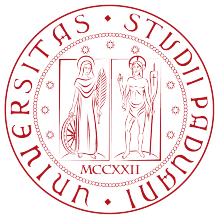

DIPARTIMENTO MATEMATICA

Università di Padova

Laurea Magistrale in Informatica

Corso di Intelligenza Artificiale

Servivi cognitivi e analisi visiva

Marco Romanelli

marco.romanelli.1@studenti.unipd.it

matricola n. 1106706

22 giugno 2017

Indice

1 Introduzione

1.1 Scopo dell'analisi

Questa analisi consiste in una panoramica sulle principali piattaforme che offrono servizi di *Cognitive Computing*, partendo con un'introduzione generale per poi focalizzarsi sull'analisi delle immagini.

1.2 I servizi cognitivi

Si tratta di una serie di servizi che consentono agli sviluppatori di realizzare applicazioni in grado di analizzare e interpretare la realtà, usando quelli che si usa chiamare col nome di Metodi di Comunicazione Naturale. Questi servizi spaziano in aree come la visione, il linguaggio, il parlato, la ricerca, la conoscenza, eccetera; per esempio, l'area di visione racchiude quelle tecniche per l'analisi, la rappresentazione di immagini e video, mentre nell'area del linguaggio fornisce strumenti per capire meglio cosa vuole l'utente tramite l'interpretazione della scrittura o del linguaggio naturale.

Dal punto di vista tecnico, i Servizi Cognitivi sono una raccolta di metodi (API) che aiutano professionisti e programmatori a realizzare programmi "intelligenti". Sostanzialmente, si effettuano delle chiamate *REST* (a *endpoint* http) passando le immagini o i dati da analizzare e il servizio restituirà un risulato (a seconda del metodo utilizzato). Questo è possibile sfruttando le complesse infrastutture cloud e lo stato dell'arte di tecniche e algoritmi.

L'articolo procederà come segue: nella Sezione 2 verranno presentate brevemente le piattaforme prese in esame, illustrando per ognuna i vari sevizi offerti. Seguiranno poi nelle sezioni 3-6 le analisi dettagliate di ciascuna piattaforma, focalizzandosi sull'analisi delle immagini. La Sezione 7 verranno presentati alcuni esempi di applicazione, mentre nella Sezione 8 saranno prese in esame situazioni reali valutando l'aspetto finanziario. Infine, la Sezione 9 riassumerà brevemente i concetti visti inserendo alcune conclusioni finali e possibili sviluppi futuri. In Appendice A sono presenti alcune tabelle riassuntive.

2 Panoramica

Molte fra le maggiori aziende nel settore tecnologico e informatico hanno cominciato a offrire servizi cognitivi. Fra queste ne sono state individuate quattro, che offrono un'ampia gamma di sevizi, tariffe diversificate in base all'esigenza e la possibilità di sopportare carichi di lavoro molto ingenti. I servizi presi in esame saranno, quindi:

- Microsoft Cognitive Services [1] (Microsoft Corporation),
- Watson Services (Bluemix) [2] (IBM: International Business Machines Corporation),
- Amazon Artificial Intelligence [3] (Amazon.com, Inc)
- Google Cloud Machine Learning Services [4] (Google Inc.)

Ognuna di queste suddivide i servizi in macro aeree che possono essere riassunte così: visione artificiale, sintesi vocale, linguaggio naturale, ricerca, tecniche di apprendimento e altro. Non tutte le piattaforme utilizzando la stessa nomenclatura e i servivi al loro interno possono variare leggermente, ma le aree di interesse coperte sono più o meno queste. L'area denominata *visione artificiale* include riconoscimento visivo di immagini e video, estrazione di informazioni, riconoscimento volti ed emozioni. Con *sintesi vocale* vengono indicati quei servizi atti all'elaborazione dell'audio e alla sua trasformazione in testo o in strutture dati adatte all'analisi.

L'area *linguaggio* permette di analizzare ed elaborare il linguaggio naturale, come ad esempio comprendere comandi ed estrapolare informazioni importanti da un dato contesto. La *ricerca* permette di sfruttare le potenzialità del motore di ricerca offerto dalla compagnia stessa (se presente) o di eseguire ricerche avanzate all'interno di collezioni create appositamente. *Tecniche di apprendimento* permette la fruizione di algoritmi e modelli, oltre che alla loro creazione, per l'apprendimento automatico e approfondito. La macro area *altro* comprende tutti quei servizi che non ricadono nelle aree precedentemente descritte, come ad esempio sistemi di raccomandazione o altro. Infine, la Tabella 1 riassume i servivi offerti (ad alto livello) da ogni piattaforma, inseriti nelle macro aree di riferimento.

È necessario aggiungere, tuttavia, che questa classificazione vuole fornire solamente una visone generale dei servizi disponibili al momento della stesura di questo documento. Per l'offerta completa si rimanda alle rispettive documentazioni. Inoltre, l'assenza di voci in una macro area per una certa piattaforma non esclude la presenza di relativi servizi; potrebbero essere, infatti, presenti sotto altri nomi, piattaforme, framework o comunque presenti negli altri servizi.

3 Microsoft Cognitive Services

Per quanto concerne il riconoscimento delle immagini, Microsoft offre diverse API per l'analisi delle immagini, raggruppate nella categoria *Vision*:

- *Computer Vision API*: che comprende diverse funzioni, dal riconoscimento di oggetti alla creazione di anteprime;
- *Content Moderator*: per aiutare i moderatori nell'analisi di immagini, testi e video (verrà analizzata solo l'analisi di immagini);
- *Emotion API*: per il riconoscimento di emozioni;
- *Face API*: per il ricongiunto di volti;
- *Video API*: per l'analisi di video (non verrà preso in considerazioni in questa analisi).

3.1 Prerequisiti

Caratteristiche immagini[1]:

- Metodo: dati grezzi (stream application/octet) o URL.
- Formati supportati: JPEG, PNG, GIF, BMP.
- Caratteristiche minime: 50x50 pixel.
- Dimensione massima: 4 MB.

Altro:

- Disponibilità: Stati Uniti occidentali, Stati Uniti Orientali 2, Stati Uniti centro-occidentali, Europa occidentale, Asia sud-orientale [2].

3.2 Computer Vision API

In base allo scopo per cui si vuole analizzare l'immagine, le *Computer Vision API* [5] mettono a disposizione diversi metodi per ottenere le informazioni desiderate.

[1] Per un confronto con i requisiti richiesti dalle altre piattaforme, vedere la Tabella 6.
[2] Per conoscere in dettaglio le aree, visitare la pagina Azure regions.

Tabella 1: Tabella riassuntiva dei servizi offerti, raggruppati per macro aree.

Macro Aree	Microsoft Cognitive Services	Watson Services	Amazon Artificial Intelligence	Google C.M. Learning Services
Visione artificiale	Computer Vision API Content Moderator Emotion API Face API Video API	Visual Recognition	Amazon Rekognition	Video Intelligence API Vision API
Sintesi vocale	Bing Speech API Custom Speech Service Speaker Recognition API	Speech to Text Text to Speech	Amazon Polly	Speech API
Linguaggio naturale	Bing Spell Check API Language Understanding Intelligent Service Linguistic Analysis API Text Analytics API Translator API Web Language Model API	AlchemyLanguage Conversation Dialog Document Conversion Language Translator Natural Language Classifier Natural Language Understanding Personality Insights Retrieve and Rank Tone Analyzer	Amazon Lex	Natural Language API Translation API
Ricerca	Bing Autosuggest API Bing Image Search API Bing News Search API Bing Video Search API Bing Web Search API Academic Knowledge API Knowledge Exploration Service	Discovery Discovery News	-	-
Apprendimento	-	-	Amazon Machine Learning Apache Spark su Amazon EMR	Machine Learning Engine
Altro	Entity Linking Intelligence Service QnA Maker Recommendations API	Tradeoff Analytics	-	Jobs API

Tagging Le API ritornano un insieme di etichette (in formato JSON) che descrivono gli oggetti presenti nell'immagine, come oggetti, esseri viventi, azioni, paesaggi; per ogni etichetta viene anche fornito il livello di *confidence* (affidabilità). I tag non sono in alcun modo organizzati fra loro e non esiste nessun tipo di ereditarietà. Nel caso un tag sia ambiguo viene fornito in aggiunta un aiuto (*hint*) che ne spiega il contenuto. Al momento la sola lingua supportata è l'inglese.

Classificazione L'immagine viene classificata in categorie che seguono una tassonomia con ereditarietà di tipo padre-figlio. Questa tassonomia prevede 86 categorie[3] e classifica gli elementi visivi in modo più o meno specifico. Per esempio, una categoria è *food*, che comprende *bread*, *pizza*, *fastfood*, ...

Identificazione del tipo È possibile classificare l'immagine come in bianco o nero o a colori, se è un disegno o se è del tipo *clip-art*; in quest'ultimo caso viene fornito un livello di qualità dell'immagine, compreso fra 0 e 3.

Riconoscimento volti Riconosce i volti umani e restituisce la posizione (coordinate) di questi all'interno dell'immagine, come anche età e sesso della persona.

Contenuto personalizzato Ideato per raffinare la tassonomia a 86 categorie utilizzando informazioni specifiche sul dominio. Attualmente sono supportati solamente il riconoscimento dei volti delle persone famose e luoghi di interesse (categorie: persone, gruppi di persone e luoghi di interesse).

Generazione di descrizioni Genera una lista di frasi (in lingua inglese) che descrivono il contenuto dell'immagine, ordinate secondo un livello di affidabilità calcolato per ogni descrizione.

Estrazione colori Identifica i colori analizzandoli in tre contesti: di sfondo, in primo piano e d'insieme; i colori sono raggruppati in 12 colori predominanti. Classifica le immagini fra in bianco e nero e a colori.

Riconoscimento contenuti non adatti ai minori Riconosce materiali pornografici e contenuti osé in generale. Può essere impostato un livello per il filtro.

Riconoscimento del testo (OCR) Rileva il testo presente nell'immagine e lo trasforma in un flusso di parole, ruota l'immagine se necessario per rendere il testo orizzontale e fornisce le coordinate per ogni parola. Al momento sono supportati 21 linguaggi, fra cui l'inglese, l'italiano, il francese, il tedesco e lo spagnolo.

L'accuratezza del riconoscimento dipende dalla qualità dell'immagine ed eventuali errori possono essere causati da immagini sfuocate, scrittura a mano, testo troppo piccolo, ecc.

Creazione anteprime Un'anteprima è una rappresentazione dell'immagine in scala ridotta. L'immagine viene prima analizzata e poi ritagliata secondo la "regione di interesse" (ROI); il rapporto dell'immagine (*aspect ratio*) può essere impostato secondo le proprie preferenze.

[3] `https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/category-taxonomy`

3.3 Content Moderator API

Le *Content Moderator's image moderation API* [6] permettono di rilevare immagini a contenuto pornografico e contenuti per adulti in generale. Si possono distinguere tre tipi di operazione.

- Evaluate: rileva se l'immagine contiene contenuti per adulti e/o contenuti osé.
- Find Faces: permette di identificare la probabilità di trovare dei volti e, in caso positivo, quanti ne sono presenti; potrebbe essere utilizzato, ad esempio, per evitare che gli utenti inviino foto personali riguardanti la propria persona.
- Match: è possibile creare e personalizzare una lista dei contenuti che si vogliono bloccare; a questo punto il sistema di moderazione confronta le immagini caricate con quelle presenti nella lista, identificando non solo i doppioni ma anche versioni leggermente modificate. Il sistema fornisce anche un insieme di etichette per identificare al meglio il tipo di contenuto, come nudità, alcool, armi, sfruttamento minorile, eccetera.

3.4 Emotion API

Dato uno o più volti in un'immagine, l'*Emotion API* [7] permette di identificare le espressioni del viso e riconoscere quali emozioni vuole trasmettere. Le emozioni rilevabili sono rabbia, paura, felicità, espressione neutra, tristezza, sorpresa, disprezzo e disgusto[4]. Per ognuna di queste, viene restituito un punteggio corrispondente alla probabilità che quella data emozione sia espressa nel volto.

L'utilizzo di questa funzione è di due tipologie:

- base: se l'utente ha già utilizzato le API per il riconoscimento facciale *Face API*, allora nella chiamata può includere le coordinate del viso (o, meglio, del rettangolo che lo include) e utilizzare questa tipologia (pagando un costo inferiore);
- standard: se non è stata effettuata nessuna chiamata con la *Face API*.

3.5 Face API

Questa funzione permette un'analisi più approfondita dei volti rispetto alle *Computer Vision API*, includendo anche funzioni di confronto e ricerca.

Rilevamento volti Rileva i volti presenti nell'immagine (fino a 64) e restituisce le coordinate del rettangolo che ingloba il viso, gli attributi facciali e i punti di riferimento. I punti di riferimento sono le coordinate degli elementi (di riferimento) del viso, come le pupille, le sopracciglia, i punti della bocca, il naso, eccetera. Fra gli attributi facciali si trovano l'eta, il sesso, l'intensità del sorriso, un valore per indicare la tipologia di barba, la posizione tridimensionale del volto espressa in gradi di Tait-Bryan (imbardata, rollio e beccheggio), la presenza di occhiali, le emozioni (vedi sezione precedente), eccetera.

Verifica di un volto Calcola la probabilità che due volti appartengano alla stessa persona. Per default, due volti sono "dichiarati" appartenenti alla stessa persona se il valore di verosimiglianza è superiore allo $0,5\%$.

Identificazione volti Può essere utilizzato per identificare le persone sulla base di un volto e di un database di persone (chiamato *person group*), creato in precedenza e che può essere aggiornato nel tempo. Dopo aver creato e addestrato il database, si può procedere all'identificazione data una nuova faccia; viene restituita una lista di candidati, ordinati per probabilità decrescente che il volto appartenga a quella persona.

[4] Disprezzo e disgusto sono sperimentali

Ricerca volti simili Fornendo un volto obbiettivo e un insieme di candidati (di volti) all'interno del quale eseguire la ricerca, questa funzione restituisce un piccolo insieme di volti che assomiglia al volto obiettivo. Questo può essere fatto con due modalità: matchPerson e matchFace. Il primo è il metodo di default e cerca di trovare volti simili della stessa persona utilizzando una soglia interna di verosimiglianza; utile per cercare altre foto di una persona conosciuta. Il secondo ignora il valore di soglia e restituisce una lista ordinata di volti simili, anche se il valore di verosimiglianza è basso; può essere utilizzato, ad esempio, per la ricerca di volti di personaggi famosi.

Aggregazione di volti Dato un insieme di volti (da 2 a 1000), questo metodo raggruppa automaticamente i volti simili fra loro creando dei sotto insiemi. Ogni gruppo è un insieme disgiunto proprio dell'insieme di partenza; ogni volto all'interno del gruppo può essere considerato come della stessa persona. È previsto un gruppo speciale, chiamato messyGroup, contente quei volti che non trovano riscontri di similarità.

3.6 Tariffe

Il piano gratuito prevede la possibilità di effettuare 5000 chiamate (all'API) al mese per le *Vision API* e 30000 per la *Emotion API* e *Face API*, con un limite di 20 chiamate al minuto tranne che per la moderazione, dove il limite è di una chiamata/secondo. I piani a pagamento, invece, variano dai 2,50$ per 1000 chiamate fino a 0,65$, in base al servizio richiesto e al numero di chiamate effettuate al mese. La Tabella 2 riassume i costi, espressi in dollari americani (USD) ogni 1000 chiamate; nelle colonne troviamo il carico di lavoro, espresso in chiamate alle API al mese.

Per la *Content Moderator API*, invece, il piano gratuito prevede 5000 chiamate al mese e un massimo di una al secondo. Il piano standard parte da 1,00$ ogni 1000 chiamate fino a 1M di chiamate/mese, 0,75$ da 1M a 5M, 0,60$ da 5M a 10M e 0,60$ fino a 20M.

Metodo	1 - 1M	1M - 5M	5M - 20M
Tag	1,00	0,80	0,65
OCR	1,50	1,00	0,65
Handwriting OCR	2,50	-	-
Describe	2,50	-	-
Adult	1,5	1,00	0,65
Face	1,00	0,80	0,65
Categories	1,00	0,80	0,65
Celebrity	1,50	1,00	0,65
Landmark	1,50	1,00	0,65
Get Thumbnail	1,00	0,80	0,65
Color	1,00	0,80	0,65
Image Type	1,00	0,80	0,65

Tabella 2: Tariffe per le Computer Vision API

Per le *Emotion API* è previsto l'utilizzo gratuito fino a 30K chiamate al mese. Alternativamente il costo è 0,10$ ogni 1000 chiamate per il piano base e 0,25 $ per quello standard[5].

Per le *Face API* il limite per il piano gratuito è lo stesso della precedente, mentre i costi variano in base al numero di transizioni: fino a 1M il costo è di 1,50$ ogni 1000 chiamate, fino a 5M di 1,10$ e fino a 20M di 0,65$. Il costo per lo spazio di archiviazione è di 0,50$ ogni 1000 immagini/mese con un massimo di 4MB a immagine.

[5]Per le differenze fra il piano base e standard si rimanda alla documentazione ufficiale.

4 IBM Watson Services

Il servizio di Visual Recognition [8] utilizza tecniche e algoritmi di *deep learning* per identificare scene, oggetti, visi di persone nell'immagine che viene fornita come input al servizio. Permette, inoltre, la creazione e l'addestramento di un classificatore personalizzato per l'identificazione di elementi in base alle necessità dello sviluppatore.

4.1 Prerequisiti

Caratteristiche immagini:

- Metodo: dati grezzi o URL all'immagine.
- Formati supportati: PNG, JPEG.
- Caratteristiche minime: 224x224 pixel.
- Dimensione massima: 2MB.

4.2 Visual Recognition

Classificazione Per ogni immagine sottoposta a classificazione viene fornito in risposta una lista di coppie classe-punteggio per ogni classificatore selezionato. Il punteggio è compreso in un intervallo 0-1, dove un valore maggiore indica una probabilità più alta che la classe descriva l'immagine; la soglia di default perché un valore sia ritornato da un classificatore è $0,5$. Le classi sono organizzate in categorie e sotto-categorie dove il livello più astratto comprende categorie quali animali, persone, cibo, sport, natura, eccetera.

Le lingue sopportate[6] nella risposta sono l'inglese, spagnolo, arabo o giapponese.

Riconoscimento dei volti Analizza i volti presenti l'immagine e ne deriva alcune informazioni, come età stimata, sesso o nome del personaggio famoso (nel caso ci sia). Anche in questo caso viene fornito un punteggio (nell'intervallo $0-1$) atto ad indicare una maggiore probabilità di correlazione.

Classificatore personalizzato Permette di creare un nuovo classificatore e di addestrarlo su un dato insieme di immagini. Queste sono inviate in un file compresso e devono comprendere o due immagini d'esempio positive o una positiva e una negativa. L'insieme contente le immagini d'esempio positive serve a creare le classi che definiscono il nuovo classificatore. Il complementare definisce invece quello che il classificatore *non* deve essere; le immagini d'esempio negative non devono contente i soggetti presenti nelle immagini positive.

Se, ad esempio, si volesse creare un classificatore "frutta" si potrebbe utilizzare un file compresso contente immagini di pere, uno contente immagini di mele e uno con immagini di banane. Per le immagini d'esempio negative si potrebbero utilizzare immagini di verdure.

Collezioni Questa funzione[7] permette di creare una nuova collezione, aggiungere immagini a questa e utilizzare la *Similarity Search* per cercare immagini simile all'interno della collezione.

Note per la privacy Per default, tutte le immagini e le informazioni inviate vengono salvate e utilizzate per migliorare il servizio. Per evitare questo è necessario impostare diversamente il parametro X-Watson-Learning-Opt-Out per ogni richiesta inviata.

[6] Al momento della stesura di questo documento.
[7] Questa funzione è ancora in fase BETA.

4.3 Tariffe

Il piano gratuito (della durata di 30 giorni) prevede la possibilità di:

1. classificare 250 immagini al giorno,
2. addestrare un solo classificatore personalizzato con massimo 5000 immagini.

Il piano *standard* prevede:

1. per la classificazione: $0,002$ dollari a immagine,
2. per il riconoscimento volti: $0,004$ dollari a immagine,
3. per l'addestramento classificatore: $0,10$ dollari a immagine,
4. per la classificazione con classificatore personalizzato: $0,004$ dollari a immagine.

5 Amazon Artificial Intelligence

Amazon Rekognition [9] è il servizio di Amazon che permette di riconoscere oggetti, volti, scene, nelle immagini e molto altro. Inoltre permette l'integrazione con le altre piattaforme offerte da Amazon, come Amazon S3, AWS Lambda e altri servizi AWS.

5.1 Prerequisiti

Caratteristiche immagini:

- Metodo: dati grezzi (stream application/octet) o oggetto Amazon S3.
- Formati supportati: PNG, JPEG.
- Caratteristiche minime: 80x80 pixel.
- Dimensione massima: 5 MB (dati grezzi), 15 MB (oggetto Amazon S3).
- Massimo numero di immagini per collezione: un milione.

Altro:

- Disponibilità: Stati Uniti orientali, Stati Uniti occidentali, Europa.

5.2 Amazon Rekognition

Prima di descrivere cosa si può fare o meno con Amazon Rekognition, è necessaria fare una distinzione; le operazioni fornite dal servizio si suddividono in due categorie:

- Operazioni volatili (*non-storage API operations*): le operazioni in questo gruppo non salvano alcuna informazione sui server Amazon.
- Operazioni persistenti (*storage-based API operations*): l'utilizzo di queste operazioni comporta il salvataggio di alcuni metadati sui server. Ad esempio nella ricerca di un volto viene salvata la rappresentazione vettoriale dei volti rilevati (mentre l'immagine di partenza no).

Rilevamento scene e oggetti Permette di rilevare automaticamente oggetti, come ad esempio veicoli, alberi, animali e con il relativo punteggio (*confidence score*) che ne indica il grado di affidabilità (probabilità che sia corretto). Permette, inoltre, il riconoscimento di scene (una spiaggia o un tramonto), eventi (matrimoni, feste di compleanno) e concetti astratti (serata, paesaggio, natura).

Analisi volti Permette il riconoscimento di volti all'interno dell'immagine, la loro localizzazione spaziale e l'analisi di attributi facciali come il sesso, l'eta, emozioni, se la persona sta sorridendo, se ha gli occhi aperti, presenza o assenza di barba/baffi, eccetera[8]. La localizzazione avviene tramite un immaginario rettangolo (sotto forma di coordinate (x, y) degli angoli) che circonda ogni viso rilevato e dei punti di rifermento su elementi come il naso, gli occhi, le orecchie, la bocca, ecc. Naturalmente, gli algoritmi di riconoscimento sono più affidabili in presenza di visi rivolti frontalmente e potrebbero non riconoscere (o farlo con un punteggio inferiore) visi oscurati o se rivolti non frontalmente.

Confronto volti A differenza del metodo precedente, questo permette di misurare la probabilità che due volti siano la stessa persona; un punteggio associato ad ogni confronto aiuta a valutarne il risultato. Da un'immagine *sorgente* contente un volto, questo viene confrontato con il volti presente nelle immagini *destinazione*. Anche in questo caso vengono forniture le coordinate spaziali di un immaginario rettangolo che circonda i visi che sono stati rilevati, assieme al grado di sicurezza che quel rettangolo contenga veramente un volto.

Riconoscimento volti Per trovare un volto all'interno di una collezione di immagini. Per prima cosa è necessaria la creazione di una collezione per il salvataggio dei volti, rappresentati come vettore di attributi. Successivamente si fornisce al servizio un'immagine che provvederà alla ricerca di volti simili all'interno della collezione precedentemente creata. Per ogni volto restituito viene associato, al solito, un livello di affidabilità la posizione del volto all'interno dell'immagine.

5.3 Tariffe

Il piano gratuito prevede ogni mese, per i primi 12 mesi, di:

- analizzare 5000 immagini,
- memorizzare 1000 metadati facciali.

Altrimenti:

- per il primo milione di immagini[9]: 1 dollaro ogni 1000 immagini[10];
- successivi 9 milioni di immagini: $0,80$ dollari ogni 1000 immagini;
- successivi 90 milioni di immagini: $0,60$ dollari ogni 1000 immagini;
- oltre i 100 milioni di immagini: $0,40$ dollari ogni 1000 immagini.

Inoltre utilizzando le API per il riconoscimento dei volti, il servizio memorizza ogni volta la rappresentazione vettoriale dei volti. Questo comporta dei costi pari a $0,01$ dollari per 1000 metadati memorizzati al mese.

6 Google Cloud Machine Learning Services

Le *Google Cloud Vision API* [10] permettono di analizzare un'immagine e classificarla in categorie, rilevare oggetti e volti, cercare parole, moderare contenuti offensivi e molto altro. Le immagini possono essere caricate assieme la richiesta, oppure utilizzare quelle già presenti nel Google Cloud Storage.

[8] Per una lista esaustiva si faccia riferimento alla documentazione ufficiale all'indirizzo: API Types.
[9] Ogni API che accetta una o più messaggi di input conta come un'immagine elaborata.
[10] Al mese.

6.1 Prerequisiti

Le immagini passate al servizio devono rispettare i seguenti requisiti:

- Metodo: dati grezzi (stream application/octet) o tramite Google Cloud Storage URIs.
- Formati supportati: JPEG, PNG8, PNG24, GIF, Animated GIF[11], BMP, WEBP, RAW, ICO.
- Caratteristiche minime: 640x480 pixel.
- Dimensione massima: 4 MB.

6.2 Cloud Vision API

Per ogni immagine all'interno di una richiesta, è possibile specificare uno o più tipi di metodi (*features*) corrispondenti alle azioni desiderate.

LABEL_DETECTION Rileva elementi all'interno di un'ampia gamma di categorie che spaziano da animali, trasporti, eccetera. Per ogni elemento rilevato viene associato un punteggio che indica il grado di affidabilità che quella categoria rispecchi veramente l'elemento.

FACE_DETECTION Rileva i volti presenti nell'immagine e restituisce un insieme di metadati (per ogni volto) che includono:

- le coordinate di due poligoni, uno che circonda tutta la testa mentre un altro che circonda solamente il viso (la parte frontale della testa ricompera di pelle),
- le coordinate per un insieme di punti di riferimento del viso, fra cui occhi, orecchie, sopracciglia, labbra, naso, bocca[12].
- il grado di rotazione del volto, espresso nella forma degli angoli di Tait-Bryan (imbardata, rollio e beccheggio),
- il grado di affidabilità (che l'elemento rilevato sia effettivamente un volto),
- il grado di affidabilità dei punti di riferimento,
- le probabilità che il volto esprima gioia, tristezza, rabbia, sorpresa e che sia presente un cappello,
- le probabilità che l'immagine sia sottoesposta o sfuocata.

Le probabilità sono espresse secondo una scala a sei valori: UNKNOWN, VERY_UNLIKELY, UNLIKELY, POSSIBLE, LIKELY e VERY_LIKELY.

TEXT_DETECTION Effettua le stesse funzioni di un OCR (Optical Character Recognition): riconosce caratteri e parole all'interno dell'immagine. Restituisce la lingua del testo rilevato, il testo e le coordinate dei poligoni, uno per l'intera frase e uno per ogni parola che la costituisce.

[11] Viene considerato solo il primo frame.

[12] Per la lista completa vedere `https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#Landmark`.

DOCUMENT_TEXT_DETECTION Esegue la stessa funzione del metodo precedente, ma è ottimizzato per immagini con molto testo. L'oggetto restituito da questo metodo è costituito dalla struttura: Page→Block→Paragraph→Word→Symbol.

Una pagina contiene il linguaggio del testo al suo interno, l'altezza, la larghezza e una lista di blocchi. Un blocco contiene la lingua, le coordinate del poligono che racchiude il blocco, il tipo di blocco e una lista di paragrafi; i tipi sono: sconosciuto, testo, tabella, figura, linea o codice a barre. Il paragrafo include il testo che a sua volta contiene le singole parole che contengono i simboli.

LANDMARK_DETECTION Questo metodo permette il riconoscimento di famose elementi naturali e artificiali, come ad esempio la Torre Eiffel, il "Paris Hotel and Casino" a Las Vegas, la Fontana di Trevi. Il valore ritornato contiene, oltre ai soliti elementi, le coordinate latitudine/longitudine.

LOGO_DETECTION Riconosce loghi e marchi di vari prodotti comuni o famosi.

SAFE_SEARCH_DETECTION Rileva contenti non adatti ai minori e restituisce la probabilità che l'immagine contenga contenuti per adulti, relativi a un ambito medico, violenti o che sia una modifica di un'immagine originale con scopi ludici o offensivi. I livelli di probabilità sono espressi secondo la scala a sei gradi illustrata precedentemente.

IMAGE_PROPERTIES Restituisce una lista di colori dominanti; per ogni colore è presente la sua descrizione in formato RGB, un valore di punteggio e il numero di pixel di quel colore presenti nell'immagine (in rapporto al totale).

CROP_HINTS Suggerisce i punti dove meglio ritagliare l'immagine. Viene restituito un valore di importanceFraction, che indica il rapporto fra l'"importanza" dell'immagine ritagliata e l'immagine originale.

WEB_DETECTION Restituisce informazioni presenti nel web rilevanti per l'immagine:

- webEntities: deduce elementi dell'immagine da immagini simili nel web;
- fullMatchingImages: immagini presenti nel web molto simili a quella di partenza (spesso sono copie);
- partialMatchingImages: immagini presenti nel web che presentano elementi chiave in comune con quella di partenza;
- pagesWithMatchingImages: pagine web che contengono immagine simili a quella di partenza.
- visuallySimilarImages: i risultati delle immagini visivamente simili.

Le funzioni LANDMARK_DETECTION, CROP_HINTS e WEB_DETECTION sono in versione beta e potrebbero variare nel tempo. Non è consigliabile l'utilizzo applicativo e/o in applicazione critiche.

6.3 Tariffe

La Cloud Vision API fornisce diversi metodi per analizzare un'immagine. Per ogni metodo corrisponde una tariffa che viene applicata a ogni utilizzo su un'immagine, chiamato *unità*; se, ad esempio, due metodi sono applicati alla stessa immagine allora ai fini tariffari verranno conteggiati separatamente. Il prezzo è determinato dal numero di unità per mese e il costo per ogni metodo è indicato per 1000 unità/mese in dollari, come indicato in tabella 3.

Metodo	1-1K unità/mese	1K-1M unità/mese	1M-5M unità/mese	5M-20M unità/mese
Label Detection	Gratis	1,50	1,50	1,00
OCR	Gratis	1,50	1,50	0,60
Explicit Content Detection	Gratis	1,50	1,50	0,60
Facial Detection	Gratis	1,50	1,50	0,60
Landmark Detection	Gratis	1,50	1,50	0,60
Logo Detection	Gratis	1,50	1,50	0,60
Image Properties	Gratis	1,50	1,50	0,60
Web Detection	Gratis	3,50	3,50	n.d.
Web Detection	Gratis	3,50	3,50	n.d.

Tabella 3: Tariffe per la Cloud Vision API.

7 Esempi

Verranno presentati ora alcuni esempi di utilizzo delle API con l'obbiettivo di eseguire un confronto il più imparziale possibile. Sono state quindi individuate alcune funzionalità comuni a tutte le piattaforme: riconoscimento oggetti e ambientazione, riconoscimento di volti (singolo e con più volti).

7.1 Riconoscimento oggetti e ambientazione

Data l'immagine in Figura 1, dopo aver interrogato le relative API per il riconscimento di oggetti/tagging, i risultati sono stati riassunti in Tabella 4; in questa si possono osservare le categorie o etichette sotto le quali gli oggetti sono stati identificati e il loro livello di affidabilità. Inoltre, utilizzando le API di Microsoft C.S. è stata generata anche una descrizione.

Figura 1: Immagine utilizzata come riferimento in questo confronto.

7.2 Riconoscimento di un volto

In questo confronto lo scopo era quello di verificare le caratteristiche del volto (*landmark*) che l'API era in grado di riconsoscere in un ambiente semplice; per questo motivo è stata scelta un'immagine di un primo piano di una ragazza[13]. Come si può notare dai risultati in Figura 2, tutte le API hanno riconosciuto e identificato correttamente il volto e, a parte il servizio offerto

[13]I diritti dell'immagine sono del legittimi proprietario.

	Microsoft C.S	IBM W.S.	Amazon A. I.	Google C.M.L.S.
Etichette	plane (97,41%) indoor (96,20%) floor (96,10%) airplane (91,19%) airport (91,11%) aircraft (72,66%) transport (67,56%)	hangar (97,9%) blue color (85,9%) steel blue color (75,5%)	Hangar (95,74%) Aircraft (89,96%) Airplane (89,96%) Warplane (66,30%) Jet (57,09%) Landing (52,80%)	Airliner (96%) Airline (95%) Airplane (95%) Vehicle (91%) Air travel (90%) Aircraft (87%) Aviation 85%)
Descrizione	a large airplane at an airport	-	-	-

Tabella 4: Tabella riassuntiva per il riconscimento oggetti e ambientazione.

da IBM, diversi elementi di questo. Il migliore di questi sembrerebbe essere le API di Google che, tuttavia, non fornisce indicazioni sulla persona specifica (età, sesso).

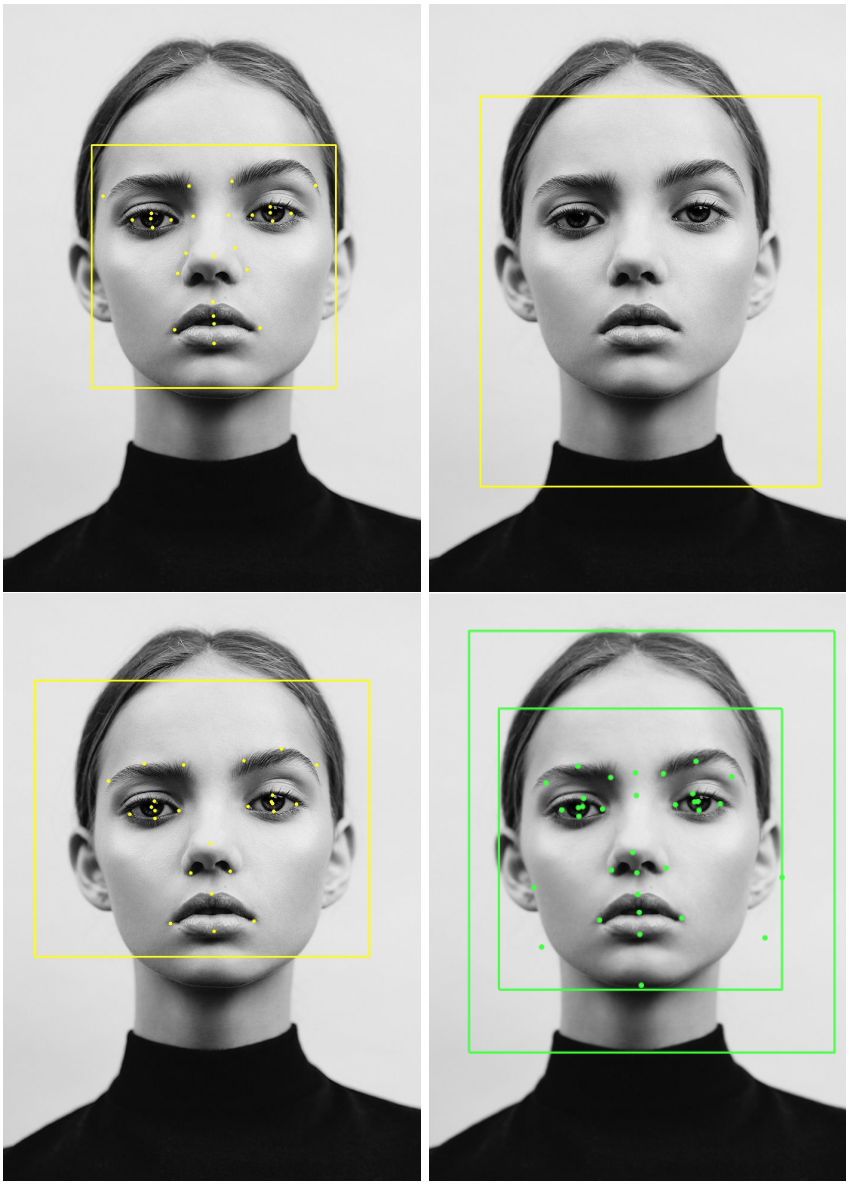

Figura 2: Riconscimento di un singolo volto utilizzando le API di (da sinistra): Microsoft, IBM, Amazon, Google. (Fonte: `http://eddienew.com`)

7.3 Riconoscimento di più volti

Come si vede dalla Figura 3, quasi tutti i volti nella foto sono stati riconosciuti. Da notare come le API Amazon riescano anche a determinare l'orientamento dei singoli volti.

Figura 3: Riconscimento di più singolo volto utilizzando le API di (da sinistra): Microsoft, IBM, Amazon, Google.

8 Applicazioni reali

I grafici che seguono rappresentano i modelli tariffari scelti dalle varie piattaforme e mostrano il loro comportamento al variare del "carico di lavoro". Nell'asse delle ascisse si trova, quindi, il numero di immagini (numero di chiamate con un'immagine ogni chiamata) processate al mese, mentre nell'asse delle ordinate il corrispondente costo in dollari americani[14]. E' stata scelta questa valuta poiché tutti i servizi coprono il territorio americano (almeno in parte), mentre non è così per altre zone come l'europa; una valuta come il dollaro, quindi, offre un metro di paragone fra tutti i servizi.

Il primo grafico in Figura 4 mostra le differenze di prezzo per riconoscimento di oggetti con un basso carico di immagini, includendo anche i piani gratuiti. Come si può notare per poche immagini il sevizio offerto da Google non è conveniente, infatti offre gratuitamente solo 1000 chiamate al mese. Bisogna però ricordare che l'IBM Visual Recognition offre mensilmente il maggior numero di chiamate ma il conteggio è giornaliero. Rimuovendo il piano gratuito iniziale, le tariffe di Google e Microsoft si equivalgono mentre quella più conveniente sembra essere quello di Amazon, come si evince dal grafico in Figura 5. Gli ultimi due grafici, invece, mostrano le tariffe per scenari più reali, con carichi di lavoro dai 50 alle 150 milioni di immagini

[14] Regione: West US

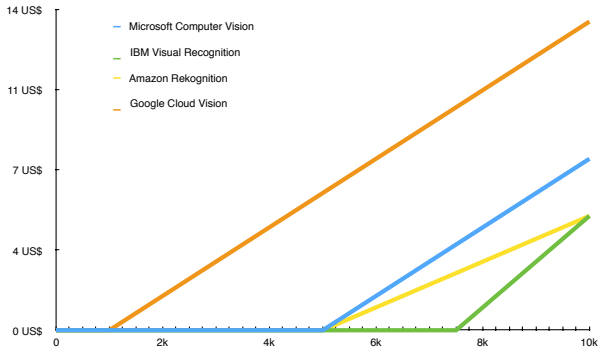

Figura 4: Riconoscimento oggetti con piano gratuito (da 0 a 10K di immagini)

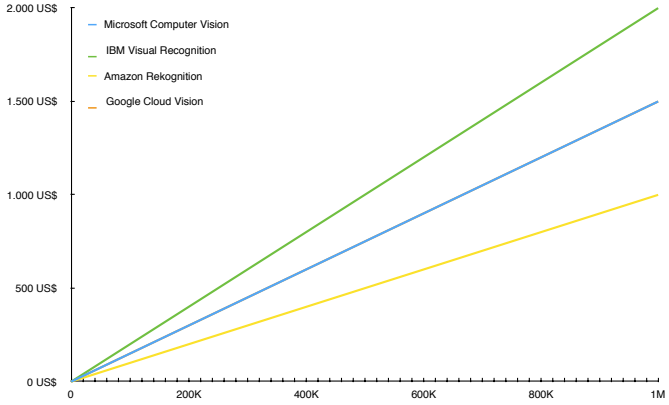

Figura 5: Riconoscimento oggetti senza piano gratuito (da 0 a 1M di immagini)

analizzate al mese. Il primo (Figura 6) rappresenta sempre il riconoscimento di oggetti mentre il secondo (Figura 7) il riconoscimento dei volti. Questo perché alcuni prevedono costi differenti per il riconoscimento di oggetti e di volti, come ad esempio IBM, dove il costo per la seconda è il doppio, e Google, quando il numero di immagini al mese superano i cinque milioni.

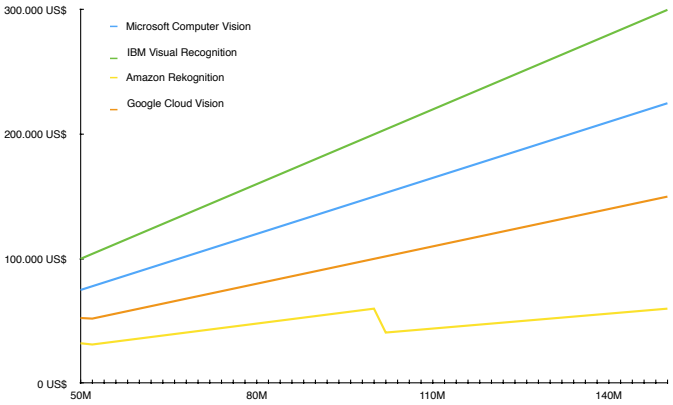

Figura 6: Riconoscimento oggetti (da 50M a 150M di immagini)

Ogni scenario è da considerarsi indipendente dagli altri e rappresenta la stima nel caso peggiore, basandosi su quando scritto nelle relative documentazioni; per applicazioni concrete, infatti, contattando i fornitori di servizi si potrebbero ottenere tariffe più agevolate. Inoltre, i

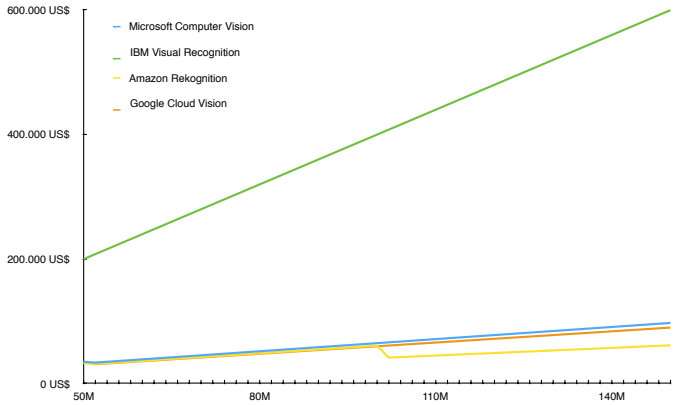

Figura 7: Riconoscimento volti (da 50M a 150M di immagini)

costi sono stati calcolati tenendo conto solo delle API analizzate in precedenza, escludendo costi aggiungivi come ad esempio il costo dell'archiviazione.

9 Conclusioni

In questo lavoro è stato possibile osservare le caratteristiche salienti dell'analisi visiva nell'ambito dei servizi cognitivi e le possibilità offerte da alcuni fra i maggiori fornitori di queste. Grazie a questo è possibile, quindi, conoscere i progressi effettuati fino a questo punto, notare le mancanze che andranno colmate e verso dove si stanno muovendo i prossimi passi.

È opinione dell'autore che i servizi disponibili a tutt'oggi (perlomeno quelli analizzati) diano buoni, e in alcuni casi anche ottimi, risultati e che permettano di conseguenza la realizzazione di applicazioni in grado di ineragire con l'ambiente reale con un buon grado di affidabilità (per sistemi non critici).

9.1 Sviluppi futuri

Inizialmente si potrebbe includere anche l'analisi di video (che non è stata coperta in questo lavoro). Il passo successivo sarebbe sicuramente lo studio delle altre macro-aree, come per esempio quella del linguaggio o della ricerca. Nonostante siano stati inclusi alcuni esempi di utilizzo della API, questi sono molto generici e danno solo un'idea di massima dei servizi. Si potrebbe, quindi, includere lo studio di un'applicazione reale (o il più possibile reale), in modo da osservare il comportamente delle API con un bisogno e problema effettivo.

A Tabelle riassuntive

Funzionalità		Microsoft Vision	IBM Visual Recognition	Amazon Rekognition	Google Cloud Vision
Riconoscimento	oggetti	✓	✓	✓	✓
	scene	✓	✓	✓	✓
	colori	✓			✓
	tipo immagine	✓			
	volti	✓		✓	✓
	celebrità	✓	✓	✓	
	loghi	✓	✓		✓
	punti interesse	✓			✓
	testo	✓			✓
	nudità/violenza				✓
Generazione	descrizioni	✓			
Rielaborazione	immagini	anteprime			ritagli
Gestione	classificatore		✓	solo volti	
Ricerca	immagini	solo volti	✓	solo volti	sul Web
Confronto	immagini	solo volti	✓	solo volti	sul Web

Tabella 5: Riassunto delle funzionalità.

	Microsoft Vision	IBM Visual Recognition	Amazon Rekognition	Google Cloud Vision
Metodi di input	dati raw URL	dati raw URL	dati raw AS3O	dati raw GCL URIs
Formati supportati	JPEG PNG GIF BMP	JPEG PNG	JPEG PNG	JPEG PNG(8—24) (Animated) GIF BMP WEBP RAW ICO
Dimensione minima [pixel]	50x50	224x224	80x80	640x480
Dimensione massima [MB]	4	2	5,15 (AS3O)	4

Tabella 6: Requisiti delle immagini fornite alle API.

Riferimenti bibliografici

[1] Microsoft Corporation, "Microsoft Cognitive Services." https://www.microsoft.com/cognitive-services/en-us/, 2017 (accessed March 22, 2017).

[2] IBM, "IBM Watson Services." https://www.ibm.com/watson/developercloud/services-catalog.html, 2017 (accessed March 30, 2017).

[3] Amazon.com, Inc., "Amazon Artificial Intelligence." https://aws.amazon.com/en/amazon-ai/, 2017 (accessed March 30, 2017).

[4] Google Inc., "Google Cloud Machine Learning Services." https://cloud.google.com/products/machine-learning/, 2017 (accessed April 10, 2017).

[5] Microsoft Corporation, "Computer Vision API (Version 1.0) Documentation." https://www.microsoft.com/cognitive-services/en-us/Computer-Vision-API/documentation, 2017 (accessed March 30, 2017).

[6] Microsoft Corporation, "Image Moderation API Documentation." https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/image-moderation-api, 2017 (accessed April 21, 2017).

[7] Microsoft Corporation, "Emotion API Documentation." https://docs.microsoft.com/en-us/azure/cognitive-services/emotion/home, 2017 (accessed April 21, 2017).

[8] IBM, "IBM Visual Recognition API Reference." https://www.ibm.com/watson/developercloud/visual-recognition/api/v3/, 2017 (accessed March 30, 2017).

[9] Amazon.com, Inc., "Amazon Rekognition Developer Guide." http://docs.aws.amazon.com/rekognition/latest/dg/what-is.html, 2017 (accessed March 30, 2017).

[10] Google Inc., "Google Cloud Vision API Documentation." https://cloud.google.com/vision/docs/, 2017 (accessed April 14, 2017).