

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI MATEMATICA

CORSO DI LAUREA IN INFORMATICA

---

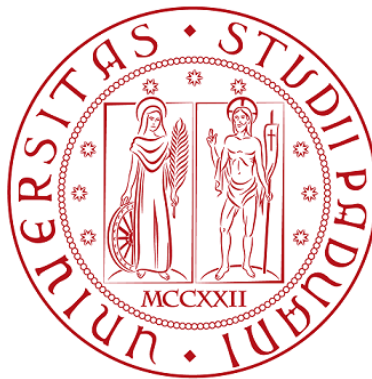
# Riconoscimento e tracciamento di elementi su video ad alta risoluzione

---

*Laureando:*  
Davide LIU

*Relatore:*  
Prof. Lamberto BALLAN

*Tutor aziendale:*  
Leonardo DAL ZOVO



Anno Accademico 2018/2019

---

# Indice

<b>1</b>	<b>Introduzione</b>	<b>4</b>
<b>2</b>	<b>Ambiente Aziendale</b>	<b>5</b>
<b>3</b>	<b>Progetto</b>	<b>6</b>
<b>4</b>	<b>Analisi dei problemi</b>	<b>6</b>
4.1	Tecniche di computer vision . . . . .	6
4.2	Computer vision applicata su immagini ad alta risoluzione . . . . .	7
4.3	Frammentazione dell'immagine . . . . .	8
<b>5</b>	<b>Tecnologie</b>	<b>9</b>
<b>6</b>	<b>Sviluppo</b>	<b>9</b>
<b>7</b>	<b>Risultati ottenuti</b>	<b>9</b>
<b>8</b>	<b>Glossario</b>	<b>9</b>
<b>9</b>	<b>Bibliografia</b>	<b>9</b>
<b>10</b>	<b>Appendice</b>	<b>9</b>

---

## Elenco delle tabelle

---

## Elenco delle figure

1	Logo di Studiomapp . . . . .	5
2	Esempio di un'immagine con label, categoria e probabilità per ogni elemento riconosciuto in essa . . . . .	6
3	Esempio di un frame di un video in 4K . . . . .	8

---

# 1 Introduzione

La computer vision è un ambito dell'intelligenza artificiale il cui scopo è quello di insegnare alle macchine non solo a vedere un'immagine, ma anche a riconoscere gli elementi che la compongono in modo da poter interpretare il suo contenuto come farebbe il cervello di un qualsiasi essere vivente. Nonostante le attuali tecniche di deep learning rendano possibile questo compito, è comunque necessaria una grande quantità di immagini e di tempo per poter allenare una rete neurale a sufficienza in modo da riuscire a riconoscere correttamente degli oggetti in un'immagine non incontrata durante il processo di allenamento.

Lo scopo del progetto di stage è stato quello di progettare e realizzare un sistema di riconoscimento e tracciamento di specifici elementi all'interno di un video ad alta risoluzione. Questo progetto ha comportato sfide e complessità aggiuntive rispetto all'analisi degli elementi presenti in una singola foto sia per il fatto che un video è composto da una sequenza di frames anziché da una singola immagine, sia per il fatto che i frames trattati erano in formato Ultra High Definition (4K) e quindi elaborare l'intero frame in una sola volta sarebbe stato troppo oneroso dal punto di vista computazionale.

Riassumendo i tre problemi principali che sono stati affrontati, in ordine sequenziale, sono i seguenti:

- Frammentazione dell'immagine;
- Tracciamento degli elementi;
- Stabilizzazione delle label degli elementi tracciati.

Ognuno dei sotto-problemi è stato analizzato a fondo, ne è stata trovata un'adeguata soluzione ed è stata applicata ai fini di realizzare un prodotto il più performante possibile.

---

## 2 Ambiente Aziendale



Figura 1: Logo di Studiomapp

STUDIOMAPP, fondata a fine 2015, è una startup innovativa con sedi a Ravenna e Roma, in Italia. Sviluppa algoritmi di intelligenza artificiale specifici per Geo-calcolo e dati geo-spaziali in modo da fornire soluzioni innovative per smart cities, mobilità, trasporti e logistica, turismo e beni culturali, immobiliare e real estate, agricoltura, territorio e gestione delle risorse naturali, adattamento ai cambiamenti climatici.

E' la prima startup dell'Emilia Romagna selezionata dall'ESA BIC Lazio, l'incubatore dell'Agenzia Spaziale Europea (ESA) ed è supportata da importanti istituzioni, acceleratori e grandi attori. Membro fondatore dal 2016 della rete Copernicus Academy, si impegna a diffondere i benefici dell'utilizzo dei dati di Osservazione della Terra per la qualità della vita dei cittadini e la competitività delle PMI.

Ha acquisito una solida esperienza nella promozione di Copernicus, il programma europeo di osservazione della Terra, nell'ecosistema Startup condividendo conoscenza e formazione. Ad oggi la società ha organizzato più di 40 eventi che hanno raggiunto più di 1000 persone che vanno dal pubblico generale, agli studenti, alle startup, ai funzionari pubblici, alle autorità locali, alle PMI.

---

## 3 Progetto

## 4 Analisi dei problemi

### 4.1 Tecniche di computer vision

Per riconoscere un singolo elemento in un'immagine viene tipicamente utilizzata una Convolutional Neural Networks (CNN) allenata con grandi quantità di immagini che possono essere tranquillamente reperite in rete già raggruppate in datasets come ad esempio ImageNet. La vera sfida salta fuori non appena ci troviamo a dover identificare nella stessa immagine diversi oggetti appartenenti a categorie diverse, di differenti dimensioni e posizioni e talvolta anche sovrapposti. Questa situazione è molto comune quando osserviamo qualsiasi foto rappresentante il mondo reale. Il risultato ottimale sarebbe quindi di avere una label di dimensioni corrette per ogni oggetto identificato mostrando anche la categoria di appartenenza dell'elemento e la probabilità che la classificazione sia effettivamente quella corretta.

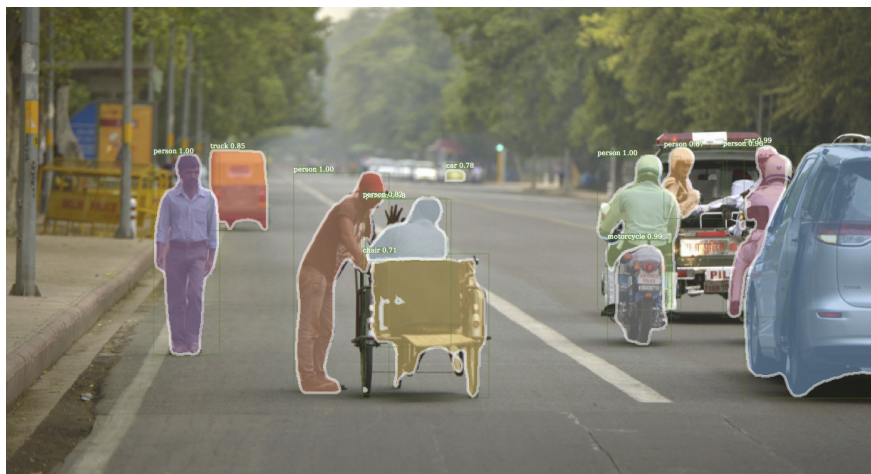


Figura 2: Esempio di un'immagine con label, categoria e probabilità per ogni elemento riconosciuto in essa

Il modo più semplice per andare incontro a questo problema è quello di utilizzare una R-CNN (Regional Convolutional Neural Network) che tramite un algoritmo greedy chiamato Selective Search estrae delle regioni di interesse nelle quali è probabile che sia presente un oggetto. Queste regioni vengono poi date singolarmente in input ad una normale CNN la quale ha il compito di estrarne le caratteristiche principali e poi utilizzare una Support

---

Vector Machine (SVM) presente nell'ultimo strato della CNN per rilevare la presenza di un oggetto ed eventualmente classificarlo. Questo tipo di soluzione presenta il difetto di richiedere molto tempo durante l'operazione di ricerca delle regioni di interesse. La versione più avanzata della R-CNN ed attualmente in uso nei sistemi di computer vision attuali è chiamata Faster R-CNN e risolve il bottleneck della sua antecedente sostituendo la Selective Search con una Region Proposal Network (RPN). Con una CNN viene prima costruita la mappa delle caratteristiche più significative dell'immagine, in seguito vengono passate queste caratteristiche ad una RPN la quale ritorna delle regioni associate ad una probabilità che esse contengano un oggetto. Infine queste regioni vengono passate ad una R-CNN che avrà come al solito il compito di riconoscere e classificare l'oggetto.

## 4.2 Computer vision applicata su immagini ad alta risoluzione

Sebbene sul web sia abbastanza facile reperire grandi quantità di immagini con cui allenare i propri modelli, però la maggior parte di questi datasets contiene solamente immagini a bassa risoluzione ed è difficile trovare in rete grandi datasets di immagini o video in 4K. Inoltre, la maggior parte dei modelli sono stati progettati per lavorare su immagini a bassa risoluzione (tra 200 e 300 pixels) sia per il fatto che una bassa risoluzione è comunque sufficiente per riconoscere e classificare un elemento, sia perchè è più efficiente lavorare su immagini di bassa qualità che su immagini in 4K.

Lo svantaggio che questo comporta è che nelle immagini in bassa risoluzione si perdono molti dei dettagli che invece potrebbero essere catturati da un'immagine o video ad alta risoluzione. In aggiunta, i video in 4K o addirittura 8K sono sempre più diffusi al giorno d'oggi perciò anche gli attuali modelli dovranno prima o poi adattarsi per trattare efficacemente immagini di tali risoluzioni. In figura si può vedere un esempio di un frame in alta risoluzione, diminuendone le dimensioni, perdendo quindi qualità, non sarebbe stato possibile riconoscere alcune delle persone individuate nel frame.





Figura 3: Esempio di un frame di un video in 4K

### 4.3 Frammentazione dell'immagine

Un'immagine in 4K ha una risoluzione di  $3840 \times 2160$  pixels con un totale di 8294400 di pixels. Una rete neurale in grado di ricevere un input così corposo dovrebbe allenare un numero molto elevato di parametri risultando così in un processo molto lungo e dispendioso. L'idea è quella di frammentare l'immagine in diverse sotto-immagini di dimensioni minori e quindi gestibili efficacemente da una singola rete neurale. (FIGURA) Tuttavia questo procedimento non è esente da difficoltà, nel caso in cui un oggetto dovesse trovarsi su più regioni diverse esso potrebbe venire identificato più volte ed essere riconosciuto ogni volta come se fosse un oggetto diverso.

Un primo approccio per risolvere questo problema è stato quello di porre maggiore attenzione alle labels degli oggetti posizionate in prossimità dei confini delle regioni in quanto con molta probabilità è possibile che l'oggetto continui invece nella regione adiacente. Se quindi presente una label anche nella regione adiacente allora si è passati al controllo che le labels possano effettivamente appartenere allo stesso elemento. Per fare ciò è stato sufficiente assicurarsi che le due labels combaciassero entro una certa soglia ed in caso affermativo fonderle in una sola label che contenesse l'oggetto intero.

Un'altra soluzione esaminata è stata quella di suddividere l'immagine intera in regioni con sovrapposizione. In questo modo un elemento giacente a ridosso tra una o più regioni avrebbe generato due o più labels sovrapposte. Tramite un algoritmo di Non-Maximum Suppression, per ogni insieme di labels parzialmente sovrapposte, sono state eliminate le labels con probabilità minore tenendo valida solo quella con probabilità massima.

---

## 5 Tecnologie

## 6 Sviluppo

## 7 Risultati ottenuti

## 8 Glossario

## 9 Bibliografia

Fonti: Sito aziendale <https://www.studiomapp.com/en/> Tecniche di computer vision, fragmentation [https://users.ece.cmu.edu/~franzf/papers/hpec\\_2018\\_vr.pdf](https://users.ece.cmu.edu/~franzf/papers/hpec_2018_vr.pdf) <https://www.mpi-inf.mpg.de/fileadmin/inf/d2/HLCV/cv-ss13-0605-sliding-window.pdf>

## 10 Appendice