

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI MATEMATICA

CORSO DI LAUREA IN INFORMATICA

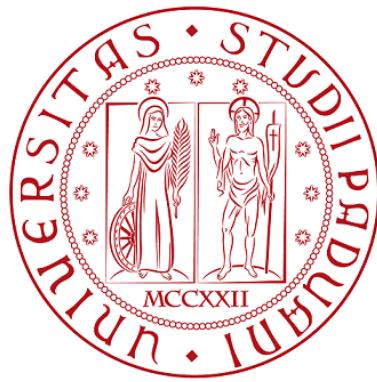
Riconoscimento e tracciamento di elementi su video ad alta risoluzione

Laureando:

Davide LIU

Relatore:
Prof. Lamberto BALLAN

Tutor aziendale:
Leonardo DAL ZOVO



Anno Accademico 2018/2019

Indice

| | |
|--|-----------|
| 1 Introduzione | 5 |
| 1.1 Note esplicative | 5 |
| 2 Ambiente Aziendale | 6 |
| 3 Analisi dei problemi | 7 |
| 3.1 Tecniche di computer vision | 7 |
| 3.2 Computer vision applicata su immagini ad alta risoluzione | 8 |
| 3.3 Frammentazione dell'immagine | 9 |
| 3.4 Tecniche di object tracking | 11 |
| 3.5 Object tracking con continue object detections | 12 |
| 4 Progettazione | 14 |
| 4.1 Progettazione algoritmo per riconoscimento di elementi in un'immagine frammentata | 14 |
| 4.1.1 Scomposizione del frame originale in regioni | 14 |
| 4.1.2 Rimozione degli elementi individuati più volte all'interno delle aree di sovrapposizione | 15 |
| 4.1.3 Creazione di raggruppamenti di labels correlate | 15 |
| 4.1.4 Miglioramento: Raggruppamenti di labels utilizzati come region proposal | 18 |
| 4.2 Progettazione algoritmo per tracciamento di elementi | 19 |
| 4.2.1 Filtro di Kalman | 19 |
| 4.2.2 Assegnazione detection-tracker | 19 |
| 4.2.3 Miglioramento: Utilizzo dell'hash dell'immagine come metrica di supporto | 19 |
| 5 Tecnologie | 20 |
| 5.1 Python | 20 |
| 5.2 Pycharm | 20 |
| 5.3 Tensorflow | 21 |
| 5.4 OpenCV | 21 |
| 5.5 Numpy | 21 |
| 5.6 Matplotlib | 22 |
| 5.7 Pytest | 22 |

| | |
|--|-----------|
| 6 Sviluppo | 23 |
| 6.1 Sviluppo algoritmo per riconoscimento di elementi in un'immagine frammentata | 23 |
| 6.1.1 Implementazione | 23 |
| 6.1.2 Test di integrazione | 25 |
| 7 Risultati ottenuti | 26 |
| 7.1 Metriche utilizzate per singole immagini | 26 |
| 7.1.1 Precision | 26 |
| 7.1.2 Recall | 26 |
| 7.1.3 Intersection over union | 27 |
| 7.1.4 Average Precision | 27 |
| 7.1.5 Mean Average Precision | 28 |
| 8 Glossario | 29 |
| 9 Bibliografia | 31 |
| 10 Appendice | 33 |

Elenco delle tabelle

Elenco delle figure

| | | |
|---|---|----|
| 1 | Logo di Studiomapp | 6 |
| 2 | Esempio di un'immagine con box, categoria e probabilità per ogni elemento riconosciuto in essa | 7 |
| 3 | Esempio di object detection in un frame di un video in 4K | 9 |
| 4 | Esempio di un'immagine in alta risoluzione suddivisa in regioni senza sovrapposizioni | 10 |
| 5 | Frames di un video nei quali viene tracciata un' auto (ordinati da sinistra a destra e dall'alto al basso) | 12 |
| 6 | Esempi di labels erroneamente individuate a causa della frammentazione: nel primo caso la ricostruzione avviene correttamente, nel secondo caso è presente un errore. | 16 |
| 7 | Logo di Python | 20 |
| 8 | Logo di Tensorflow | 21 |
| 9 | Esempio di intersezione e di unione | 27 |

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 umano. 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 anzichè 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 labels 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.

1.1 Note esplicative

Allo scopo di evitare ambiguità a lettori esterni al gruppo, si specifica che all'interno del documento verranno inseriti dei termini con un carattere 'G' come pedice, questo significa che il significato inteso in quella situazione è stato inserito nel Glossario

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 Analisi dei problemi

3.1 Tecniche di computer vision

La computer vision ha come obiettivo quello di riconoscere e classificare alcuni specifici elementi presenti in un'immagine. Identificare un oggetto significa localizzarne la posizione esatta all'interno dell'immagine. Una volta trovata la sua locazione, l'oggetto viene messo in evidenza disegnando un rettangolo attorno ad esso, detto anche bounding box, in modo che lo racchiuda con la maggiore precisione possibile. La classificazione ha invece come scopo quello individuare la categoria di appartenenza di un oggetto e la probabilità che essa sia realmente quella corretta. Per classificare 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 e classificare nella stessa immagine diversi oggetti appartenenti a categorie diverse, di differenti dimensioni e posizioni e talvolta anche sovrapposti. Questa situazione è molto comune quando ci troviamo ad osservare qualsiasi foto rappresentante il mondo reale. Il risultato ottimale sarebbe quindi di avere una bounding box di dimensioni corrette per ogni oggetto identificato mostrando anche la categoria di appartenenza dell'elemento e la sua probabilità.

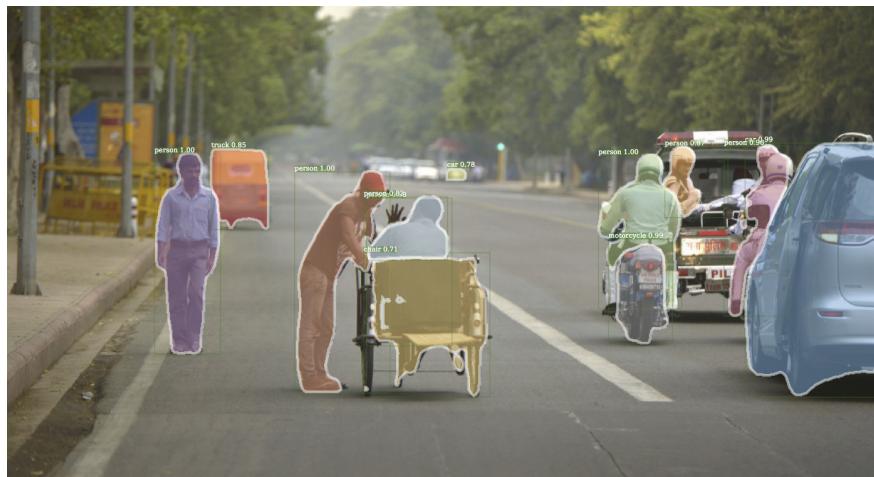


Figura 2: Esempio di un'immagine con box, categoria e probabilità per ogni elemento riconosciuto in essa

Il modo più semplice per andare incontro a questo problema è quello di utilizzare una

R-CNN (Regional Convolutional Neural Network) la quale tramite un algoritmo greedy chiamato Selective Search estrae delle regioni di interesse nelle quali è probabile che vi 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 per 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 Research con una Region Proposal Network (RPN). Essa prende come input un'immagine di qualsiasi dimensione e restituisce come output un insieme di rettangoli associati ad una probabilità che essi contengano un oggetto o meno. Tramite una CNN viene prima costruita la mappa delle caratteristiche più significative dell'immagine, in seguito viene utilizzata una sliding window per scorrere la mappa delle caratteristiche e darle in input a due fully-connected layers dei quali uno serve per individuare le coordinate del box dell'oggetto¹ mentre l'altro serve per ritornare la probabilità che nel box vi sia effettivamente un oggetto². Infine queste regioni vengono passate ad una R-CNN che avrà come al solito il compito di riconoscere e classificare l'oggetto.

3.2 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, tuttavia 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 i 200 e i 600 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 con una risoluzione molto alta.

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 da un video ad alta risoluzione. In aggiunta, i video in 4K o addirittura 8K sono al giorno d'oggi sempre più diffusi perciò anche gli attuali modelli dovranno prima o poi adattarsi per trattare efficacemente immagini di tali risoluzioni. Nella figura sottostante si può vedere un esempio di un frame in alta risoluzione nel quale, diminuendone le dimensioni e perdendo quindi qualità, non sarebbe stato possibile riconoscere alcune delle persone individuate nel frame.

¹Box-regression layer

²Box-classification layer



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

3.3 Frammentazione dell'immagine

Un'immagine in 4K ha una risoluzione di 3840 x 2160 pixels per 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 tanto che molti dei modelli attuali effettuano un ridimensionamento dell'immagine per adattarla al meglio al loro input. L'idea per aggirare il problema è quella di frammentare l'immagine in diverse sotto-immagini di dimensioni minori e quindi gestibili efficacemente da una singola rete neurale. Questa particolare strategia è chiamata frammentazione dell'immagine e risulta essere molto utile in quanto permette di analizzare un'immagine ad alta risoluzione scomponendola in frammenti di dimensioni minori anziché gestirla nella sua interezza.

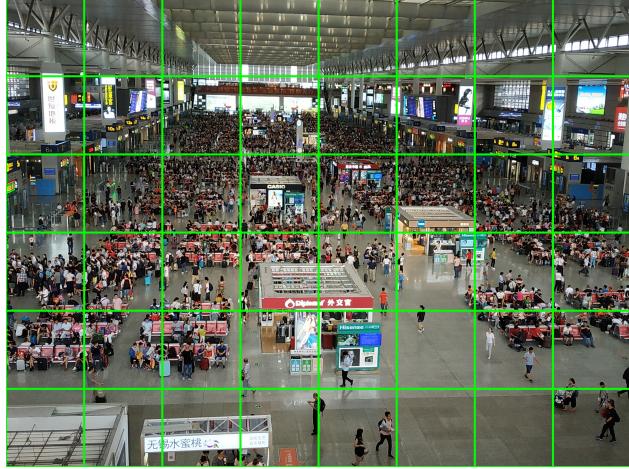


Figura 4: Esempio di un’immagine in alta risoluzione suddivisa in regioni senza sovrapposizioni

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 o addirittura essere riconosciuto ogni volta come se fosse un oggetto appartenente ad una categoria diversa.

Un primo approccio per risolvere questo problema è quello di porre maggiore attenzione alle labels degli oggetti posizionati in prossimità dei confini delle regioni in quanto con molta probabilità è possibile che l’oggetto continui invece nella regione adiacente piuttosto che essere interamente contenuto nella regione esaminata. Se è quindi presente una label anche in una regione adiacente si passa allora alla verifica che le due labels possano effettivamente appartenere allo stesso elemento. Per fare ciò bisogna assicurarsi che le due labels siano compatibili sia in termini di categoria che di posizione entro una certa soglia ed in caso affermativo fonderle in una sola label contenente l’oggetto intero.

Un’altra soluzione esaminata è quella di suddividere l’immagine intera in regioni con sovrapposizione il cui scopo è quello di generare intersezioni tra labels in prossimità dei confini. In questo modo un elemento contenuto all’interno di un’area di sovrapposizione tra più regioni genererebbe due o più labels quasi completamente sovrapposte. Il problema può essere facilmente gestito con un algoritmo di Non-Maximum Suppression, il quale, per ogni insieme di labels parzialmente sovrapposte e con stessa categoria, tende ad eliminare le labels con probabilità minore tenendo valida solo quella con probabilità massima. Tuttavia il caso più insidioso ed allo stesso tempo più frequente avviene quando due o più labels non solo hanno un’intersezione dentro un’area di sovrapposizione ma la loro box si estende anche al di fuori

di esse. E' proprio per far fronte a questo problema che è stato ideato ed implementato un algoritmo il quale verrà descritto nella *sezione 4.1*.

3.4 Tecniche di object tracking

Il tracciamento di oggetti in un video è uno di quei problemi presenti nell'ambito dell'informatica che non sono ancora stati risolti ottenendo risultati soddisfacenti. Il tracciamento consiste non solo nel localizzare l'oggetto tracciato in una sequenza di frames ma anche nel riconoscere che l'oggetto è sempre lo stesso.

Il problema non è per niente banale se pensiamo che in un video uno stesso oggetto può spostarsi, nascondersi dietro qualche altro elemento, deformarsi, cambiare le sue dimensioni, la sua illuminazione, la sua velocità ed il tipo di background. Nel caso ancora peggiore un oggetto tracciato potrebbe addirittura scomparire in un frame per poi ripresentarsi solamente dopo che sono trascorsi un numero casuale di frames.

Un algoritmo di tracking ottimale dovrebbe essere in grado di far fronte a tutti questi problemi riconoscendo quindi l'oggetto da esso tracciato tramite per esempio l'assegnazione di un ID univoco.

Allo stato dell'arte, questo problema viene affrontando eseguendo inizialmente una normale object detection sul primo frame del video in modo tale da individuarne tutti gli elementi presenti e le rispettive labels. In seguito, ognuno di questi elementi viene assegnato ad un tracker che avrà il compito di tracciare l'elemento nei frames successivi.

Tracciare un elemento è un' operazione meno onerosa rispetto alla sua individuazione in quanto il tracker conosce già alcune informazioni relative all'oggetto tracciato acquisite nei frames precedenti potendo così tenere in memoria uno storico del suo stato, come per esempio, le sue ultime locazioni. Per aumentare la sua precisione, un tracker non tiene conto solamente della locazione dell'elemento osservato ma può anche conservare altre utili informazioni aggiuntive come la sua direzione, una previsione delle locazioni future analizzandone la sua traiettoria oppure può memorizzare un hash³ dei pixels presenti all'interno del bounding box dell'oggetto per poi confrontarlo con l'hash dello stesso oggetto calcolato nel frame successivo. I trackers più performanti come CSK, MOSSE e GOTURN utlizzano alcune delle informazioni sopra descritte per costruire un filtro di correlazione in modo da localizzare l'oggetto nel frame successivo e migliorare la precisione del filtro con le successive individuazioni, lo scopo del filtro è quello di minimizzare la differenza tra l'output ricostruito e quello originale.

Nonostante tutti questi accorgimenti è però inevitabile che col trascorrere dei frames i trac-

³Al contrario degli hash usati in crittografia, un hash applicato alle immagini è progettato in modo tale che piccole variazioni dei pixels dell'immagine non risultino in un hash molto diverso da quello originale

kers cominceranno ad essere sempre più imprecisi nell'individuare il loro oggetto tracciato. In particolare, questa situazione può accadere molto velocemente in quei video dove avvengono molti spostamenti e sovrapposizioni tra elementi. E' quindi buona norma aggiornare i trackers con le locazioni corrette effettuando una nuova detection ogni fissato numero di frames.

E' qui che si presenta il problema di maggiore rilevanza: una nuova detection effettuata su un nuovo frame non tiene conto delle informazioni acquisite in precedenza in quanto queste con molta probabilità sarebbero errate o imprecise. Il problema è quindi quello di riassegnare a ciascun oggetto individuato lo stesso ID che possedeva in precedenza ed assegnare un nuovo ID ad ogni oggetto comparso nel video per la prima volta.



Figura 5: Frames di un video nei quali viene tracciata un'auto (ordinati da sinistra a destra e dall'alto al basso)

3.5 Object tracking con continue object detections

Considerata la scarsa affidabilità degli attuali algoritmi di tracciamento, nella soluzione individuata gli oggetti vengono identificati effettuando una nuova detection per ogni frame del video in modo da assicurarsi di avere sempre una buona accuratezza ed allo stesso tempo migliorare l'efficacia dei trackers migliorando gradualmente la precisione dei loro filtri di predizione. Questo metodo sacrifica l'efficienza del processo per migliorarne l'efficacia. A meno che non venga utilizzato un algoritmo di detection molto veloce come SSD (Single Shot Detector) non è possibile applicare il tracking sui video in tempo reale in quanto il frame rate che ne risulterebbe sarebbe troppo basso. A seguito di una nuova detection, per effettuare una corretta riassegnazione degli ID viene fatto un confronto tra le labels presenti nel frame precedente con quello corrente con lo scopo di trovare una corrispondenza biunivoca tra due

labels e capire quando entrambe si riferiscono allo stesso elemento in modo da garantire un corretto trasferimento dell'ID. Per rendere tutto ciò possibile viene utilizzato un filtro in grado di predire le locazioni future di un oggetto tenendo traccia dei suoi bounding boxes e poi correggersi tramite misurazioni successive. Anche per realizzare questa soluzione è stato ideato ed implementato un algoritmo descritto nella *sezione 4.2*.

4 Progettazione

4.1 Progettazione algoritmo per riconoscimento di elementi in un'immagine frammentata

Per quanto riguarda il problema della frammentazione dei frames in 4K è stato individuato un apposito algoritmo in grado di effettuare il riconoscimento degli oggetti nei frames presenti in un video senza doverli ridimensionare ma utilizzando la tecnica della frammentazione dell'immagine.

4.1.1 Scomposizione del frame originale in regioni

Un frame in 4K viene quindi scomposto in una matrice di $R \times C$ sotto-immagini chiamate regione_G in modo tale che ogni regione sia efficacemente analizzabile da un modello come Faster R-CNN. Per facilitare l'operazione di riconoscimento degli elementi da parte della rete, ogni regione si sovrappone leggermente con le sue regioni adiacenti. Per definire la quantità di pixels da coinvolgere nella sovrapposizione viene definito uno stride_G che indica quanti pixels della regione tralasciare, sia in verticale che in orizzontale, prima che cominci quella successiva, ovviamente lo stride deve essere minore della larghezza di una regione. Una Faster R-CNN dopo aver elaborato singolarmente ogni regione come se fosse una singola immagine darà in output una lista di label_G con le seguenti caratteristiche:

- (**max-x, max-y**): coordinate del vertice in alto a sinistra del rettangolo rappresentante la label dell'oggetto riconosciuto;
- (**min-x, min-y**): coordinate del vertice in basso a destra del rettangolo rappresentante la label dell'oggetto riconosciuto;
- **Categoria_G** : è un numero naturale che indica la categoria di appartenenza dell'elemento individuato;
- **Score_G** : rappresenta la misura di probabilità che la classificazione ottenuta sia effettivamente quella corretta.

Successivamente viene aggiustata la posizione delle labels individuate in modo da trasstrarle nella loro posizione corretta all'interno dell'immagine originale non frammentata. Questo viene fatto aggiungendo un adeguato offset alle coordinate dei vertici dei box delle labels sulla base della loro regione di appartenenza.

4.1.2 Rimozione degli elementi individuati più volte all'interno delle aree di sovrapposizione

A causa della presenza delle aree di sovrapposizione dovute alla struttura delle regioni, gli elementi giacenti in queste particolari zone del frame verranno individuati tante volte quante sono le regioni che si sovrappongono in quella determinata area. Per eliminare le copie duplicate e tenerne solo una viene utilizzato un algoritmo chiamato Average Non-Max Suppression (ANMS) che è una variante del Non-Max Suppression tipicamente utilizzato dai modelli di visione artificiale. Invece che tenere la label con lo score maggiore ed eliminare tutte le altre, come box viene calcolato il box che deriva dalla media dei box di tutte labels e lo score viene calcolato come la media dei loro score. Questo metodo è fondato sul ragionamento che non bisognerebbe buttare via delle informazioni già possedute ma piuttosto riutilizzarle per scoprire qualcosa di nuovo. Per esempio ad uno stesso elemento visualizzato dentro due sezioni differenti di un'immagine potrebbero venirgli assegnati due score diversi. Mentre NMS conserverebbe solo il valore più alto tra i due, ANMS li utilizzerebbe entrambi per ottenere un valore ancora più affidabile aumentando quindi la veridicità della classificazione.

4.1.3 Creazione di raggruppamenti di labels correlate

A questo punto tutti gli elementi sono stati individuati e classificati ma rimane comunque il problema che, a causa della precedente scomposizione, gli oggetti situati all'interno delle aree di sovrapposizione risulterebbero individuati due o più volte. Come si può vedere in figura 5, questo numero varia in base al numero di regioni sulle quali giace l'oggetto. Il secondo problema è che due elementi *vicini*⁴, anche se classificati nella stessa categoria, non è detto che necessariamente debbano rappresentare lo stesso elemento. Un esempio di questo caso lo si può sempre notare in figura 5. Un caso ancora peggiore lo si ha quando non solo l'elemento è situato su più regioni differenti ma sussiste anche il problema che ogni parte dell'elemento verrebbe classificata in modo diverso a causa della loro ambiguità. Infine, è anche possibile che un oggetto si distribuisca su molte regioni ed ogni sua label presenta dimensioni diverse. La soluzione individuata consiste nel raggruppare labels correlate tra loro in insiemi di labels dette raggruppamenti_G per poi racchiuderli in una label che identifica l'elemento rappresentato dal raggruppamento. Due labels sono in correlazione tra di loro se soddisfano una condizione di correlazione sotto riportata.

Condizione di correlazione: Per effettuare un corretto raggruppamento delle labels viene anche tenuta in considerazione la categoria a loro associata tramite la classificazione insieme allo score assegnato. Per definire il risultato della condizione è inoltre necessario stabilire una

⁴Due label sono considerate vicine se sono intersecate tra di loro ed intersecano lo stesso confine di regione

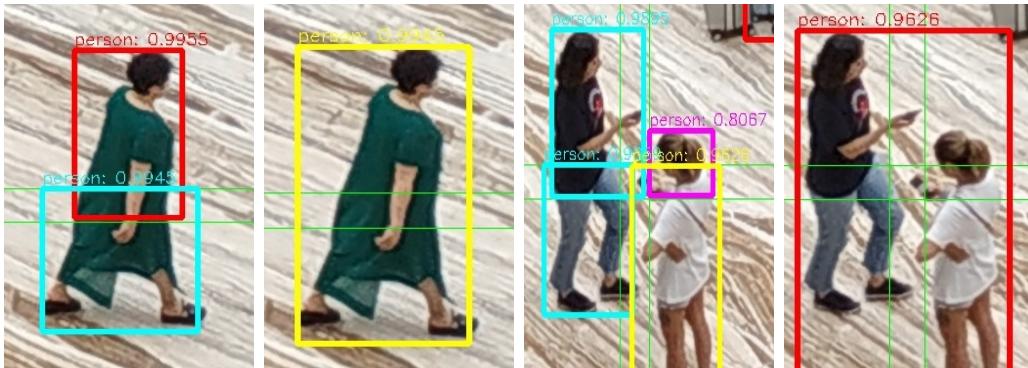


Figura 6: Esempi di labels erroneamente individuate a causa della frammentazione: nel primo caso la ricostruzione avviene correttamente, nel secondo caso è presente un errore.

soglia_G di probabilità per considerare una label come affidabile o meno. Di seguito vengono riportati i vari casi per decidere se la condizione è vera o falsa.

- **True:** Le due labels hanno la stessa categoria ed entrambe con score uguale o maggiore della soglia;
- **True:** Le due labels hanno la stessa categoria ma almeno una delle due ha score minore della soglia;
- **True:** Le due labels hanno categoria diversa ma almeno una delle due ha score minore della soglia;
- **False:** Le due labels hanno categoria diversa ed entrambe con score uguale o maggiore della soglia;

Prima di cominciare con il raggruppamento, vengono inizialmente individuate tutte le labels che intersecano i confini di regioni causati dalle aree di sovrapposizione o che non distino più di un fissato numero di pixels, detto overlap_G, da esse. L'overlap viene definito in quanto anche se nell'immagine reale un oggetto interseca un confine di regione è possibile che a causa di errori di imprecisione del modello, il box della label risulti leggermente distaccato dalla linea che rappresenta il confine. E' quindi possibile ovviare a questo problema aumentando lo spessore del confine di tanti pixels quanti indicati nell'overlap. Per lo stesso motivo precedente, ai fini di controllare se una label interseca un'altra entità o meno viene anche tenuta in considerazione una tolleranza_G che indica quanto il box della label può essere distante da quell'entità affinché questa venga comunque considerata come intersecata. Per creare i raggruppamenti di labels è stato ideato il seguente algoritmo:

-
1. Vengono tenute solo le labels che intersecano almeno un confine di regione e sono inizializzate come *non controllate*;
 2. Viene selezionata una label qualsiasi *non controllata* e la si imposta come *controllata*;
 3. Per ogni label *controllata* ma non ancora *Raggruppata* controlla se ci sono altre labels *non controllate* che rispettino ognuna delle seguenti condizioni:
 - Devono essere *vicine* o distanti entro la tolleranza_G fissata;
 - Devono rispettare la *condizione di correlazione*;
 - La loro box non deve intersecare una regione al di fuori delle aree di sovrapposizione che sia già intersecata da una qualsiasi label *controllata*⁵;
 - Deve essere rispettata una **condizione di matching**: La posizione delle loro aree deve essere compatibile entro una certa soglia ovvero che, i lati lungo il quale le due labels vengono unite siano tali che la differenza tra quello maggiore e quello minore sia inferiore ad una certa soglia che può essere sia definita come proporzione rispetto ad uno dei due lati oppure un valore in pixels.
 4. Le labels così trovate diventano a loro volta *controllate*;
 5. Si ripetono i punti 3 e 4 fino a che non sia più possibile trovare ulteriori labels;
 6. Tutte le label *controllate* vengono ora classificate come *raggruppate* e viene assegnato un numero progressivo ad ogni label *raggruppata* in modo da identificarne il gruppo di appartenenza;
 7. Si ripetono i punti da 2 a 6 fino a che tutte le labels non vengano raggruppate. L'algoritmo in questo modo termina sempre ed è possibile che un raggruppamento comprenda una sola label.

E' da notare che labels intersecanti ma interamente comprese in una sola regione non sono motivo di interesse in quanto l'algoritmo di Non-Maximum Suppression utilizzato dalla rete in fase di post-processing ci assicura che labels intersecanti individuino elementi diversi. In seguito bisogna trasformare ogni raggruppamento in una nuova label che racchiuda tutte le labels che lo compongono. Per fare questo vengono esaminate le coordinate di ogni vertice di tutte le labels di un raggruppamento in modo tale da trovare quattro nuovi vertici di un

⁵Questo perché se due labels sono state individuate come elementi distinti all'interno di una regione allora è probabile che lo siano anche nell'immagine intera in quanto viene supposto che un modello non commetta errori di riconoscimento

rettangolo che soddisfi i requisiti sopra discussi. Il nuovo box così creato andrà a sostituire le labels del rispettivo raggruppamento e per deciderne la categoria e lo score viene applicata una **regola di classificazione**: categoria e score assegnati saranno pari alla categoria e allo score posseduti dalla label con score maggiore.

A questo punto l'algoritmo può dirsi concluso ed è in grado di riconoscere gli elementi in un'immagine in 4K con un'accuratezza accettabile e buona velocità. Tuttavia in casi particolari come quello mostrato in figura figura 5 l'algoritmo commetterebbe un errore di classificazione in quanto individuerebbe due elementi distinti con una sola label comune.

4.1.4 miglioramento: Raggruppamenti di labels utilizzati come region proposal

Un ulteriore miglioramento dell'algoritmo potrebbe essere ottenuto utilizzando le labels ottenute dal procedimento descritto in precedenza come nuove regioni sulle quali applicare nuovamente Faster R-CNN per identificare nuovamente gli elementi contenuti nella regione ma con maggiore precisione in quanto questa volta l'area non verrà affetta da problemi di frammentazione dando quindi la possibilità alla rete di esaminare l'oggetto per intero. La regione viene prima inizializzata rimuovendo la sua label e poi ripopolata con le nuove labels identificate dalla rete. Il primo problema che salta fuori è che durante questo procedimento la rete identificherà nuovamente anche quegli elementi che casualmente si trovavano dentro la regione coinvolta ma che erano già stati trovati anche in precedenza. Tuttavia questo problema viene tranquillamente risolto applicando un algoritmo di Average Non-Max Suppression, utilizzato già in precedenza, per eliminare oggetti quasi completamente sovrapposti. Il secondo problema riguarda ancora gli oggetti che stanno a cavallo tra la regione interessata e l'immagine originale, questa volta però, avendoli già individuati nella loro interezza durante la prima fase è quindi solamente necessario fondere la nuova label con quella già trovata in precedenza. Un caso particolare lo si ha quando la la regione in esame risulti essere così estesa da vanificare i vantaggi ottenuti dalla frammentazione. Per far fronte a questo problema basta ridimensionare l'area coinvolta fino a portarla ad avere dimensioni gestibili da un modello. In questo caso la perdita di risoluzione e quindi di dettagli non comporterebbe un grave problema in quanto gli elementi visibili solo grazie all'alta definizione sono già stati individuati nella fase precedente. Nel caso in cui dovessero venire nuovamente identificati verrebbero gestiti dall'ANMS per ottenerne una migliore approssimazione. Questa funzionalità permette di migliorare l'accuratezza quando si vogliono identificare oggetti che si estendono su due o più regioni o per migliorare il riconoscimento di gruppi di elementi sovrapposti e molto vicini tra loro in prossimità di un confine.

4.2 Progettazione algoritmo per tracciamento di elementi

Per affrontare il problema, viene utilizzato un algoritmo di tracking supportato da una detection applicata ad ogni frame al fine di garantire una migliore accuratezza. Essendo i video in qualità 4K viene sempre utilizzato l'algoritmo descritto in sezione 4.1 in modo da non ridurre la qualità dei frames.

4.2.1 Filtro di Kalman

4.2.2 Assegnazione detection-tracker

4.2.3 miglioramento: Utilizzo dell'hash dell'immagine come metrica di supporto

5 Tecnologie

5.1 Python

Il linguaggio di programmazione utilizzato per perseguire gli obiettivi del progetto è stato Python v3.7, si tratta di un linguaggio di programmazione di alto livello il cui obiettivo è quello di facilitare la leggibilità del codice ed adattarsi a diversi paradigmi di programmazione come quello procedurale, ad oggetti e funzionale. La motivazione per la scelta dell'utilizzo di questo linguaggio, oltre alla sua semplicità, è per il suo supporto di numeri frameworks e moduli relativi al deep learning e alla computer vision tra i quali Tensorflow e OpenCV. Qualsiasi modulo aggiuntiva può essere semplicemente installata eseguendo il comando:

```
pip install nome_modulo
```

Per lanciare un programma viene utilizzato il comando:

```
python nome_file.py
```

Altri moduli che sono stati utilizzati comprendono Numpy, Matplotlib e Pytest.



Figura 7: Logo di Python

5.2 Pycharm

Pycharm è un IDE per programmare in Python sviluppato da JetBrains. Le sue caratteristiche più importanti includono:

- Un sistema intelligente di completamento automatico del codice;
- Analisi statica del codice eseguita a tempo di esecuzione;
- Individuazione e risoluzione veloce degli errori tramite proposte di correzione;
- Possibilità di lavorare in un ambiente di sviluppo virtuale dove per ogni progetto vengono installate solamente le proprie dipendenze e i propri moduli.

5.3 Tensorflow

Tensorflow è un framework gratuito e open-source per lo sviluppo e l'allenamento di modelli di machine learning come le reti neurali. Ha la particolarità che i dati vengono gestiti attraverso dei grafi computazionali dove i nodi rappresentano delle operazioni matematiche da eseguire e gli archi rappresentano degli array multidimensionali contenenti i dati sui quali svolgere le operazioni (tensori). La sua architettura permette di svolgere le operazioni sia usando la CPUs che le GPUs in modo da eseguire operazioni con alto livello di parallelismo.

NON ANCORA USATO



Figura 8: Logo di Tensorflow

5.4 OpenCV

OpenCV è una libreria open-source orientata allo sviluppo di applicazioni di computer vision in tempo reale. E' stata scritta originariamente in C++ ma offre anche il supporto ad altri linguaggi di programmazione come Python. OpenCV è un' ottima libreria quando si ha bisogno di lavorare su immagini e video permettendo di compiere operazioni sia di alto livello che di basso livello operando sui singoli pixels. Sono inoltre presenti diversi algoritmi di object detection ed object tracking già implementati permettendo quindi di sperimentarli tutti senza apportare troppe modifiche al codice esistente. La libreria è stata anche utilizzata per scomporre un video nei suoi singoli frames oltre che per disegnare su di essi.

5.5 Numpy

Numpy è un modulo di Python che fornisce il supporto per la gestione di matrici e array multidimensionali di grandi dimensioni. Dispone anche di una vasta collezione funzioni

matematiche per lavorare ad alto livello su di essi. Viene spesso utilizzato per operare su grandi quantità di dati in modo rapido ed efficiente.

5.6 Matplotlib

Matplotlib è una libreria grafica sviluppata per Python impiegata principalmente per disegnare grafici o altre figure con lo scopo di rappresentare dei dati utilizzando il minor numero di righe di codice possibile. In fase di allenamento di una rete neurale viene spesso usato per mostrare l'andamento della variazione dell'errore e dell'accuratezza.

5.7 Pytest

Pytest è un framework per Python che permette di scrivere dei test per testare il codice permettendone la loro esecuzione in modo automatico. Per installarlo viene utilizzato il comando:

```
pip install pytest
```

mentre per lanciarne l'esecuzione viene usato il comando:

```
pytest nome_file
```

Per convenzione il file contenente i test relativi ad un solo modulo è stato chiamato "nome_modulo_test.py" dove "nome_modulo.py" è il modulo ad esso associato. I metodi che eseguono un test devono iniziare con "test_," , in caso contrario non verranno riconosciuti come tali e la loro esecuzione non verrà lanciata. Sono stati implementati dei test di unità per tutte le funzioni più complesse o critiche in modo da verificarne il corretto funzionamento. Sono anche stati sviluppati dei test di integrazione per ogni algoritmo implementato in modo da assicurarsi del loro corretto funzionamento. Per valutare l'esecuzione dell'intero sistema sono invece state utilizzate delle metriche per misurare la bontà del risultato in quanto i test di sistema non sarebbero stati adatti per misurare un risultato non deterministico come il riconoscimento di oggetti.

6 Sviluppo

6.1 Sviluppo algoritmo per riconoscimento di elementi in un'immagine frammentata

Al fine di implementare l'algoritmo è stata creata una libreria contenente numerose funzioni per lavorare con le labels ritornate dal modello come per esempio ottenerne la posizione o controllare le loro intersezioni con altre entità. Una labels è composta da un array formato da quattro numero interi caratterizzanti il vertice in alto a sinistra e quello in basso a destra del bounding box, un intero per rappresentare la categoria dell'elemento ed infine un numero decimale compreso tra 1 e 0 per indicarne lo score. Viene data la possibilità all'utente di ridefinire la propria *condizione di correlazione*, la *regola di classificazione* e la *condizione di matching* passandole come parametro direttamente nella funzione. Altri parametri passabili includono: le labels, le dimensioni delle regioni e dello stride e la quantità di overlap, sia in verticale che in orizzontale, e di tolleranza in pixels. Alla fine l'algoritmo ritornerà una lista di tutte le labels presenti nell'immagine.

Nella seguente sezione ne viene riportata una sua implementazione.

6.1.1 Implementazione

```
1 from faster_nms import faster_nms
2 import numpy as np
3 import labels as lb
4
5 def process_image_labels(boxes, region_size=(300, 300), stride_size=(270, 270),
6     , overlap=(0, 0), tol=0, condition=None, find_category=None, matching=None):
7     if condition is None:
8         condition = lb.condition
9     if find_category is None:
10        find_category = lb.find_category
11
12     w, h = lb.img_dim_from_boxes(boxes)
13     regions = lb.generate_regions(w, h, region_size, stride_size)
14     boxes = faster_nms(boxes, overlapThresh=0.8)
15
16     # Keeps only the labels intersecting one or more region edges
17     lb.get_intersect_edges_labels(boxes, regions, tol, overlap)
18     # All labels are set as not-checked and not-grouped
19     checked = np.zeros((np.ma.size(ne_boxes, 0)))
20     grouped = np.ones((np.ma.size(ne_boxes, 0))) * (-1)
21     n_group = 0
```

```

21     for i in range(len(ne_boxes)):
22         # Select a label and sets it as checked
23         if not checked[i]:
24             checked[i] = True
25             while True:
26                 found = 0
27                 # If a label is checked then it means it is not-grouped
28                 for j in range(len(ne_boxes)):
29                     if checked[j] and grouped[j] == -1:
30                         for k in range(len(ne_boxes)):
31                             # check for the conditions to be true
32                                 if not checked[k] and lb.intersection(ne_boxes[j, :],
33                                     ne_boxes[k, :], tol) and condition(ne_boxes[j, :], ne_boxes[k, :]) and
34                                     lb.intersect_common_edge(ne_boxes[j, :], ne_boxes[k, :], regions, tol,
35                                     overlap) and lb.matched(ne_boxes[j, :], ne_boxes[k, :]) and lb.matched(
36                                     ne_boxes[j, :], ne_boxes[k, :]) and not lb.belong_same_region_strict_group(
37                                     ne_boxes[k, :], checked_boxes, regions, tol):
38                                     # The new label is set as checked
39                                     checked[k] = True
40                                     found += 1
41                 # The cycle is repeated until no new labels can be found
42                 if found == 0:
43                     break
44             # All the checked labels are set as grouped and is them assigned a
45             # number with the purpose of identify their group ID
46             for j in range(len(ne_boxes)):
47                 if checked[j] and grouped[j] == -1:
48                     grouped[j] = n_group
49             n_group += 1
50             # Now we have that each label only belongs to a single group
51
52             # Now that we have grouped the labels each group is merged into a label
53             # containing all of them
54             new_boxes = np.zeros((0, 6))
55             for i in range(n_group):
56                 # For each group
57                 v = np.zeros((0, 6))
58                 count = 0
59                 for j in range(len(ne_boxes)):
60                     if grouped[j] == i:
61                         v = np.append(v, ne_boxes[j, :])
62                         count += 1
63                 v = v.reshape(count, 6)
64                 # Find the coordinates of the labels containing the whole group
65                 xx1 = np.amin(v, axis=0)[0]

```

```

59     yy1 = np.amin(v, axis=0)[1]
60     xx2 = np.amax(v, axis=0)[2]
61     yy2 = np.amax(v, axis=0)[3]
62     # find_category can be redefined by the user
63     categories, scores = find_category(v)
64     for j in range(len(categories)):
65         new_boxes = np.append(new_boxes, np.array([xx1, yy1, xx2, yy2,
categories[j], scores[j]]))
66     # new_boxes are the labels resulting from grouping algorithm
67     new_boxes = new_boxes.reshape(int(len(new_boxes) / 6), 6)
68
69     # The algorithm is over now thus we can return the results
70     final_boxes = np.append(new_boxes, old_boxes, 0)
71
72     return final_boxes

```

Listing 1: Python example

6.1.2 Test di integrazione

Al fine di verificarne la correttezza sono stati implementati 80 test di integrazione. Essi hanno il compito di testare l'algoritmo su diverse disposizioni e quantità di labels impostando i parametri con differenti valori in modo tale da riconoscerne il maggior numero di errori possibile. Sono state in tutto testate in totale cinque disposizioni di labels differenti con due diversi valori di overlap, tolleranza, stride e dimensione delle regioni arrivando ad un totale 80 test passati tutti con esito positivo.

7 Risultati ottenuti

7.1 Metriche utilizzate per singole immagini

Nell'ambito della computer vision vengono tipicamente utilizzate due tipi di metriche per misurare due diverse proprietà:

- Corretta determinazione della posizione degli oggetti;
- Rilevamento l'esistenza degli oggetti nell'immagine e la loro corretta classificazione.

Le metriche utilizzate per misurare la bontà dei risultati del progetto sono AP (Average Precision) e mAP (mean Average Precision) le quali misurano la seconda proprietà sopra elencata. Prima di iniziare a descrivere le due metriche è necessario definire tre concetti che saranno poi coinvolti per il calcolo dei valori delle metriche.

7.1.1 Precision

La precision misura l'accuratezza delle classificazioni ed indica la percentuale di classificazioni corrette sulla base di quelle totali e viene calcolata come:

$$precision = \frac{TP}{TP + FP}$$

Dove TP (True Positives) è il numero di classificazioni corrette e FP (False Positives) è il numero di classificazioni errate. E' da sottolineare che se uno stesso elemento viene individuato più di una volta, solo la prima volta conterà come TP mentre il resto delle volte conterà come FP. Questa metrica fornisce un'idea sulla correttezza dei risultati.

7.1.2 Recall

La recall misura quanti oggetti sono stati individuati e classificati correttamente sulla base degli oggetti totali, viene calcolata come:

$$precision = \frac{TP}{TP + FN}$$

Dove TP (True Positives) è sempre il numero di classificazioni corrette mentre FN (False Negatives) è il numero di oggetti che non sono stati individuati ma che se sarebbero stato corretto individuare. Questa metrica fornisce un'idea sulla completezza dei risultati.

7.1.3 Intersection over union

L'intersection over union (IoU) misura il grado di sovrapposizione tra due aree e viene usato per misuare il grado di sovrapposizione tra la box dell'oggetto individuato e la box dell'oggetto reale. Viene calcolata come:

$$IoU = \frac{AI}{AU}$$

Dove AI (Area dell'Intersezione) corrisponde all'area dell'intersezione tra le due box e AU (Area dell'Unione) corrisponde all'area dell'unione tra le due box. Una label per essere considerata come corretta deve avere un valore di IoU maggiore di una soglia stabilità (per esempio 0.5).

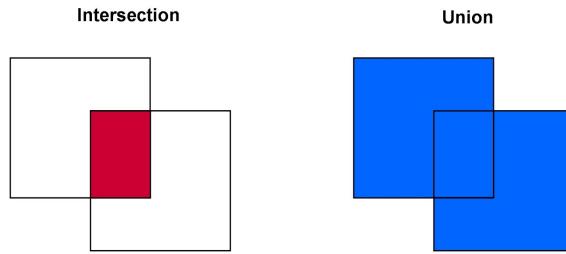


Figura 9: Esempio di intersezione e di unione

7.1.4 Average Precision

Una delle due metriche utilizzate è l'average precision (AP) e per calcolarla si fa la media delle precisioni di undici diversi valori di recall equamente distribuiti. Questa metrica viene applicata ad una sola categoria di elementi e viene calcolata tramite la seguente formula:

$$AP = \frac{1}{11} \sum_{Recall_i} Precision(Recall_i)$$

Dove $i=[0, 0.1, 0.2, \dots, 1.0]$ in quanto $0 < recall < 1$. Inoltre la precisione di uno specifico valore di recall viene calcolata nel seguente modo:

$$Precision(Recall_i) = \max Precision(Recall_j) \quad \text{and} \quad j \geq i$$

L'AP è un valore che riassume la forma della curva precision/recall per una data categoria.

7.1.5 Mean Average Precision

La mean average precision (mAP) è la media delle AP di tutte le categorie calcolata su diverse soglie di IoU:

$$mAP_{IoU=x\%} = \frac{1}{n} \sum_{i=0}^n AP_i$$

Dove i è il numero totale di categorie sulle quali è stata calcolata la propria AP e x rappresenta diverse soglie di IoU.

8 Glossario

B

Box E' un rettangolo che identifica il perimetro entro quale l'oggetto riconosciuto si trova.

C

Categoria L'obiettivo della classificazione è quello di assegnare all'oggetto individuato la sua categoria di appartenenza.

L

Label Etichetta che viene data ad ogni elemento riconosciuto e ne indica la categoria di appartenenza, una bounding box ed uno score.

O

Overlap L'overlap indica lo spessore in pixels dei confini delle regioni. E' usato per sapere se un' entità interseca un confine o meno.

R

Raggruppamento Insieme di labels correlate tra loro tale che da una qualsiasi label del gruppo sia possibile raggiungere qualsiasi altra label dello stesso insieme passando solo per label *vicine*⁶.

Regione Una sezione di un'immagine con un'area di dimensione minore dell'area dell'immagine originale.

S

Score La probabilità che la categoria associata all'elemento riconosciuto sia quella corretta.

⁶Due label sono considerate vicine se sono intersecate tra di loro ed intersecano lo stesso confine di regione

Soglia E' il valore che lo score di una label deve eguagliare o superare per essere considerata affidabile.

Stride E' un attributo da tenere conto in fase di frammentazione di un'immagine ed indica quanti pixels della regione tralasciare, sia in verticale che in orizzontale, prima che cominci quella successiva. Lo stride orizzontale può avere dimensione diversa da quello verticale. La sua applicazione ha come scopo quello di creare delle zone di sovrapposizione tra le varie regioni.

T

Tolleranza La distanza in pixels per la quale una label può discostare da una zona di interesse per cui sia ancora considerata essere intersecata con la zona di interesse.

9 Bibliografia

Riferimenti bibliografici

- [1] Sito web dell'azienda Studiomapp.
<https://www.studiomapp.com/en/>
- [2] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. University of Toronto.
- [3] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. UC Berkeley.
- [4] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. Microsoft Research.
- [5] Vit Ruzicka, Franz Franchetti. *Fast and accurate object detection in high resolution 4K and 8K video using GPUs*. Carnegie Mellon University.
- [6] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, Michael Felsberg *Accurate Scale Estimation for Robust Visual Tracking*. Linköping University.
- [7] David Held, Sebastian Thrun, Silvio Savarese *Learning to Track at 100 FPS with Deep Regression Networks*. Stanford University.
- [8] Imagenet database.
<http://www.image-net.org/>
- [9] Guida all'utilizzo del framework Tensorflow.
<https://www.tensorflow.org/>
- [10] Guida all'utilizzo del modulo matplotlib.
<https://matplotlib.org/>
- [11] Guida all'utilizzo del modulo numpy.
<https://www.numpy.org/>
- [12] Guida all'utilizzo del modulo OpenCV.
<https://www.pyimagesearch.com/>

[13] Studio delle metriche.

[https://medium.com/@jonathan_hui/
map-mean-average-precision-for-object-detection-45c121a31173](https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173)

10 Appendice