzando i tag felicità, sorpresa, tristezza, rabbia, disgusto, paura e neutralità. Questo dataset risultante comprende oltre 260.000 immagini, la distribuzione dei dati sulle diverse categorie è mostrata in Fig. 21.

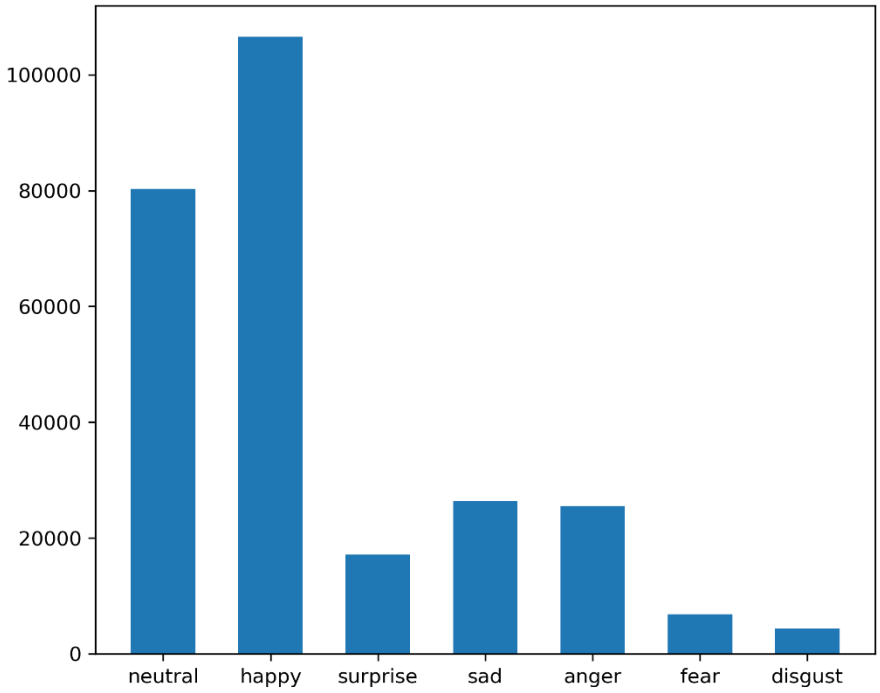

Figura 21: La distribuzione delle categorie di immagini del dataset

Per migliorare l'accuratezza del modello sono state utilizzate anche tecniche di allineamento facciale, come per le fasi di Face Detection precedentemente esposte. Dato un insieme di

punti di riferimento facciali (landmarks), l'allineatore facciale trasforma l'immagine in uno spazio di coordinate: tutti i volti dell'intero dataset in questo modo sono centrati rispetto al resto dell'immagine. Le immagini sono poi ruotate facendo in modo che gli occhi siano allineati ad un asse orizzontale e poi scalati in 64x64 pixel, così che le dimensioni di tutte le immagini siano approssimativamente identiche.
Per comprendere quale fosse la migliore architettura da utilizzare tra quelle disponibili in Keras, sono stati testati diversi modelli generati utilizzando VGG13, VGG16, VGG19, InceptionV2 ed InceptionV3, così da comparare le varie performance nel riconoscimento delle emozioni di ognuna di esse. I test di accuratezza dei vari modelli addestrati usando sempre gli stessi parametri di configurazione sono riportati nella tabella sottostante.

Architectures	**Accuracy (%)**
VGG13	**75.48**
VGG16	74.48
VGG19	73.14
InceptionV2	75.26
InceptionV3	67.20

Tabella 3: Le performance dei modelli generati tramite le diverse architetture

Come riscontrabile dalla tabella, il miglior guadagno di performance è stato raggiunto utilizzando l'architettura VGG13, mentre sotto sono riportate le accuracy ottenute in fase di addestramento e validazione del modello ottenuto, rappresentate come riferimento per gli esperimenti condotti.

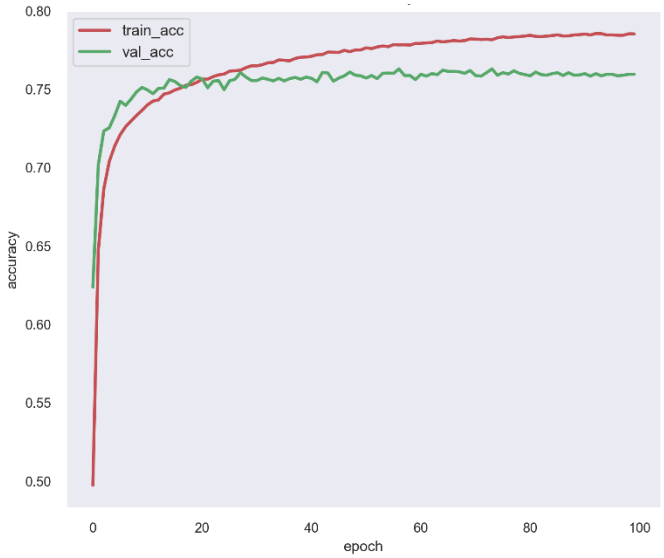

Figura 22: Accuracy di addestramento e validazione per la VGG13

L'accuratezza del test di ogni categoria di emozione, rappresentata nella matrice di confusione sottostante tramite mappa di calore, mostra degli interessanti risultati. Le immagini con tag relativi alla paura sono stati classificati erroneamente come sorpresa, mentre i tag relativi al disgusto sono stati confusi con la rabbia con una percentuale di più del 20%. Di fatto, queste emozioni sono difficili da differenziare tramite le espressioni facciali, anche per un essere umano.

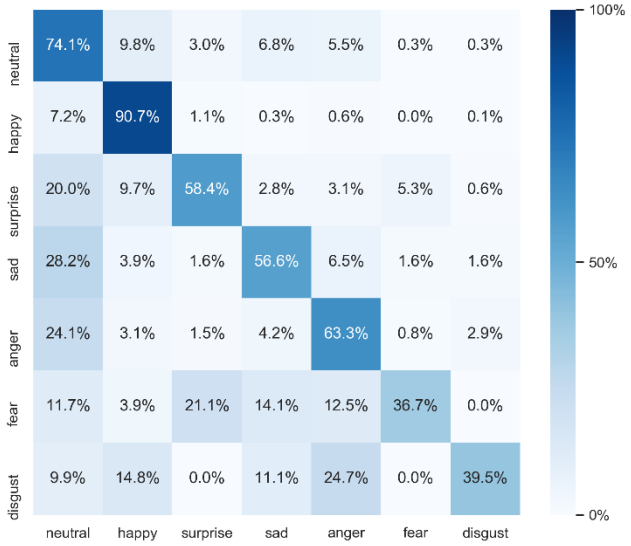

Figura 23: la matrice di confusione normalizzata

Il modello addestrato tramite l'architettura VGG13 è stato successivamente valutato sulla base del dataset EmotioNet relativo alla challenge del 2018: l'Ohio State University, sul proprio sito web, ha infatti reso disponibile il loro dataset per dare a chiunque l'opportunità di comparare i propri risultati con quelli ottenuti durante le challenge del 2017 e del 2018; la tabella sottostante mostra i risultati della valutazione, con Accuracy ed F1 Score, che tiene conto di precisione e recupero dei test tramite la formula di media armonica sottostante:

$$F_1 = \frac{2}{\frac{1}{r} + \frac{1}{p}} = 2 \cdot \frac{p \cdot r}{p + r}.$$

Come preannunciato, è possibile notare che i risultati peggiori sono raggiunti per il disgusto a causa della scarsità di immagini nel dataset di riferimento usato per la valutazione, mentre analogamente per la paura il risultato non è stato neppure riportato a causa della poca significatività dei punteggi riscontrati.

Categories	Accuracy	F1
happy	**0.9770**	**0.9799**
anger	0.75	0.1198
disgust	0.0128	0.0099
sad	0.5955	0.2888
surprise	0.7059	0.3944

Tabella 4: Accuracy e F1 Score risultanti dalla valutazione

4.5.3 La rete per il riconoscimento dello sguardo su laptop

La rete implementata in questo progetto adotta un'architettura simile a quella proposta da Krafka et al. che si basa sul modello AlexNet (Francia et al., 2018) ossia iTracker, la rete presa a modello nel paragrafo 4.3 per il riconoscimento dello sguardo in ambienti mobile e più specificatamente per iPhone.
Scopo di questa implementazione è stato dunque quello di prendere iTracker come modello per sviluppare qualcosa che si adattasse anche al mondo del web, quindi a foto acquisite tramite webcam stando di fronte ad un laptop o pc.
Questo modello è stato implementato e addestrato in Python, usando il già citato framework TensorFlow e CUDA per supportare l'accelerazione su GPU.
Gli input di questa CNN sono l'immagine del volto ritagliato (dimensioni 224x224 pixel), le immagini dei due occhi ritagliati (dimensioni 224x224 pixel) e un vettore di griglia facciale 1x4, che esprime la porzione dell'intera immagine occupata dal volto. L'output è un vettore 2D contenente le coordinate x e y del punto stimato dello schermo (in centimetri).
I volti e gli occhi vengono rilevati e ritagliati rispettivamente tramite il motore di Face Detetion frontale (HoG) e il predittore di 68 landmark di Dlib. Quando viene trovata una casella con un volto, le coordinate dello schermo del suo angolo in alto a sinistra e la sua larghezza e altezza sono memorizzate nel vettore della griglia facciale; in questo modo è possibile recuperare informazioni implicite sulla posa tenuta dalla testa dell'utente relativamente allo schermo.

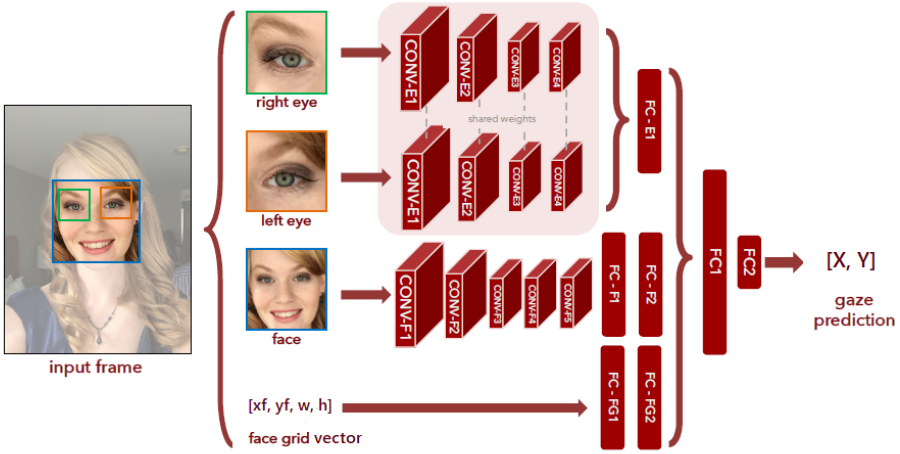

Figura 24: Rappresentazione schematica dell'architettura della CNN

Nella rappresentazione soprastante CONV rappresenta i layer convoluzionali, con filtro dimensione/numero di kernel: CONV-E1, CONV-F1: 11 X 11/96, CONV-E2, CONV-F2: 5 X 5/256, CONV-E3, CONV-F3: 3 X 3/384, CONV-E4, CONV-F4: 1 X 1/64, mentre FC rappresenta layer Fully Connected (con dimensioni: FC-E1: 128, FC-F1: 128, FC-F2: 64, FC-FG1: 256, FC-FG2: 128, FC1: 128, FC2: 2). E' stato poi aggiunto un layer di droput a destra di ognuno degli ultimi layer convoluzionali.
Gli elementi del vettore per la face grid (la griglia facciale) rappresentano, in pixel: xf (coordinata x del punto in alto a sinistra della face box, il riquadro del volto), yf (coordinata y del punto in alto a sinistra della face box), w (larghezza della face box), h (altezza della face box). Gli output riguardano le distanze, in cm dalla telecamera.
Successivamente è stato sviluppato un dataset per l'addestramento, raccogliendo immagini da volontari informati che hanno deciso di contribuire. Ad ogni partecipante è stato chiesto di fissare un punto che compariva casualmente in 30 diverse posizioni dello schermo, mentre la webcam scattava una foto per ciascuna di esse; sono state memorizzate anche le coordinate dei punti corrispondenti, per garantire l'associazione delle coordinate dell'immagine sullo schermo.
Per quanto riguarda le dimensioni dello schermo, lo scopo era che questa applicazione cercasse di coprire più casistiche possibili (cioè fosse in grado di gestire diversi tipi di display). Tuttavia, per limitazioni tecniche dovute all'utilizzo di linguaggi lato client come Javascript per reperire i dati necessari, non è possibile recuperare le dimensioni fisiche di uno schermo da un'applicazione web. Per questo motivo, nell'interfaccia mostrata in figura 25, all'utente viene chiesto di scegliere la dimensione del proprio schermo, tra quelle visualizzate in un elenco a tendina (da 13 a 30 pollici). Conoscendo le dimensioni fisiche dello schermo e la risoluzione di visualizzazione (che invece può essere determinata automaticamente), è ovviamente possibile esprimere le coordinate dei punti, nativamente in pixel, in cm.

Tale applicazione è stata sviluppata all'interno dell'ambiente Unity e poi portata in una pagina web, ospitata da un server remoto Amazon, in cui sono state memorizzate immagini e dati.

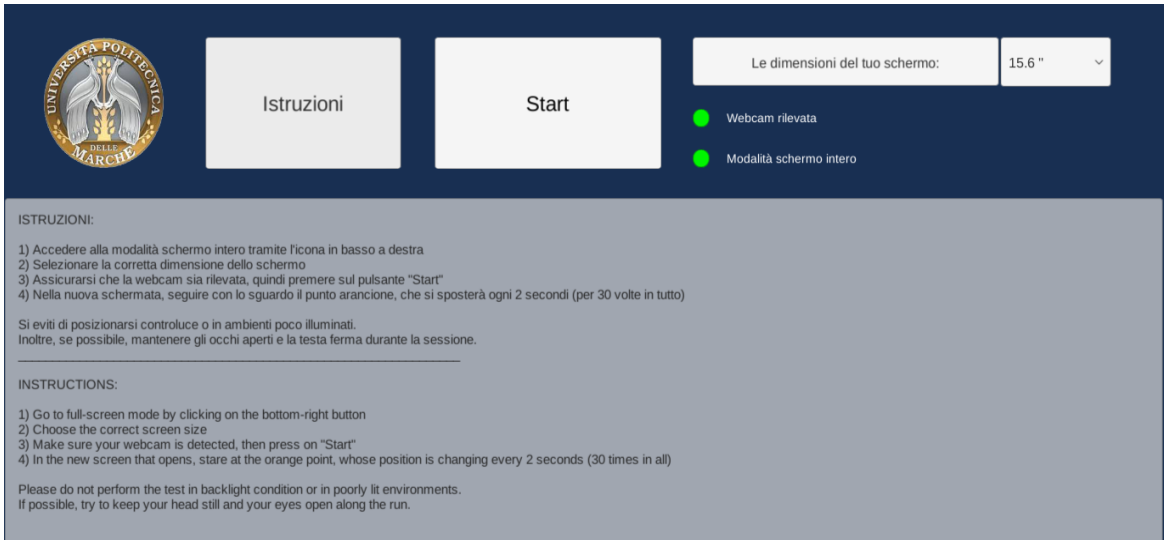

Figura 25: L'interfaccia di avvio del software per la raccolta di immagini

La funzione di errore usata riguarda semplicemente la distanza euclidea 2D tra il punto previsto e quello effettivamente visualizzato:

$$Error = \sqrt{(x_{pred} - \hat{x})^2 + (y_{pred} - \hat{y})^2}$$

dove x_pred e y_pred sono le coordinate del punto previsto, mentre xˆ e e yˆ sono quelle effettivamente visualizzate.
Adadelta è stato scelto tra gli ottimizzatori implementati in Keras.
Per poter testare questa rete è stato realizzato uno script in Python che visualizzasse il percorso di scansione degli occhi su un'immagine visualizzata; all'avvio viene visualizzata a schermo intero un'immagine preimpostata, mentre la webcam scatta ripetutamente foto dell'utente e alimenta la rete con esse. Tutte le coordinate vengono memorizzate ed eventualmente, quando l'utente chiude l'immagine, vengono automaticamente elaborate in modo che i punti vicini siano raggruppati insieme, separando così le fissazioni dalle saccadi. Alla fine, i cluster colorati sono rappresentati sopra l'immagine visualizzata, insieme alle linee rette che collegano quelle successive, per dare un'idea del percorso che gli occhi hanno seguito nel complesso.

Capitolo 5.

Casi studio

Una volta implementata la tecnologia questa è stata applicata e testata in diverse tipologie di casi studio, così da dimostrarne l'efficacia nel raggiungere gli scopi che ci si era prefissati, in particolare l'analisi dei clienti/utenti in prossimità dei touchpoint in maniera automatizzata e al contempo accurata.

5.1 L'installazione del sistema in un piccolo negozio di abbigliamento

Per determinare l'efficacia del sistema di rilevazione di sesso, età ed emozioni dei clienti rispetto ad una tradizionale analisi video, è stato condotto un primo caso studio all'interno di un piccolo negozio di abbigliamento uomo/donna durato tre giorni interi.
Il sistema è stato dunque correttamente integrato nel negozio per monitorare la CX corrispondente ai seguenti tre touchpoint chiave:

- Touchpoint 1: ingresso del cliente nel negozio;
- Touchpoint 2: prova dell'abito davanti allo specchio;
- Touchpoint 3: il cliente sosta alla cassa per il check out.

La configurazione del sistema implementato (Figura 26) è costituita da:

- Una IpCamera Foscam R2 1080p posizionata su una staffa ad un'altezza di 2,10m, sull'angolo frontale destro all'ingresso e dietro il bancone;
- Una webcam Logitech Quickcam Pro 9000 e una webcam Logitech Brio 4k, posizionate su uno scaffale ad un'altezza di 1,70 m, a sinistra dello specchio. Tali videocamere sono orientate rispettivamente per inquadrare correttamente i clienti quando rivolgono il volto e le spalle verso lo specchio;
- Un Asus VivoMini UN62 con processore Intel Core i5-4210U, Intel HD Graphics 4400, RAM 4 GB, Dual Channel, DDR3L a 1600MHz con Linux Ubuntu 14.01 posizionato sullo stesso ripiano che ospita le due webcam USB;
- Un router Sitecom N300 per consentire la connessione Wi-Fi tra la IpCamera e l'Asus VivoMini.

Figura 26: La configurazione del sistema

Sono stati coinvolti nel test un totale di 30 clienti (15 maschi, 15 femmine, equamente distribuiti tra i 3 gruppi di età: 24-34, 35-44, 45-50 anni), scelti a caso tra quelli con la carta fedeltà del negozio. Essi sono stati precedentemente informati sugli obiettivi e sulle modalità dello studio e hanno dato il consenso informato per il trattamento dei loro dati personali e sensibili, per poi venir premiati con buoni sconto di 10€. I test sono stati effettuati in tre normali giorni di chiusura. Ai clienti selezionati è stato chiesto di visitare il negozio in un giorno scelto liberamente tra le date disponibili. Prima dei test, è stato chiesto ai clienti di scegliere e provare almeno un vestito ma al di là di questo, sono stati liberi di comportarsi come meglio credevano e quindi anche acquistare o no ciò che preferivano. In ogni caso, prima di lasciare il negozio questi si sono dovuti recare alla cassa per ricevere il voucher.
Un esperto di psicologia comportamentale ha effettuato un'analisi sui video registrati, al fine di verificare la correttezza dei dati di output forniti dal sistema. In particolare, per il Touchpoint 1 sono stati considerati i primi 30 secondi di permanenza dei clienti nel negozio. I momenti in cui il cliente si trova di fronte allo specchio sono relativi al secondo touchpoint, mentre i tempi che il cliente ha trascorso alla cassa per ricevere il voucher ed eventualmente pagare sono relativi al terzo.
L'analisi video è stata effettuata analizzando sessioni video corrispondenti ai touchpoint considerati. In totale sono stati identificati 1116 spezzoni video (217 relativi al Touchpoint 1, 551 al Touchpoint 2 e 348 al Touchpoint 3), corrispondenti ai momenti in cui venivano rilevati dei volti umani e questi mostravano delle espressioni facciali riconoscibili dall'occhio umano. A ciascuno di essi sono stati associati:

- Una delle principali emozioni di Ekman (o uno stato emotivo "neutro/non rivelato");
- Sesso e fascia d'età del cliente inquadrato dal frame video.

I risultati sono stati confrontati con i dati forniti dal sistema di riconoscimento installato: nel complesso, esso è stato in grado di rilevare il volto nel 96% degli spezzoni video considerati registrati dalla Logitech Quickcam e nel 94% di quelli registrati dalla webcam

della Logitech, mentre i volti sono stati rilevati solo nell'82% dei frame registrati dalla IpCamera. Ciò è dovuto principalmente al fatto che la IpCamera, posizionata vicino all'ingresso del negozio, ha risentito dei cambiamenti di luce ambientale, per cui diversi frame risultano sovraesposti.

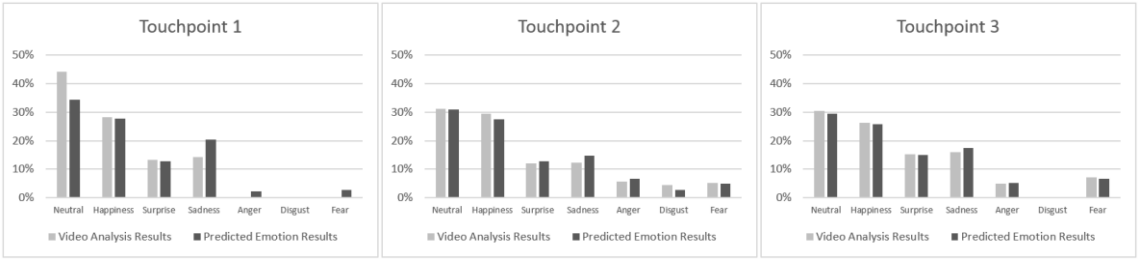

Figura 27: Confronto tra le emozioni rilevate dal sistema e quelle reali

La figura 27 permette di confrontare le emozioni identificate attraverso l'analisi video nei diversi touchpoint, con quelle previste dal sistema. Come si può osservare, gli stati emozionali che più caratterizzavano le CX erano quelli neutrali, di felicità, sorpresa e tristezza, mentre pochissime espressioni facciali sono state relazionate a stati di rabbia, disgusto o paura. La Figura 28 fornisce i risultati della matrice di confusione normalizzata del sistema nella classificazione delle emozioni durante il test.

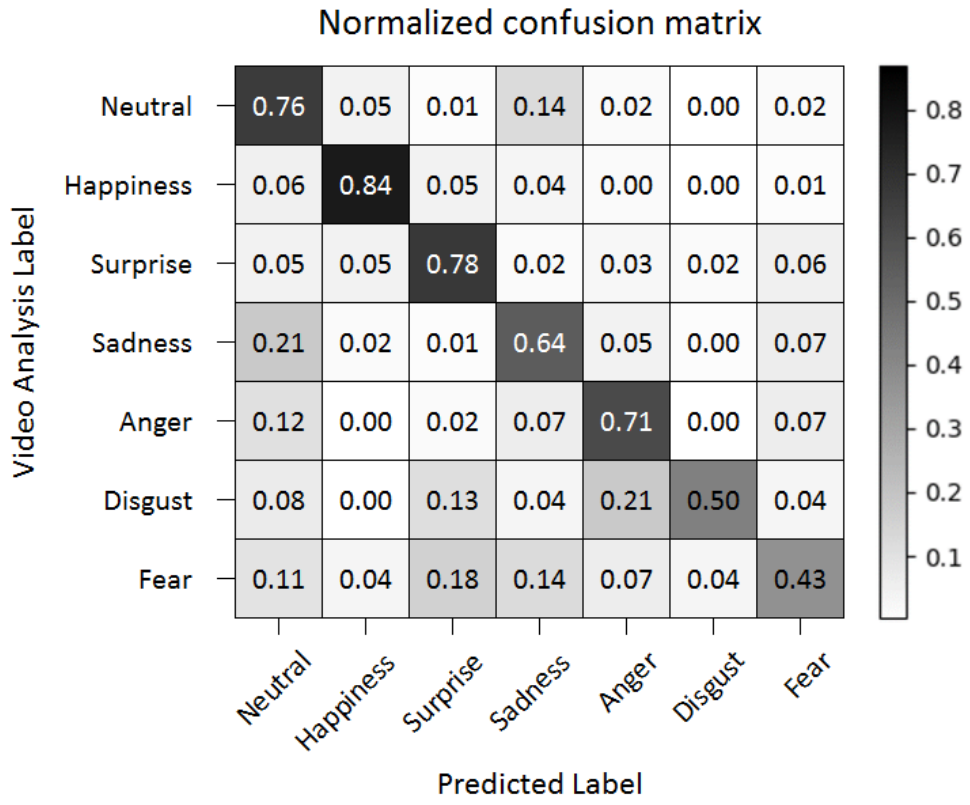

Figura 28: La matrice di confusione normalizzata dell'esperimento

Come è possibile osservare, il sistema è in grado di prevedere efficacemente la felicità e la sorpresa, tuttavia, è possibile notare diversi errori di classificazione comuni. Ad esempio, il sistema ha predetto in diversi casi "neutrale" invece di "tristezza" e viceversa, "rabbia" invece di "disgusto" e "sorpresa" o "tristezza" invece di "paura".

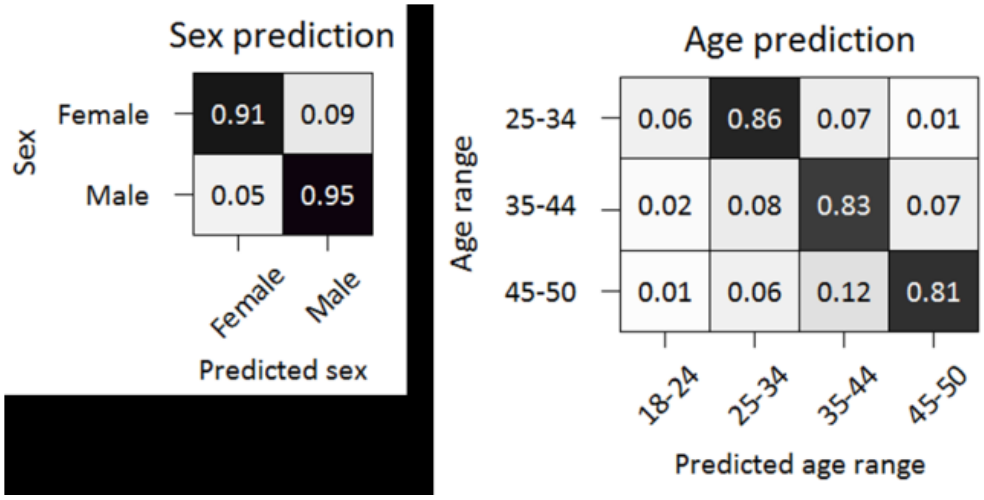

Figura 29: Le matrici di confusione per sesso ed età

I risultati delle attività di classificazione del sistema per sesso e per fasce di età sono riportati nella figura 29. Come si può osservare, il sistema è in grado di prevedere efficacemente il sesso del cliente (l'accuratezza complessiva è pari a 0,92), nonostante sia stato registrato un maggior numero di errori di classificazione per le donne. Per quanto riguarda la fascia di età, l'accuratezza della previsione risulta pari a 0,83. Come mostrato in figura 29 dunque, il sistema è più affidabile nel prevedere l'età delle persone di età compresa tra i 25 e i 34 anni, mentre è meno affidabile nel prevedere l'età delle persone appartenenti alla fascia di età 45-50 anni: ciò è probabilmente dovuto alle caratteristiche dei dataset di training utilizzati in questa fase.

5.2 L'implementazione di una Bayesian Belief Network per predire il comportamento del cliente in un negozio di abbigliamento

Anche in questo caso è stato studiato il comportamento del sistema di rilevamento di emozioni, sesso ed età, questa volta però coadiuvato da un sistema predittivo basato su reti Bayesiane, che implementano il noto teorema di Bayes che stabilisce la correlazione di probabilità tra due eventi A e B secondo la formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Il modello proposto ha lo scopo di prevedere la personalità del cliente in base al suo comportamento in negozio, al fine di fornire indicazioni sulle aspettative del cliente.
A tal fine, sono state considerate quattro personalità del cliente medio secondo i quattro stereotipi definiti Allesandra et al., nel 1998, e riportati nella Tabella 5.

Categoria	Personalità	Aspettative
Director	Persone che hanno ben chiaro cosa vogliono, prendono decisioni il più rapidamente possibile e non sono particolarmente interessati ad altri pareri ed opinioni al di fuori del loro	Non hanno bisogno/vogliono aiuto da parte dei commessi, a meno che non siano loro stessi a richiederlo
Thinker	Persone analitiche che ricercano ed analizzano tutte le possibilità prima di prendere una decisione.	Non vengono particolarmente influenzati dai piccoli consigli forniti dai commessi durante le loro visite in negozio
Relator	Persone che sentono il forte bisogno di far parte di un gruppo, facendo molto affidamento all'opinione altrui.	Prendono decisioni molto lentamente e solitamente desiderano includere altre persone nel loro processo decisionale
Socializer	Persone che amano parlare e fare nuove amicizie	Vogliono innanzitutto socializzare con i commessi e sono molto concentrati su loro stessi

Tabella 5: Le categorie di personalità dei compratori

La definizione della rete bayesiana si svolge in tre fasi:
- Raccolta dati;
- Costruzione e compilazione della rete sulla base dei risultati dell'analisi dei dati;
- Definizione di relazioni probabilistiche tra le variabili definite in base al set di dati.

Per conoscere il comportamento del cliente in un negozio e osservare le diverse strategie adottate da venditori esperti, è stata effettuata un'analisi video nell'ambito di un negozio di moda monomarca caratterizzato da prodotti di fascia media e target di clientela tra i 18 e i 30 anni.
Per gestire le BBN è stato utilizzato il software di simulazione Netica, prodotto da Norsys Software Corporation.

5.2.1 Raccolta dati

Due venditori (una donna e un uomo) con più di 5 anni di esperienza sono stati coinvolti nello studio: sono stati precedentemente informati sui metodi e gli obiettivi e hanno dato il loro

consenso informato. Per evitare che il comportamento dei clienti fosse influenzato dalle condizioni sperimentali, i clienti sono stati reclutati all'uscita dal negozio: sono stati informati sui metodi e gli scopi dello studio e sono stati invitati a dare il loro permesso di analizzare i relativi video di sicurezza firmando il consenso informato, ed è stato chiesto loro di rispondere al test di personalità tratto dal survey di Littauer (Littauer, 1922). Lo studio ha coinvolto un totale di 60 clienti (30 uomini e 30 donne). Le osservazioni sono state fatte in otto giorni.
Attraverso l'analisi video da parte di un esperto marketing e uno di psicologia cognitiva, sono stati raccolti i seguenti dati:

- Genere ed età: sono stati considerati due gruppi di età, ≤24 e >24; è stata riscontrata una percentuale maggiore di donne che rispondono alla personalità Relator rispetto agli uomini.
- Comportamento decisionale: sono stati considerati i primi 90 secondi di permanenza del cliente nel negozio. Analizzando i risultati dell'osservazione, sono stati identificati due principali comportamenti dei clienti: *Straight to the goal* e *Wandering*. La tipologia di cliente "dritto all'obiettivo" si muove all'interno di un piccolo numero di aree del negozio e trascorre in ogni area più di 30 secondi (ad esempio, un cliente che va direttamente verso un particolare scaffale e osserva un particolare prodotto). Il cliente si comporta invece in maniera "vagante" se si muove in più di 2 aree e rimane meno di 30 secondi in ognuna di esse (ad esempio, un cliente che gira intorno al negozio senza focalizzarsi su un particolare prodotto).
- Interazione con il commesso: sono stati considerati tre livelli di interazione. Un livello di interazione è "Alto" se il cliente rimane più di 5 minuti vicino al commesso, "Medio" se il tempo rilevato è compreso tra 1 e 5 minuti e "Basso" nel caso in cui il tempo totale è inferiore a 60 secondi. Le interazioni tra cliente e il commesso alla cassa non vengono prese in considerazione.
- Personalità: è la variabile dipendente di questo studio. Mentre gli altri dati vengono raccolti attraverso il SW di riconoscimento facciale, la personalità del cliente viene analizzata attraverso un questionario.

5.2.2 Sviluppo della rete Bayesiana

La BBN è stata progettata sulla base delle conoscenze degli esperti e dell'analisi dei dati precedentemente raccolti (con analisi di correlazione e test di multicollinearità). Il diagramma casuale risultante è riportato in Fig. 30

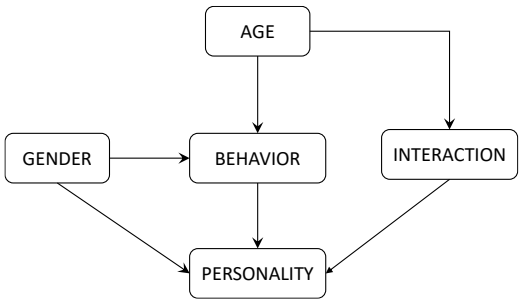

Figura 30: L'architettura della rete bayesiana

Per esaminare le relazioni tra le caratteristiche del cliente sono stati utilizzati dei test di indipendenza Chi-Square. I risultati mostrano che esiste una relazione significativa tra genere e personalità del cliente, $\chi 2$ (3, N = 50) = 9.513, p = .023. Inoltre, i risultati mostrano che esiste una relazione significativa tra la personalità e il comportamento del cliente, $\chi 2$ (3, N = 50) = 9.669, p = .022. Le persone che rispondono alle personalità di Director o del Relator di solito si comportano "Straight to the goal", mentre le persone "thinker" sono più abituate a vagare per il negozio, quindi "wandering".

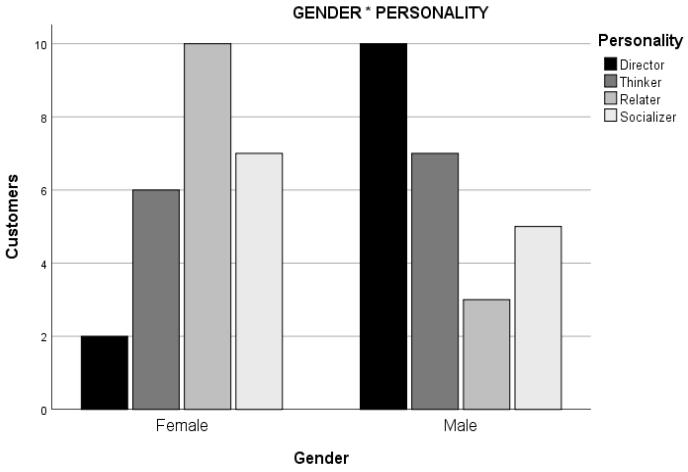

Figura 31: La correlazione tra sesso e personalità del cliente

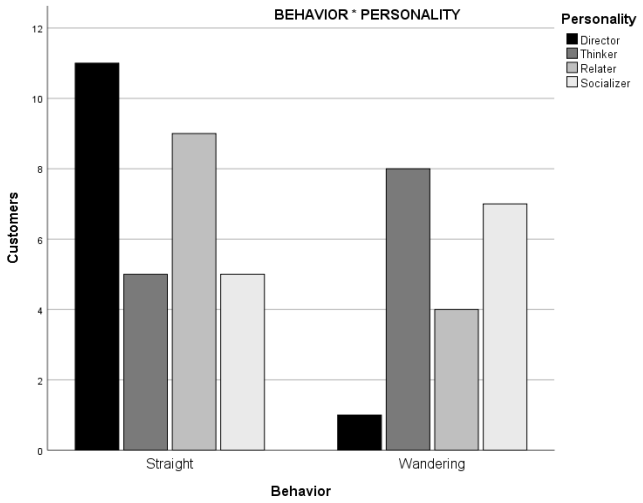

Figura 32: La correlazione tra comportamento e personalità del cliente

Un'altra caratteristica legata alla personalità è il livello di interazioni con il commesso, $\chi 2$ (6, N = 50) = 35.648, p < .001. Questa tendenza è evidente negli studi di Alessandra et al., in quanto le personalità Director difficilmente preferiranno interagire con gli assistenti (sanno già cosa vogliono) mentre Socializer e Relater raramente perderanno l'opportunità di chiedere consigli o di parlare. Al contrario non c'è relazione tra genere e comportamento del cliente, $\chi 2$ (1, N = 50) = 3.000, p = .083.

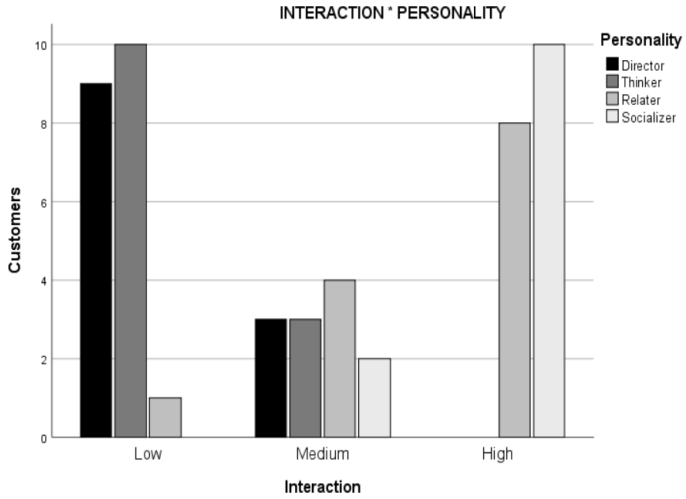

Figura 33: La correlazione tra interazioni col commesso e personalità del cliente

5.2.3 Definizione delle relazioni

Le conoscenze fornite dalla BBN possono essere utilizzate per fornire indicazioni per gestire il servizio clienti all'interno del negozio.
In questo caso studio, ci si è basati sul comportamento dei commessi considerando due principali fasi temporali: *Approach* e *Sell* (Vada et al., 2006). Per quanto riguarda la fase di approccio, possiamo supporre di avere due comportamenti principali per l'assistente nel negozio: *Waiting* (ad esempio, l'assistente del negozio saluta il cliente e poi aspetta che il cliente stesso chieda direttamente aiuto) e *Supporting* (l'assistente di negozio saluta il cliente e lo avvicina per offrire aiuto). Allo stesso modo, nella fase di vendita, si possono identificare due comportamenti principali: *Answering* (cioè, l'assistente si limita a soddisfare le richieste esplicite del cliente, come ad esempio la taglia richiesta, la posizione di un particolare prodotto, ecc.) e *Proposing* (l'assistente di negozio cerca di rispondere alle richieste esplicite dei clienti, tentando però di vendere anche altro rispetto a ciò che è specificamente richiesto). Sulla base delle previsioni della BBN, è quindi possibile definire una vera e propria CX Strategy.
Ad esempio, la probabilità condizionata per la personalità *director* è molto bassa nel caso delle donne più giovani (Fig. 34). In questo caso, potrebbe essere conveniente che i commessi adottino un comportamento di supporto-proposta.

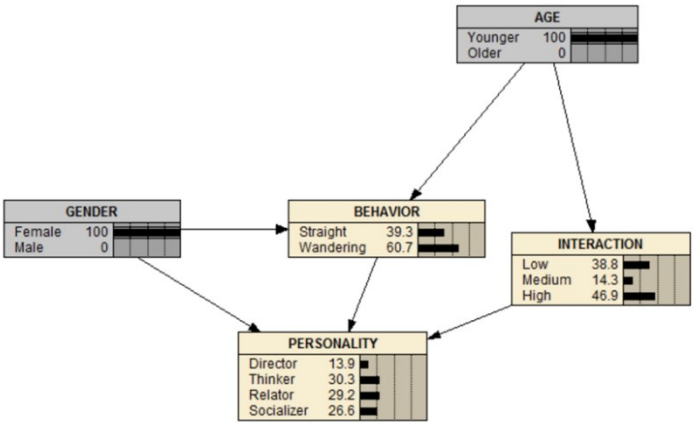

Figura 34: Predizione della BBN in caso di donne sotto i 24 anni

Quando è invece un uomo ad entrare nel negozio, è molto importante osservare il suo comportamento: se va "dritto verso l'obiettivo" (Fig. 35) la probabilità che corrisponda al profilo di un director è alta, per cui il commesso dovrebbe aspettare che chieda per un aiuto prima di intervenire.

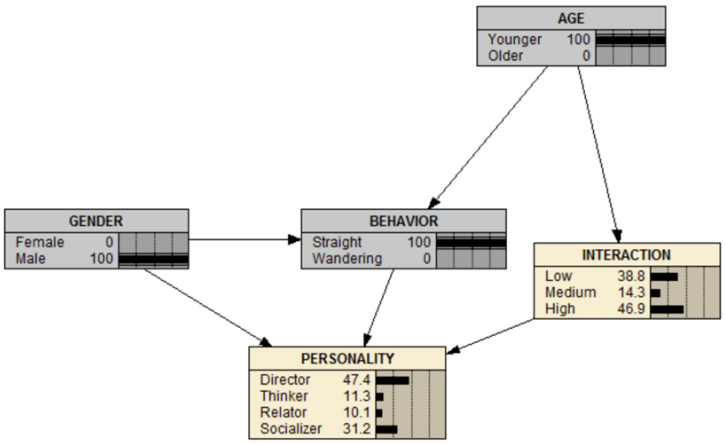

Figura 35: Predizione della BBN in caso di donne sopra i 24 anni e con comportamento "straight"

Al contrario, se quest'ultima tipologia di soggetto preferisce "vagare" per il negozio (Fig. 36), la probabilità che sia un "Thinker" è alta, per cui in questo caso i commessi dovrebbero immediatamente avvicinarsi e sostenerlo.

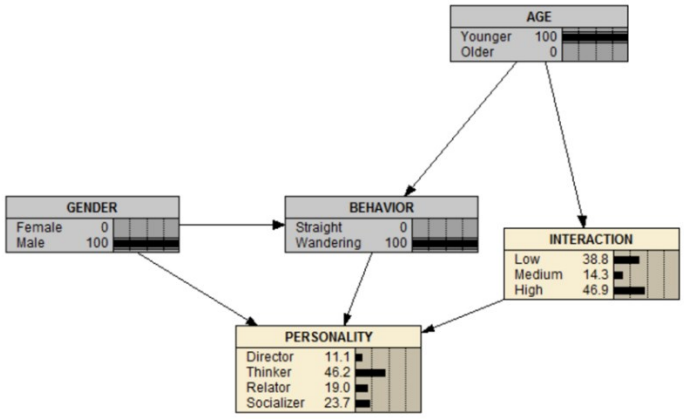

Figura 36: Predizione della BBN in caso di donne sotto i 24 anni e con comportamento "wandering"

Per convalidare la CX Strategy definita, sono state utilizzate le emozioni del cliente fornite dalla tecnologia Deep Learning implementata. Lo scopo principale dell'utilizzo di questo strumento è quello di avere un sistema di feedback in tempo reale, in grado di misurare l'efficacia della CX strategy definita e permettere di correggerla in modo retroattivo. A tal fine, viene fornito il calcolo di un valore medio ponderato (che va da -100 a 100) basato sulla positività o negatività delle emozioni registrate. La felicità e la sorpresa sono considerate emozioni positive e tutte le altre (cioè paura, disgusto, rabbia e tristezza) come emozioni negative. In questo modo si definiscono due soglie principali, che dividono i valori continui

in tre range: positivo, negativo e neutro. Quando il valore medio emotivo rimane in un range da -15 a 15 la situazione rimane stabile e il cliente è per lo più in uno stato d'animo neutro. Se il cliente ha uno stato d'animo positivo (valore compreso tra 15 e 100), o negativo (valore compreso tra -15 e -100), la CX Strategy può essere considerata rispettivamente efficace o dannosa e quindi adeguata di conseguenza.

5.2.4 Risultati dei test

Dopo aver implementato il sistema sono state ricreate le medesime condizioni di test indicate nel paragrafo 5.2.1 per validare la piattaforma, che questa volta include uno script sviluppato in Python che sfrutta per il processo decisionale le soglie definite tramite le reti bayesiane sopra esposte: l'output di questo script è la percentuale di probabilità che un cliente corrisponda o meno ad un particolare profilo. Nel test, l'adottare uno dei comportamenti di assistenza e vendita sopra esposti è poi stato a discrezione del commesso, che dunque aveva come unico scopo quello di testare o meno l'efficacia della predizione della personalità del cliente. Essendo la valutazione degli output della piattaforma affidata totalmente all'esperienza del commesso è stato complicato fornire dei dati oggettivi a supporto dell'efficacia della piattaforma, senza che tali risultati siano potuti essere supportati da un effettivo confronto con i dati di vendita (non accessibili) in seguito all'applicazione della tecnologia, ma il riscontro è stato positivo, e grossolanamente è stato riferito un successo nella vendita in circa il 78% dei casi.

5.3 L'applicazione del motore di riconoscimento delle emozioni per analizzare il gradimento del pubblico durante una serata di opera

Questo caso studio è stato portato avanti in collaborazione con lo spin-off dell'Università Politecnica delle Marche Emoj e l'Arena Sferisterio di Macerata: scopo del progetto era quello di monitorare parte dell'audience all'Opera Festival dell'estate 2019 analizzandone lo stato emotivo al fine di misurare il livello di soddisfazione e apprezzamento così da strutturare un'offerta capace di attrarre meglio gli spettatori nelle prossime edizioni del festival.
Sebbene all'apparenza questo caso studio non riguardi specificatamente l'analisi della Customer Experience durante un processo di interazione cliente/brand nel retail, in realtà analizzare il livello di gradimento del pubblico durante delle serate di opera è stato un fondamentale banco di prova per testare la tecnologia di riconoscimento del volto e delle emozioni in condizioni ambientali avverse (gli spettacoli avvenivano totalmente al buio e all'aperto), inoltre, lo spettacolo stesso costituisce un fondamentale touchpoint tra i clienti che hanno acquistato il biglietto e tutti i brand coinvolti nell'organizzazione delle serate: di fatto lo spettacolo stesso è un prodotto/servizio venduto ad un cliente.

Per questo scopo è stato monitorato un campione di dodici persone tra il pubblico durante le dieci rappresentazioni della Carmen, Macbeth e Rigoletto, al fine di analizzare gli aspetti emotivi della visione degli spettacoli tramite la tecnologia oggetto dei capitoli precedenti.

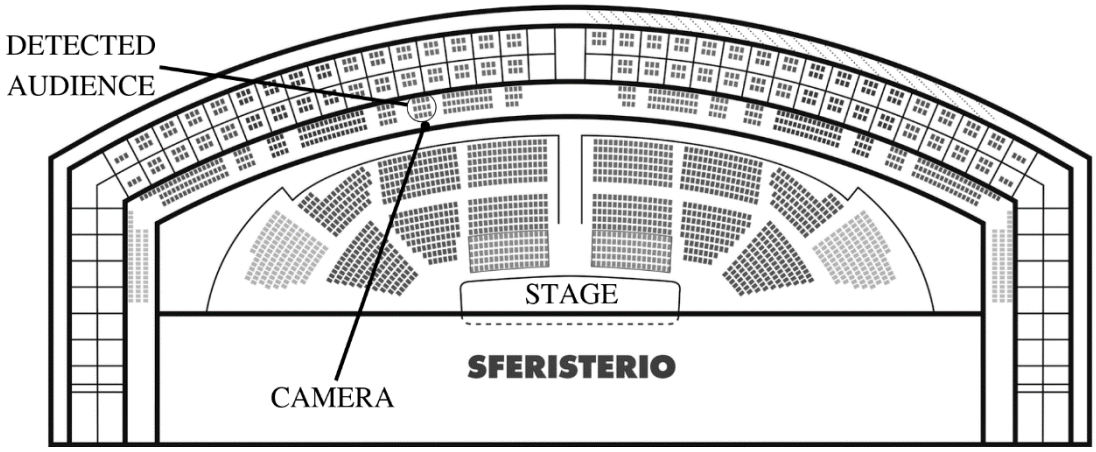

Figura 37: Mappa dello Sferisterio e area monitorata

I dati monitorati relative alle reazioni del pubblico sono mappati con la durata di ogni atto lirico, di ogni aria e con le caratteristiche dell'opera. È quindi possibile comprendere l'impatto reale delle scelte dell'art director e del direttore d'orchestra sul pubblico e il livello della qualità percepita. I dati raccolti dal monitoraggio delle dodici persone sono aggregati per ogni opera e per ogni tipo di spettacolo (cioè Macbeth, Carmen e Rigoletto).
La selezione del pubblico da monitorare non è stato scelto casualmente; i singoli individui sono stati selezionati per corrispondere ai cinque buyer personas dello Sferisterio, trovati grazie ai sondaggi presentati nelle ultime due stagioni del Macerata Opera Festival e alle ricerche di mercato sui teatri all'aperto.
I risultati dell'analisi emozionale del pubblico sono il punto centrale dello studio sull'esperienza vissuta dai clienti dello Sferisterio, fornendo indicazioni precise su ogni momento di tutti gli spettacoli, al fine di confrontare l'apprezzamento e le emozioni suscitate da ogni evento che si svolge sul palco.
La rilevazione delle emozioni rappresenta l'indagine qualitativa, mentre quella quantitativa è stata portata avanti tramite un'indagine di pubblico per ogni spettacolo e un'altra indagine al botteghino (nel momento dell'acquisto), per indagare il livello di apprezzamento di un maggior numero di persone tra il pubblico.
L'integrazione dei risultati delle due analisi ha permesso all'Associazione Arena Sferisterio di capire in quali elementi poter migliorare la propria offerta, sia sotto l'aspetto prettamente relativo alle rappresentazioni, che in tutti gli altri elementi che possono portare ad un aumento della qualità percepita.
Il monitoraggio si è svolto durante le fasce orarie tra le nove e mezzanotte, quindi in condizioni di totale assenza di luce naturale e contemporaneamente artificiale, dato che l'opera va seguita quasi totalmente al buio, con le luci di scena che, assieme alle stelle, rappresentano l'unica fonte luminosa durante gli spettacoli. In queste condizioni si è dunque

reso necessario utilizzare una camera ad infrarossi, montata su una delle ringhiere di fronte al pubblico e collegata ad un pc posizionato in una stanzetta all'interno dell'arena tramite connessione wifi, così da evitare cablaggi problematici. Nello specifico, la configurazione tecnica adottata è stata la seguente:

- Una IpCamera Foscam R2 1080p posizionata su una ringhiera in posizione perfettamente frontale rispetto al campione monitorato, ad un'altezza da terra di un metro e distante circa 3 m in linea d'aria dal più distante soggetto da monitorare tra i dodici seduti in gradinata;
- Un Asus VivoMini UN62 con processore Intel Core i5-4210U, Intel HD Graphics 4400, RAM 4 GB, Dual Channel, DDR3L a 1600MHz con Linux Ubuntu 14.01;
- Un router Sitecom N300 per consentire la connessione Wi-Fi tra la IpCamera e l'Asus VivoMini.

Questo test è servito anche per mostrare il funzionamento della piattaforma anche con camere ad infrarossi e quindi con immagini in input in bianco e nero: questo non ha causato particolari problemi alla rete neurale presentata nel paragrafo 4.5.2, addestrata con immagini convertite in scala di grigio, sebbene si è notato un lieve calo dell'accuracy in fase di riconoscimento, probabilmente dovuta alla mancanza di definizione nel acquisizione dei tratti del volto in molti dei casi in cui i led a infrarossi non sono stati sufficienti ad illuminare i volti degli spettatori.

Le figure seguenti (Figure 38 - 48) mostrano i risultati dell'analisi emotiva effettuata sul campione di dodici persone per tutti gli spettacoli monitorati. Il grafico a torta a sinistra mostra la prevalenza delle emozioni provate per l'intero spettacolo. La curva blu a destra rappresenta la Valence emotiva.

Il valore della valenza è il valore medio delle emozioni provate da tutte le persone monitorate. Questo significa che la valenza a zero può rappresentare non solo un momento di rilassamento generale, ma anche il risultato di un monitoraggio contrastante, come una scena che suscita emozioni positive per una persona e sentimenti negativi per un'altra, allo stesso tempo.

Nelle figure seguenti sono riportati 14 grafici in totale: 11 si riferiscono ai singoli spettacoli (quattro per Carmen, quattro per Rigoletto e tre per Macbeth), e gli ultimi tre mostrano i risultati aggregati per ogni opera.

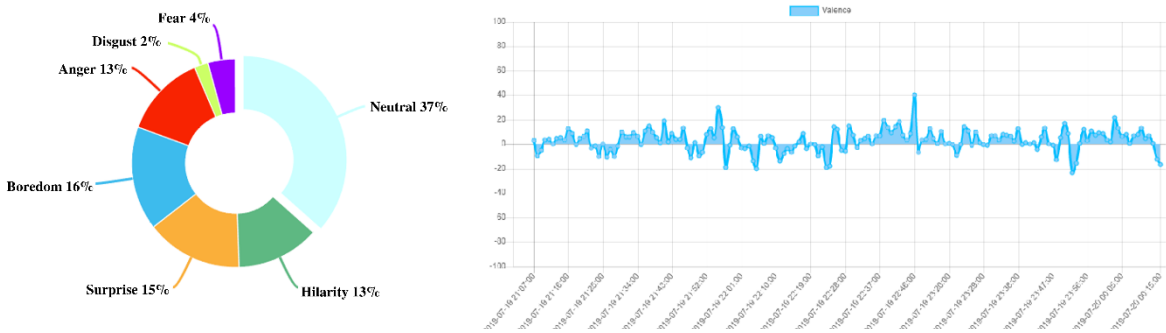

Figura 38: Grafico a torta e curva di valence per la Carmen del 19 Luglio

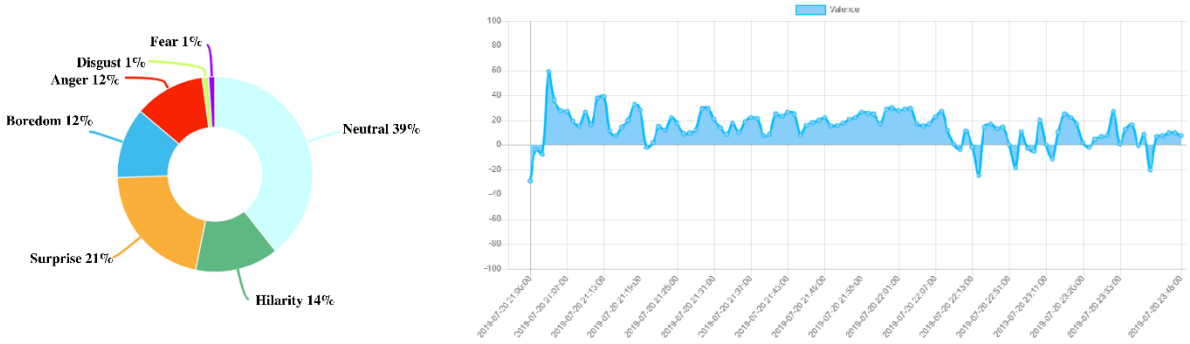

Figura 39: Grafico a torta e curva di valence per il Macbeth del 20 Luglio

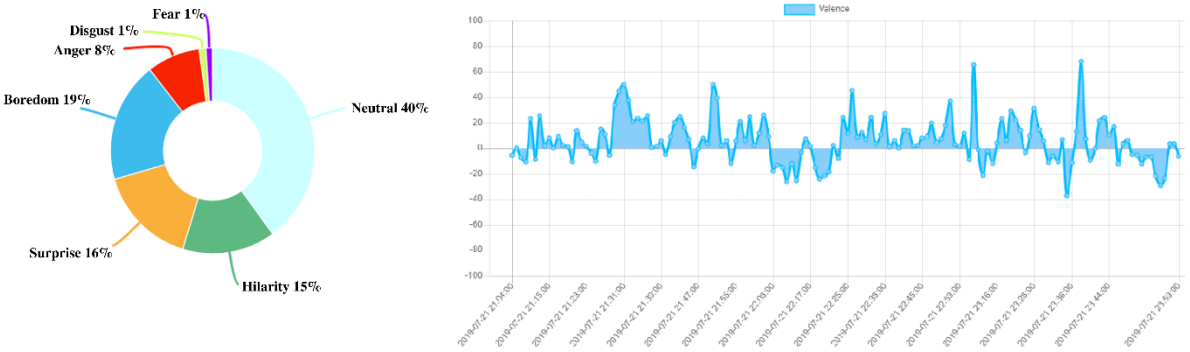

Figura 40: Grafico a torta e curva di valence per il Rigoletto del 21 Luglio

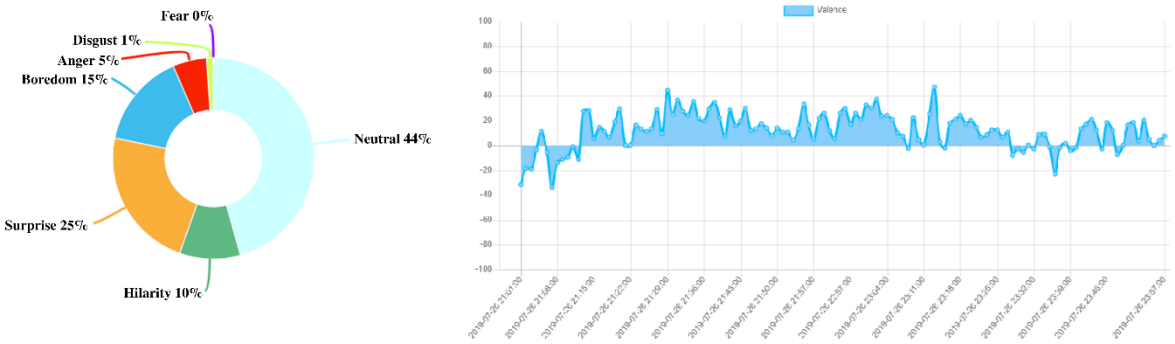

Figura 41: Grafico a torta e curva di valence per il Macbeth del 26 Luglio

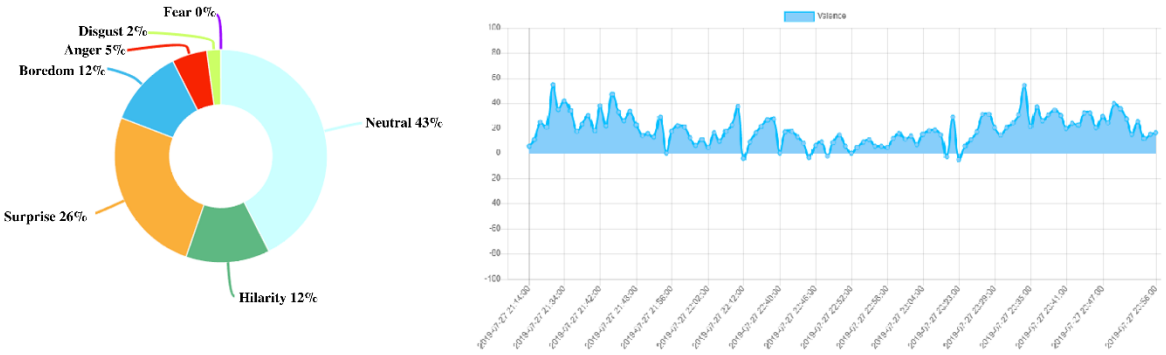

Figura 42: Grafico a torta e curva di valence per il Rigoletto del 27 Luglio

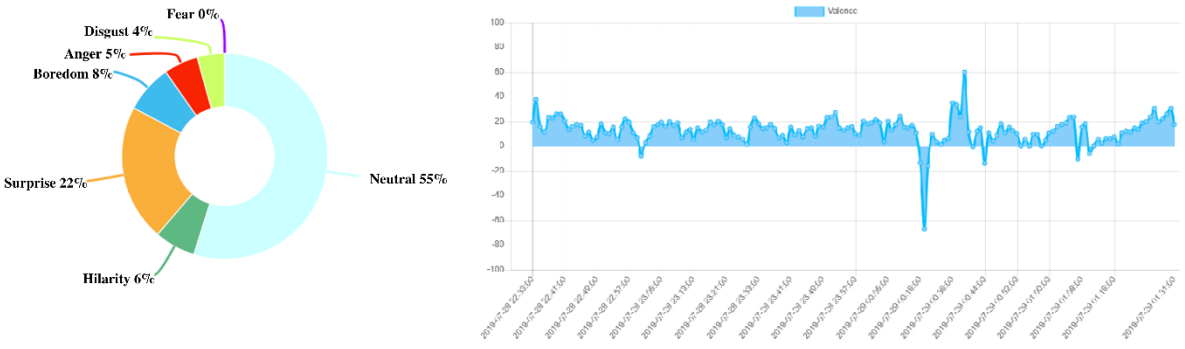

Figura 43: Grafico a torta e curva di valence per la Carmen del 28 Luglio

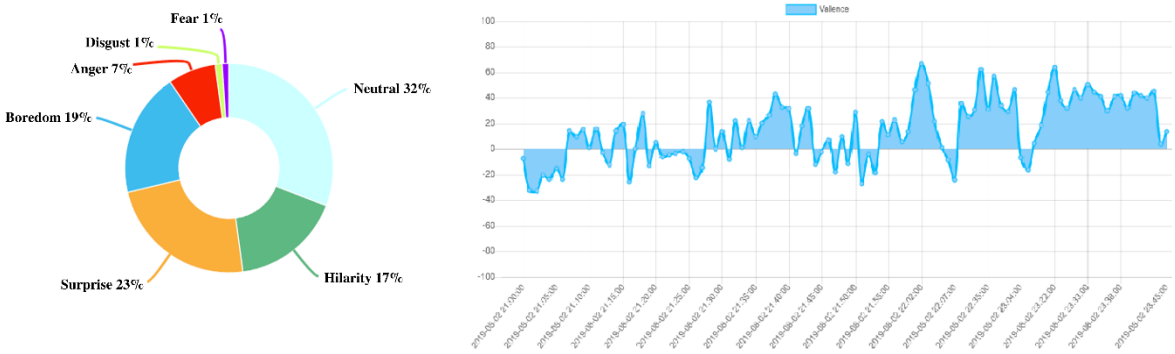

Figura 44: Grafico a torta e curva di valence per il Rigoletto del 02 Agosto

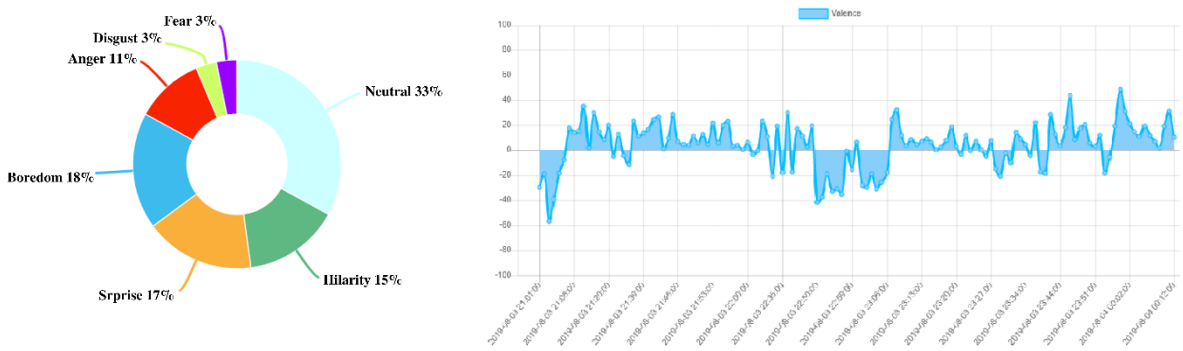

Figura 45: Grafico a torta e curva di valence per la Carmen del 03 Agosto

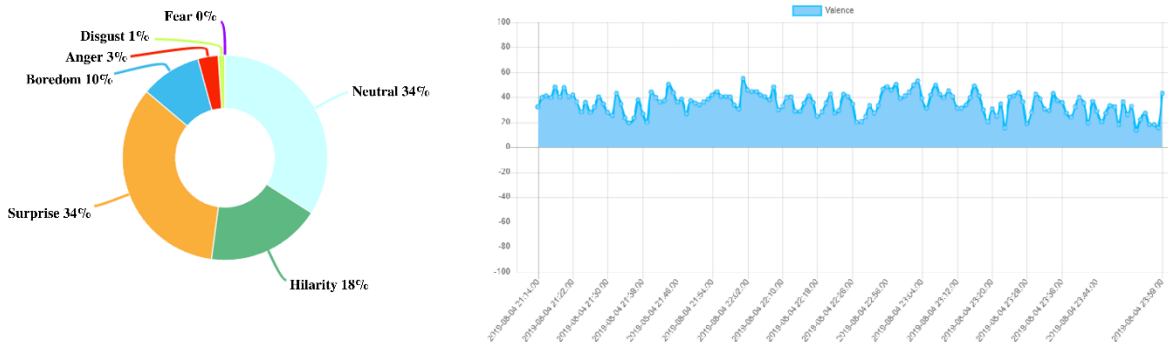

Figura 46: Grafico a torta e curva di valence per il Macbeth del 04 Agosto

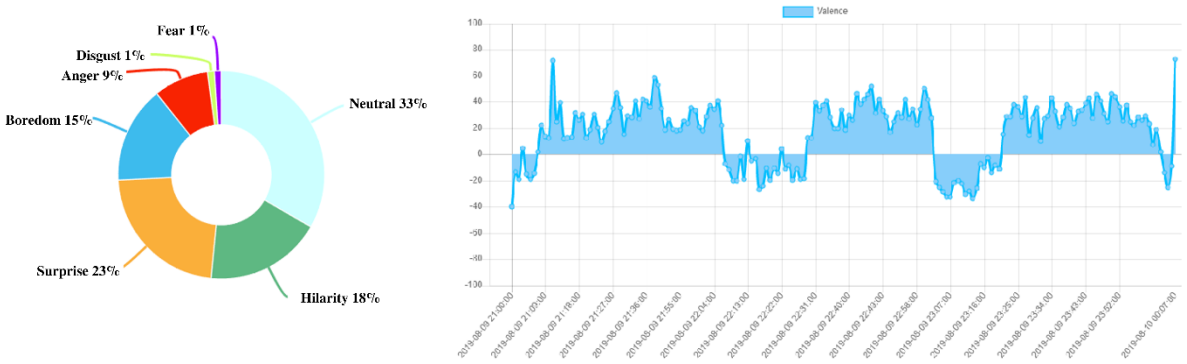

Figura 47: Grafico a torta e curva di valence per il Rigoletto del 09 Agosto

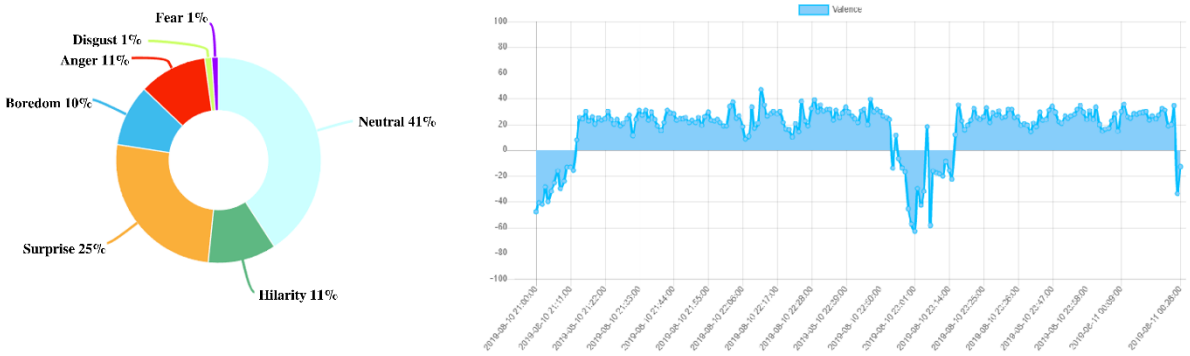

Figura 48: Grafico a torta e curva di valence per la Carmen del 10 Agosto

Il confronto tra le curve di valence della stessa opera nello stesso grafico evidenzia le differenze della performance artistica nei diversi eventi (Figure 49, 50 e 51) e quale aria suscita entusiasmo e gioia. Ad esempio, in Carmen, la compagnia dell'Opera ha mantenuto lo stesso livello di eccitazione del pubblico negli atti N.1, 2 e 4, mentre l'esperienza del pubblico varia nell'atto N.3. Inoltre alcuni fattori contestuali possono influenzare le emozioni. Un esempio è il tempo (ad esempio, freddo o pioggia inattesi). La pioggia ha colpito la Carmen del 28 luglio (curva verde nella figura 43). Ciò è dimostrato dai picchi più negativi all'inizio della seconda metà dello spettacolo. Altri fattori che possono influenzare l'esperienza emotiva riguardano gli eventi principali di una scena (es. un bel bacio tra i due attori principali, il momento dell'omicidio, l'ingresso di tutta la compagnia), l'espressività degli attori, la performance orchestrale, gli elementi scenografici che cambiano durante le diverse sessioni di opera, ecc.

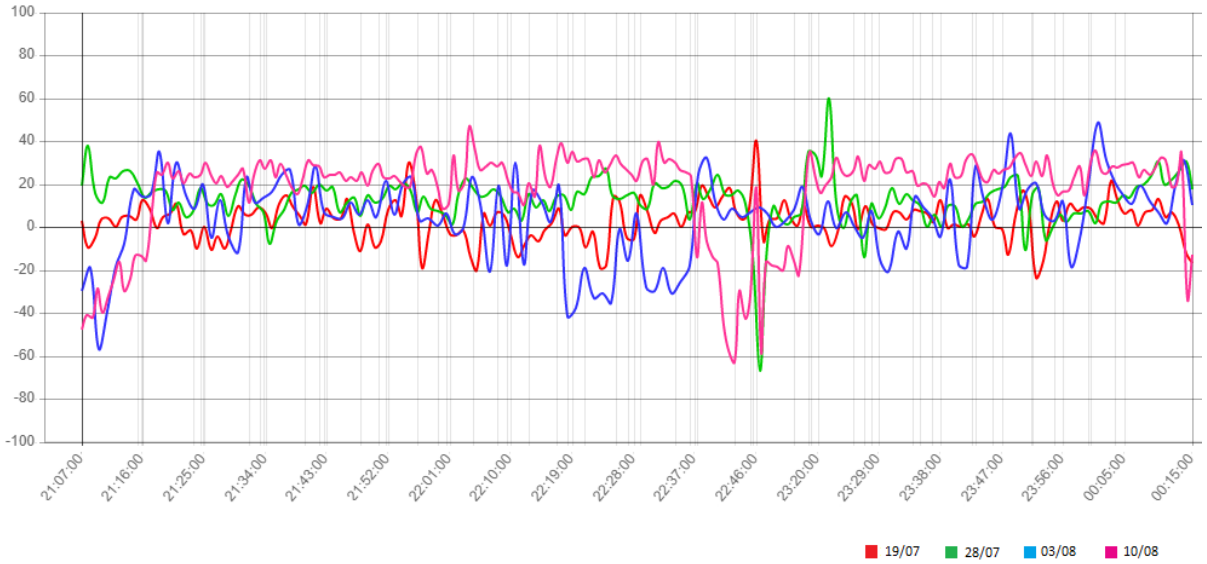

Figura 49: Confronto tra tutti gli spettacoli della Carmen

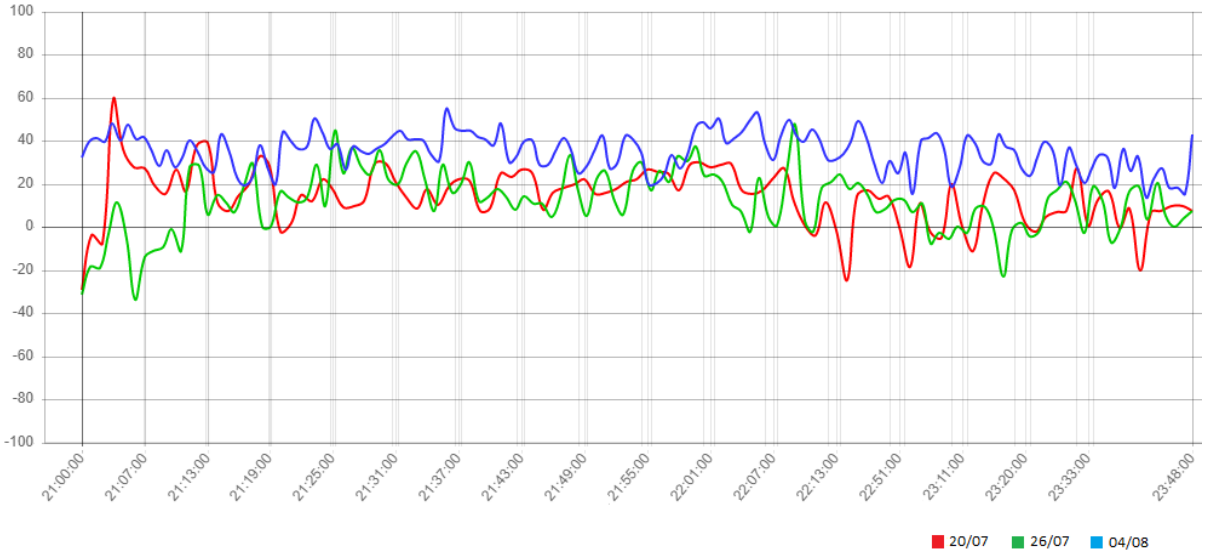

Figura 50: Confronto tra tutti gli spettacoli del Macbeth

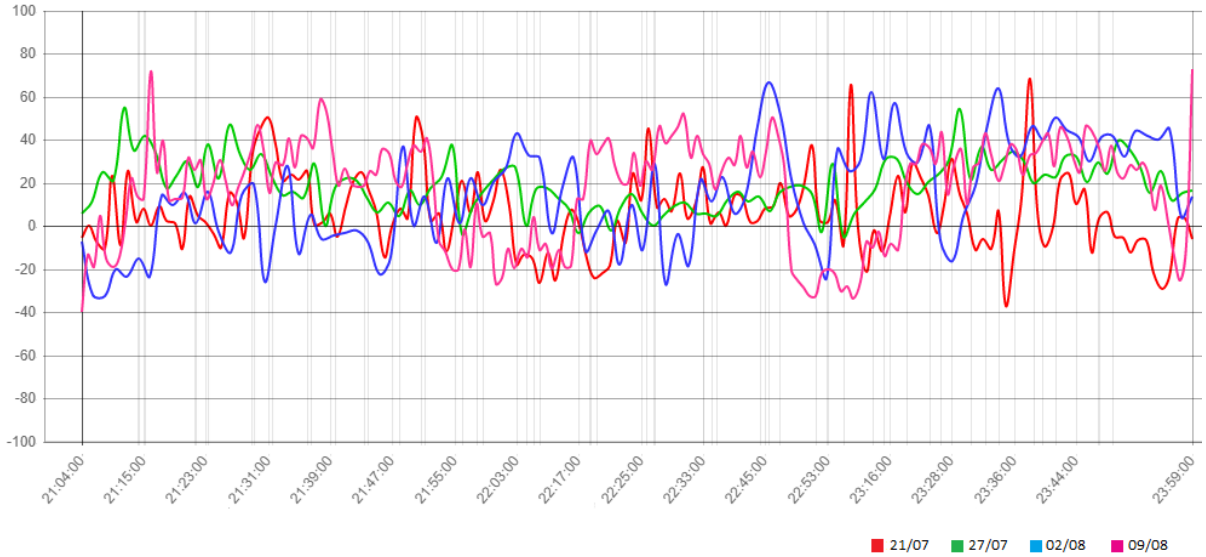

Figura 51: Confronto tra tutti gli spettacoli del Rigoletto

Un'altra tendenza generale registrata riguarda un valore più alto di valence nelle parti cantate che in quelle parlate. Questo ci suggerisce che l'orchestra è uno degli elementi più importanti che definiscono il livello di coinvolgimento del pubblico.
Inoltre, i picchi positivi sono registrati in corrispondenza delle arie più famose, o nelle scene più significative per la trama.
Il confronto della valence media delle tre opere permette allo studio di focalizzarsi su come il pubblico vive le emozioni e di indicare quale opera è stata maggiormente apprezzata (Figura 52). I risultati mostrano che le opere più apprezzate sono Macbeth, con un valori di valence mediamente più alti e il Rigoletto, che ha entusiasmato il pubblico maggiormente

rispetto alle altre opere (mostrando il valore più basso di percentuale di stato neutro).

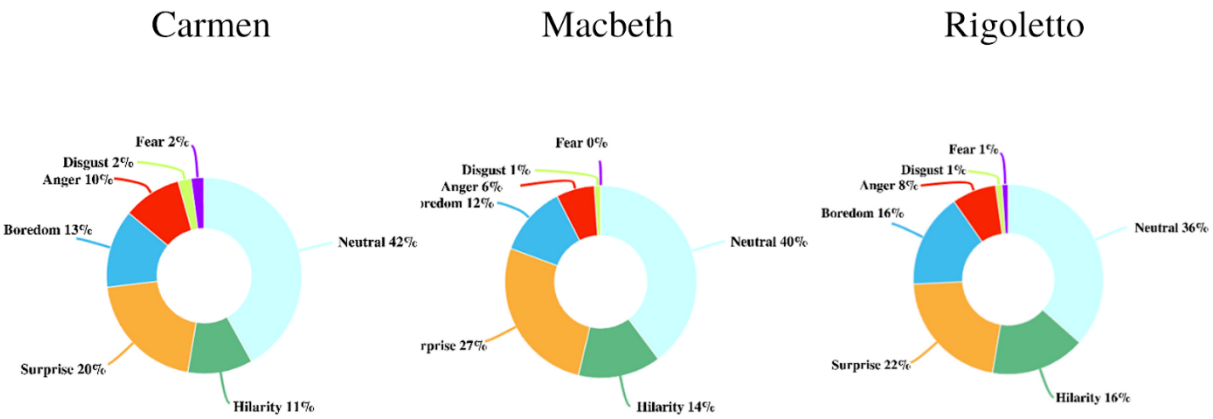

Figura 52: Le emozioni medie per tutte le tipologie di spettacolo

Un'analisi più accurata è possibile mappando le curve emozionali con lo streaming video dello stesso spettacolo. Questo permetterebbe di mettere in relazione le emozioni del pubblico con quanto accade scena per scena. Una specifica è necessaria per la completa comprensione dei grafici sopra riportati. Ci sono alcune differenze nel tempo di monitoraggio degli spettacoli. Dipendono da alcuni fattori operativi come l'avvio e l'arresto della registrazione video durante le pause degli atti che erano manuali.
Di conseguenza, le curve emotive della stessa opera non sono perfettamente sincronizzate. Il confronto e i risultati della ricerca non sono state tuttavia fortemente influenzate da questa lacuna.

Capitolo 6.

Conclusioni

Questa tesi ha presentato un sistema per analizzare ed ottimizzare l'esperienza del cliente nel contesto dello smart retail, per canali sia digitali che fisici. L'approccio seguito ha lo scopo di supportare la definizione dei requisiti e guidare lo sviluppo di ogni strategia e la progettazione di prodotti/servizi, che caratterizza il punto vendita in modo completo. A tal fine, un ruolo centrale per questa piattaforma è ricoperto dall'analisi delle emozioni vissute da ogni cliente ad ogni touchpoint che caratterizza il Customer Journey, la mappatura di queste emozioni al comportamento del cliente e la pianificazione di una strategia di CX, che implica la definizione di azioni adeguate in tempo reale, a breve termine e a lungo termine. Per supportare l'azienda nella realizzazione di questi compiti e per rendere reattiva l'esperienza d'acquisto, sulla base del comportamento e dello stato emotivo del cliente riconosciuto nei vari punti di contatto, è stato proposto un sistema basato su tecnologie di Deep Learning che permettono una categorizzazione automatica di immagini provenienti da dispositivi di acquisizione video. Nel contesto del retail, questo sistema introduce diverse innovazioni:

- **Monitoraggio automatico della reale esperienza del cliente in tutti i touchpoint.** Il sistema implementa un'innovativa piattaforma di riconoscimento emozionale in tempo reale in grado di monitorare i clienti in modo non intrusivo. Di conseguenza, il sistema sarà in grado di fornire all'azienda un'enorme quantità di dati sui consumatori, comprese le loro emozioni spontanee, che le tecniche etnografiche tradizionali non potrebbero mai raccogliere. In particolare, l'utilizzo di questa tecnologia consentirà di ottenere un profilo del cliente più elaborato di quello ottenibile da semplici dati personali o da sondaggi, e renderà possibile la proposta di offerte personalizzabili.
- **Sistema di supporto decisionale basato sulle emozioni dei clienti**. I DSS tradizionali non possono considerare l'impatto delle decisioni di gestione sulle emozioni dei clienti.
- **Soluzione non invasiva che integra diverse tecnologie**. Lo strumento presentato permetterà di monitorare lo stato emotivo dei clienti senza che essi ne abbiano una reale percezione, sebbene debbano ovviamente esserne informati per questioni sia di privacy che etiche.
- **Utilizzo di tecnologie pervasive**, come le camere, presenti già nella maggior parte dei dispositivi di uso quotidiano come gli smartphone e i laptop
- **Architettura totalmente modulare**. Ogni tecnologia utilizzata per il riconoscimento di comportamenti, emozioni e caratteristiche dell'utente costituisce un modulo indipendente. In questo modo, ogni modulo può funzionare come strumento autonomo, in modo che la funzionalità del sistema non venga compromessa in caso di moduli mancanti.

- **Tecnologie applicabili in contesti retail omnicanale**. Un altro dei punti forti di questa piattaforma è la sua applicabilità in praticamente tutti i canali retail, sia fisici che digitali

Altri miglioramenti possono essere applicati a questa piattaforma e ulteriori aspetti possono venire indagati ulteriormente, soprattutto per quanto riguardano le applicazioni dei dati acquisiti in fase di rilevamento e registrati nel database: le reazioni e l'analisi tramite tecniche e strumentazioni di analytics, non implementate (o solo parzialmente per uno specifico caso studio) e dunque trattate marginalmente in questo progetto di tesi.

Bibliografia

Achar, C., So, J., Agrawal, N., and Duhachek, A., 2016, "What we feel and why we buy: the influence of emotions on consumer decision-making, " Current Opinion in Psychology, Vol. 10, pp. 166-170.

Affectiva Media Analytics, Retrieved November 29, 2019, from https://www.affectiva.com/product/affdex-for-market-research/

Alessandra, T., & O'Connor, M. J. (2008). The Platinum Rule: Discover the Four Basic Business Personalities andHow They Can Lead You to Success. Grand Central Publishing.

Amazon Mechanical Turk. https://www.mturk.com/ [Retrieved November 29, 2019].

Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C. A., ... & John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. International journal of human-computer studies, 66(5), 303-317.

Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016, October). Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 279-283). ACM.

Berry, L. L., Carbone, L. P., and Haeckel, S. H., 2002, "Managing the total customer experience, " MIT Sloan management review, Vol. 43, No. 3, pp. 85-90.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. Emotion, 1(3), 276.
Chamberlain, L. (2007). Eye tracking methodology; Theory and practice. Qualitative Market Research: An International Journal.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008.

Chen-Yu, H. J., and Kincade, D. H., 2001, "Effects of product image at three stages of the consumer decision process for apparel products: Alternative evaluation, purchase and post-purchase, " Journal of Fashion Marketing and Management: An International Journal, Vol. 5, No. 1, pp. 29-43.

Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015, November). Video and image based emotion recognition challenges in the wild: Emotiw 2015.

In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 423-426). ACM.

Dhall, A., Goecke, R., Joshi, J., Hoey, J., & Gedeon, T. (2016, October). Emotiw 2016: Video and group-level emotion recognition challenges. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 427-432). ACM.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017, November). From individual to group-level emotion recognition: EmotiW 5.0. In Proceedings of the 19th ACM international conference on multimodal interaction (pp. 524-528). ACM.

Eiter, T., Ianni, G., Polleres, A., Schindlauer, R., & Tompits, H. (2006, September). Reasoning with rules and ontologies. In Reasoning Web International Summer School (pp. 93-127). Springer, Berlin, Heidelberg.

Ekman, P., and Keltner, D., 1970, "Universal facial expressions of emotion," California Mental Health Research Digest, Vol. 8, No. 4, pp. 151-158.

Ekman, P., and Wallace W. V., 1977, Manual for the Facial Action Coding System, Palo Alto: Consulting Psychologists Press.

EmotioNet Challenge, Retrieved April 15, 2019 from https://cbcsl.ece.ohio-state.edu/EmotionNetChallenge/index.html

Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5562-5570).

Gailly, F., & Geerts, G. L. (2013). Ontology-driven business rule specification. Journal of Information Systems, 27(1), 79-104.

Giraldi, L., Mengoni, M., & Bevilacqua, M., 2016, "How to Enhance Customer Experience in Retail: Investigations Through a Case Study, " In Transdisciplinary Engineering: Crossing Boundaries: Proceedings of the 23rd ISPE Inc. International Conference on Transdisciplinary Engineering October 3-7, Vol. 4, p. 381.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2013, November). Challenges in representation learning: A report on three machine learning contests. In International Conference on Neural Information Processing (pp. 117-124). Springer, Berlin, Heidelberg.

Goolsbee, A. D., & Klenow, P. J. (2018, May). Internet rising, prices falling: Measuring inflation in a world of e-commerce. In *AEA Papers and Proceedings* (Vol. 108, pp. 488-92).

Gonzalez-Sanchez, J., Baydogan, M., Chavez-Echeagaray, M. E., Atkinson, R. K., & Burleson, W. (2017). Affect measurement: A roadmap through approaches, technologies, and data analysis. In Emotions and Affect in Human Factors and Human-Computer Interaction (pp. 255-288). Academic Press.

Hamdi, H., Richard, P., Suteau, A., & Allain, P. (2012, June). Emotion assessment for affective computing based on physiological responses. In 2012 IEEE International Conference on Fuzzy Systems (pp. 1-8). IEEE.

Hansen, D. W., & Ji, Q. (2009). In the eye of the beholder: A survey of models for eyes and gaze. IEEE transactions on pattern analysis and machine intelligence, 32(3), 478-500.

Hay, D., Healy, K. A., & Hall, J. (2000). Defining business rules-what are they really. Final report, 34.

Hassenzahl, Marc. "User experience (UX) towards an experiential perspective on product quality." Proceedings of the 20th Conference on l'Interaction Homme-Machine. 2008.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hepp, M. (2008, September). Goodrelations: An ontology for describing products and services offers on the web. In International conference on knowledge engineering and knowledge management (pp. 329-346). Springer, Berlin, Heidelberg.

Hui, T., & Sherratt, R. (2018). Coverage of emotion recognition for common wearable biosensors. Biosensors, 8(2), 30.

IMDB-WIKI, https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/ [Retrieved November 28, 2019].

Jarrold, W. L. (2004). Towards a theory of affective mind: computationally modeling the generativity of goal appraisal (Doctoral dissertation).

Khuong, M. N., and Tram, V. N. B., 2015, "The Effects of Emotional Marketing on Consumer Product Perception, Brand Awareness and Purchase Decision-A Study in Ho Chi Minh City, Vietnam, " Journal of Economics, Business and Management, Vol. 3, No. 5, pp. 524-530.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2176-2184).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Lambda Labs. (n.d.). Retrieved November 29, 2019 from https://lambdalabs.com/api-documentation

Li, S., & Deng, W. (2018). Deep facial expression recognition: A survey. arXiv preprint arXiv:1804.08348.

Lang, P., & Bradley, M. M. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. Handbook of emotion elicitation and assessment, 29.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In Proceedings of the iEEE conference on computer vision and pattern recognition workshops (pp. 34-42).

Littauer, F. (1992). Personality plus. Revell.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (pp. 94-101). IEEE.

Maguire, M. (2001). Methods to support human-centred design. International journal of human-computer studies, 55(4), 587-634.

Manning, Harley, and Kerry Bodine. Outside in: The power of putting customers at the center of your business. Houghton Mifflin Harcourt, 2012.

Maria, E., Matthias, L., & Sten, H. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. Electronic Notes in Theoretical Computer Science, 343, 35-55.

Max-Neef, M. A. (1992). Human scale development: conception, application and further reflections (No. 04; HC125, M3.).

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261-292

Meyer, C., and Schwager, A., 2007, "Understanding Customer Experience," Harvard Business Review, Vol. 85, No. 2, pp.117-126.

Microworker. https://ttv.microworkers.com/index/template [Retrieved May 21, 2018].
Kagan, J., & Kagan, L. (2002). Surprise, uncertainty, and mental structures. Harvard University Press.

PeakProfiling, Retrieved November 05, 2019
from https://www.peakprofiling.com/technology.html

Quintana, D. S., Guastella, A. J., Outhred, T., Hickie, I. B., & Kemp, A. H. (2012). Heart rate variability is associated with emotion recognition: direct evidence for a relationship between the autonomic nervous system and social cognition. International Journal of Psychophysiology, 86(2), 168-172.

Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.

Savran, A., Gur, R., & Verma, R. (2013). Automatic detection of emotion valence on faces using consumer depth cameras. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 75-82).

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

Cloud Speech–To–Text – Riconoscimento Vocale | Google Cloud Platform. (n.d.). Retrieved November 09, 2019, from https://cloud.google.com/speech-to-text/

Stauss, B., & Weinlich, B. (1997). Process-oriented measurement of service quality: Applying the sequential incident technique. European Journal of Marketing, 31(1), 33-55.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

Teixeira, J., Patrício, L., Nunes, N. J., Nóbrega, L., Fisk, R. P., & Constantine, L. (2012). Customer experience modeling: from customer experience to service design. Journal of Service Management, 23(3), 362-376.

Ullah, A. S., Sato, M., Watanabe, M., & Rashid, M. M. (2016). Integrating CAD, TRIZ, and customer needs. Int. J. Automation Technol, 10(2).

Vadi, M., & Suuroja, M. (2006). Training retail sales personnel in transition economies: Applying a model of customer-oriented communication. Journal of Retailing and Consumer Services, 13(5), 339-349.

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. Journal of retailing, 85(1), 31-41.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), 1162-1181.

VPGLIB (2017, January 23). Pi-null-mezon/vpglib. Retrieved November 09, 2019, from https://github.com/pi-null-mezon/vpglib

Watson Developer Cloud, Retrieved February 09, 2019, from https://cloud.ibm.com/docs/services/tone-analyzer?topic=tone-analyzer-ssbts

World Health Organization. (2001). International classification of functioning, disability and health: ICF. Geneva: World Health Organization.

Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4511-4520).

Zomerdijk, L.G., and Voss, C.A., 2010, "Service design for experience centric services, " Journal of Service Research, Vol. 13, No. 1, pp. 67-82.