UNIVERSITÀ
DEGLI STUDI
FIRENZE

Rilevazione di volti con event camera tramite dati simulati da RGB

Candidato: **Gianmarco Pastore**

Relatore: **Prof. Alberto del Bimbo**

Correlatore: **Prof. Federico Becattini**

UNIVERSITÀ DEGLI STUDI FIRENZE

Scopo del progetto

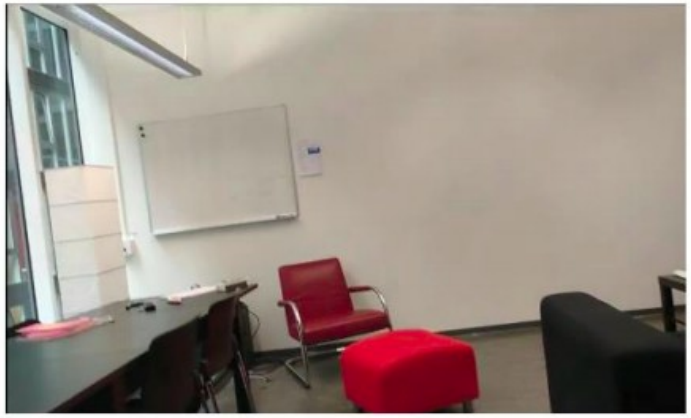
Lo scopo di questo progetto è quello di sviluppare un face detector per video prodotti con event camera.

I dati utilizzati per addestrare il face detector sono video ad eventi “sintetici” ottenuti a seguito di un’opportuna conversione di video *RGB*.

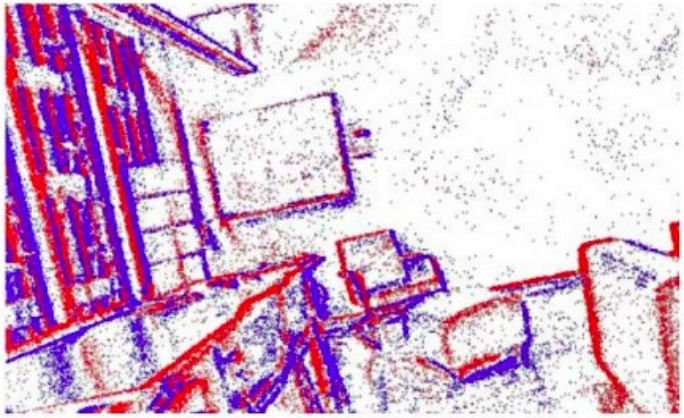
Introduzione sulle event camera

L'event camera è un nuovo tipo di sensore il cui output non è una sequenza di frames, ma uno stream di “eventi” asincroni, che indicano quando i singoli pixel registrano un cambiamento di intensità luminosa.

Standard Camera



Event Camera (ON, OFF events)

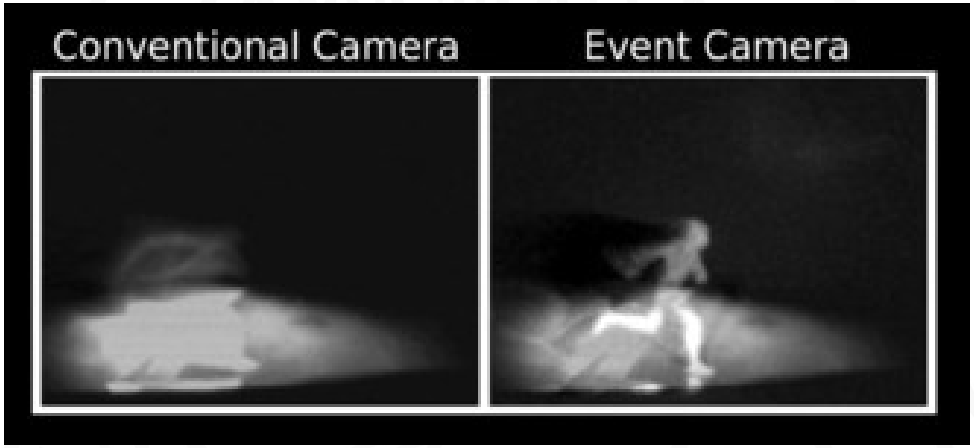


UNIVERSITÀ DEGLI STUDI FIRENZE

Introduzione sulle event camera

Offrono vantaggi significativi rispetto alle telecamere tradizionali:

- Alta risoluzione temporale
- Alta gamma dinamica
- Assenza di motion blur



UNIVERSITÀ DEGLI STUDI FIRENZE

Panoramica del progetto

Durante lo sviluppo di qualsiasi tecnologia che operi con video ad eventi, compreso un face detector, si va incontro a due problemi:

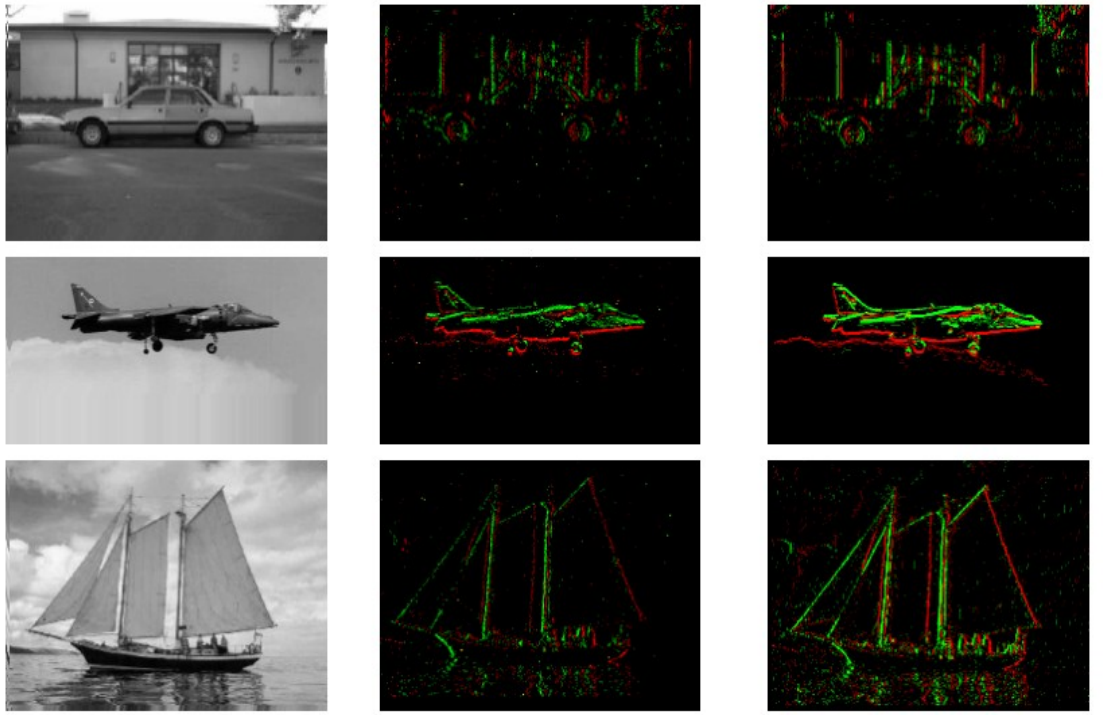
- È richiesta una grande quantità di dati ad eventi che non sono reperibili a causa della novità di questi sensori

- Utilizzare dati generati da telecamere ad eventi è difficile a causa del loro output asincrono e irregolare

UNIVERSITÀ DEGLI STUDI FIRENZE

Panoramica del progetto

- Per compensare la mancanza di dataset è stato usato un simulatore, ESIM, capace di convertire video RGB in eventi “sintetici”.



(a) preview (b) real events (c) synthetic events

UNIVERSITÀ DEGLI STUDI FIRENZE

Panoramica del progetto

- Per quanto riguarda l'operazione di *face detection,* essa è stata eseguita sui video RGB originali.

- Poi le detections sono state trasferite pixel per pixel sui frame sintetici, andando a creare così un dataset ad eventi annotato.

- La fase finale del lavoro è consistita nell'addestramento di YOLOv3, un sistema di real time object detection, sfruttando il dataset ottenuto dalla precedente conversione

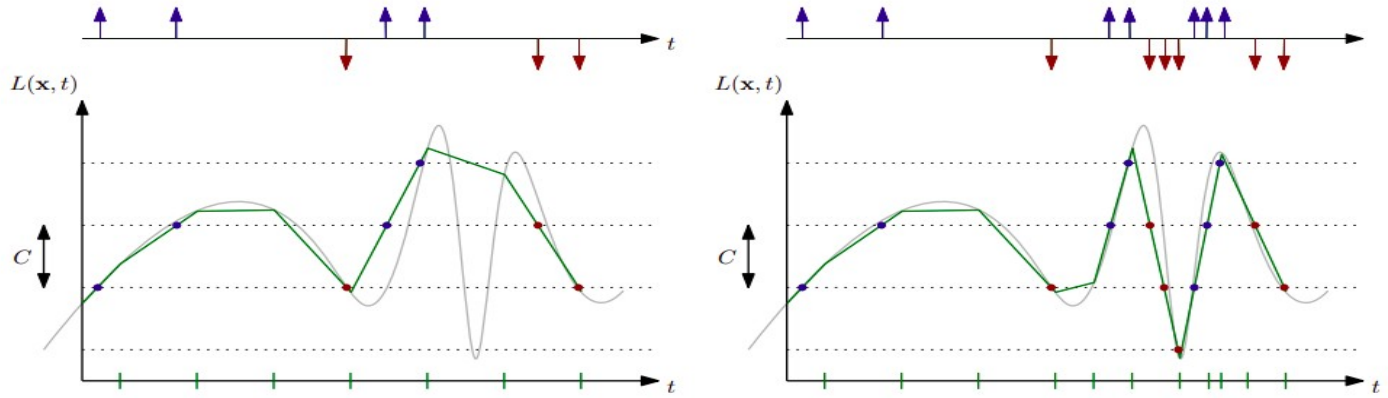
UNIVERSITÀ DEGLI STUDI FIRENZE

Tecnologie impiegate - ESIM

ESIM:

- È un simulatore open source capace di generare grandi quantità di dati ad eventi in maniera affidabile.

- Richiede un *adaptive upsampling* dei dati in input

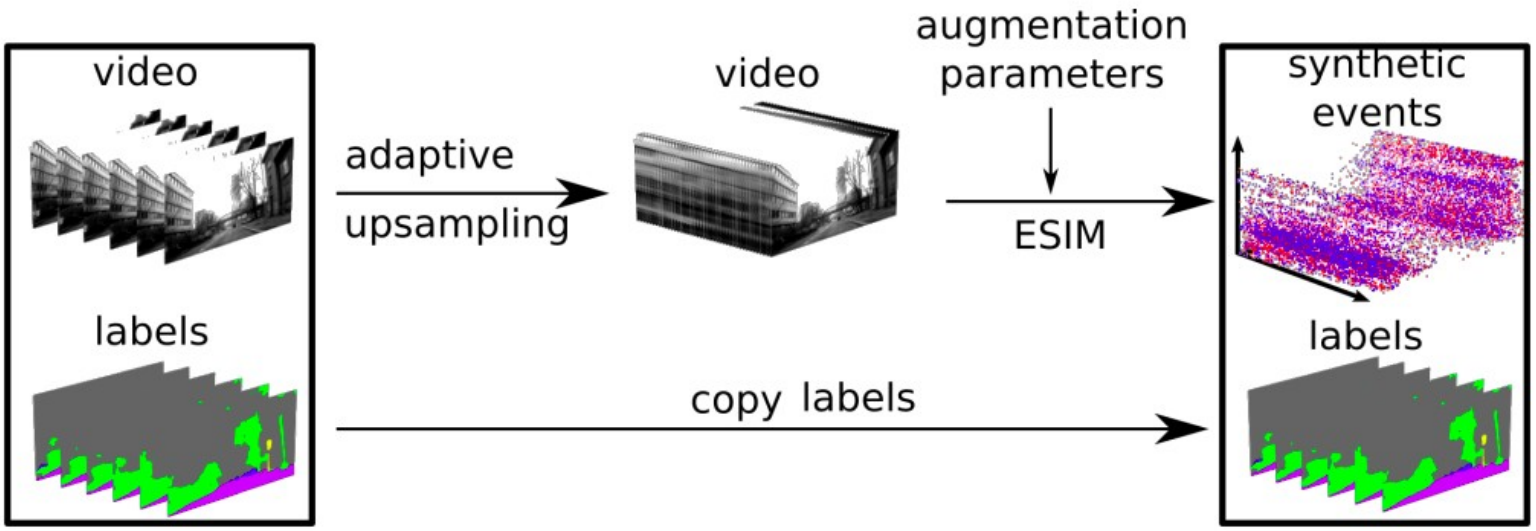


upsampling classico vs adaptive upsampling

Tecnologie impiegate - ESIM

- Mette in atto un'interpolazione lineare dei frames per ricostruire un'approssimazione del segnale visivo continuo

- Usa il segnale approssimato per simulare il comportamento di una event camera.

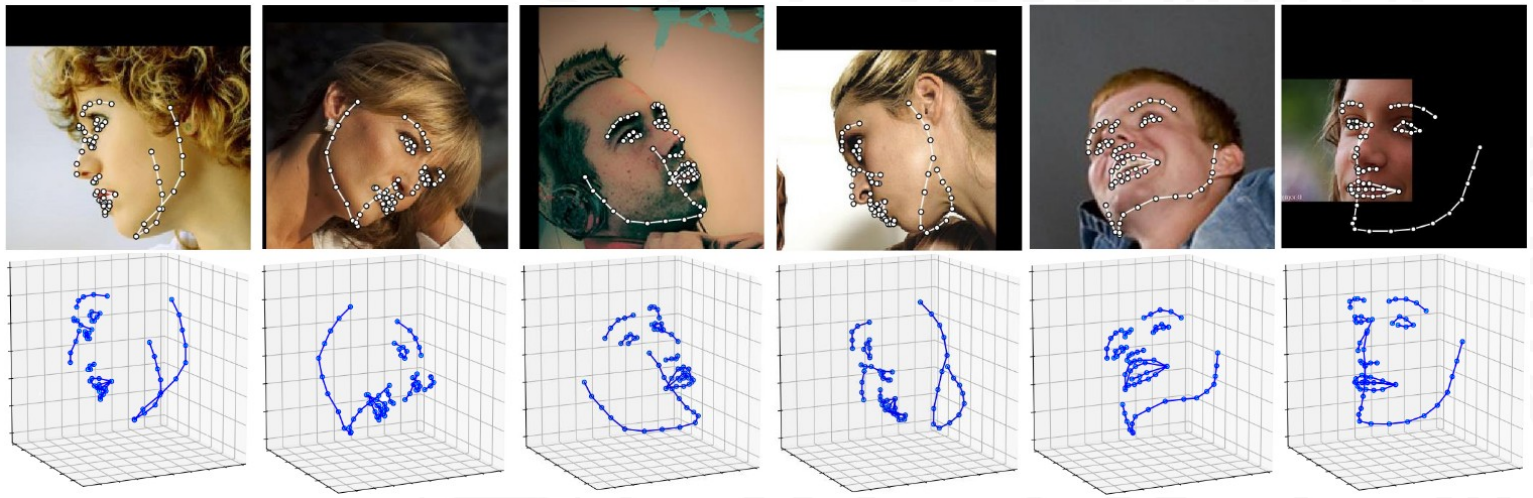


UNIVERSITÀ DEGLI STUDI FIRENZE

Tecnologie impiegate - Face-alignment

Per la rilevazione dei volti, eseguita su dati RGB, è stato utilizzato il framework face-alignment.

È in grado di rilevare punti sia in coordinate 2D che 3D, utilizzando metodi di face-alignment basati sul deep learning.



UNIVERSITÀ DEGLI STUDI FIRENZE

Tecnologie impiegate - YOLOv3

YOLOv3:

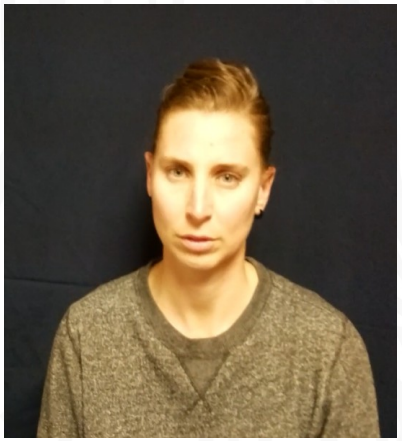
- È il sistema di real-time object detection che è stato allenato con i video ad eventi sintetici.
- Applica una singola rete neurale all'immagine completa
- Lavorando con una sola rete e su tutta l'immagine riesce ad essere più veloce rispetto ad altri modelli simili.

UNIVERSITÀ DEGLI STUDI FIRENZE

Dataset impiegato

Il dataset che è stato utilizzato per il progetto è Tufts-Face-Database.
Contiene 112 video, di 74 femmine e 38 maschi, provenienti da più di 15 paesi con una fascia di età compresa tra i 4 e i 70 anni.

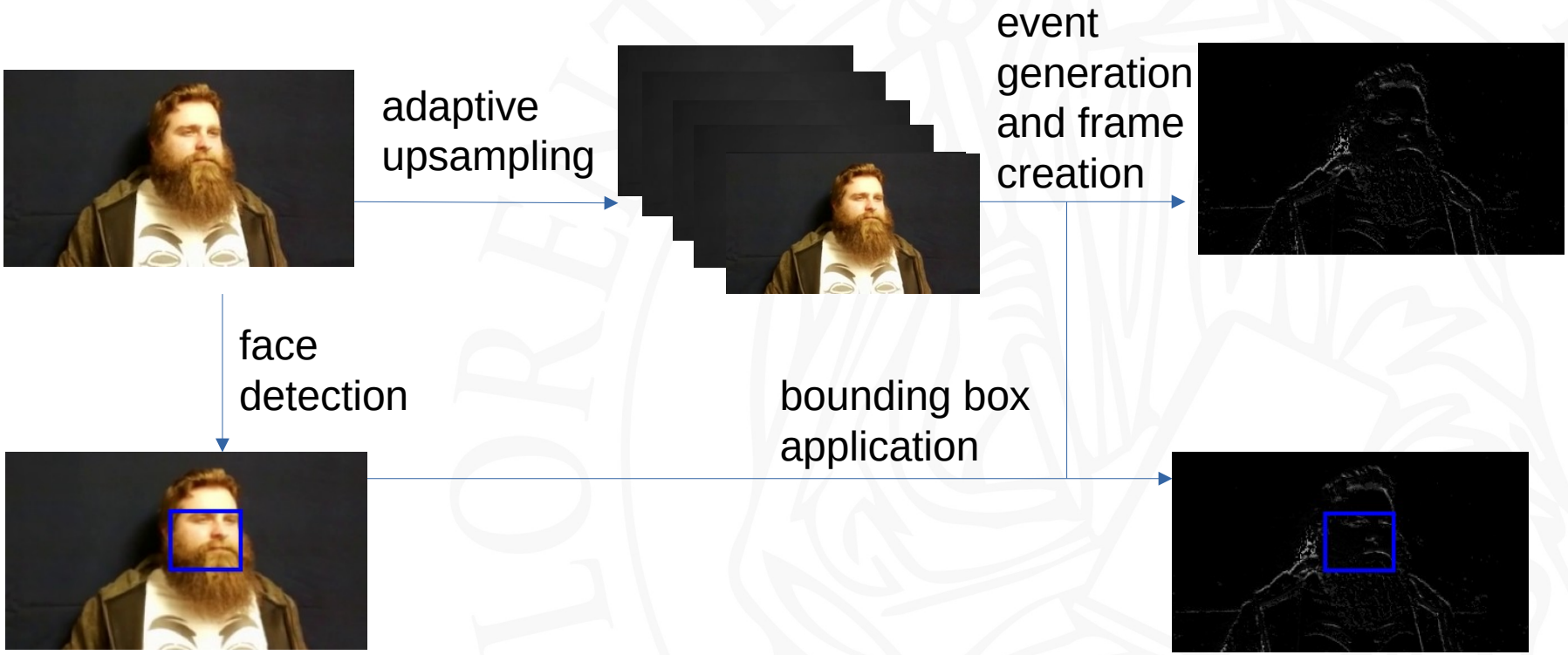


UNIVERSITÀ DEGLI STUDI FIRENZE

Esperimenti e Risultati – Conversione

In figura è mostrato il processo di creazione del dataset annotato.



Esperimenti e Risultati – Conversione

Per la creazione dei frame è stato scelto un accumulation time di **0.001** secondi che ha portato ad avere come risultato circa **18300** frames sintetici per ogni video di partenza.

UNIVERSITÀ DEGLI STUDI FIRENZE

Esperimenti e Risultati – Conversione

Risultati della conversione con trasferimento delle bounding box:

UNIVERSITÀ DEGLI STUDI FIRENZE

Esperimenti e Risultati – Addestramento

Il dataset annotato risultante dalla fase di conversione consiste in un totale di circa 2 milioni di frames, di cui un decimo sono stati usati nell'addestramento.

Inoltre il Il dataset è stato diviso in due parti:

- **training set** - 80% dei dati totali
- **validation set** - 20% restante

Esperimenti e Risultati – Addestramento

YOLOv3 durante l'addestramento fornisce un mAP (mean average precision) tramite cui si può valutare la qualità delle performance del modello allenato.

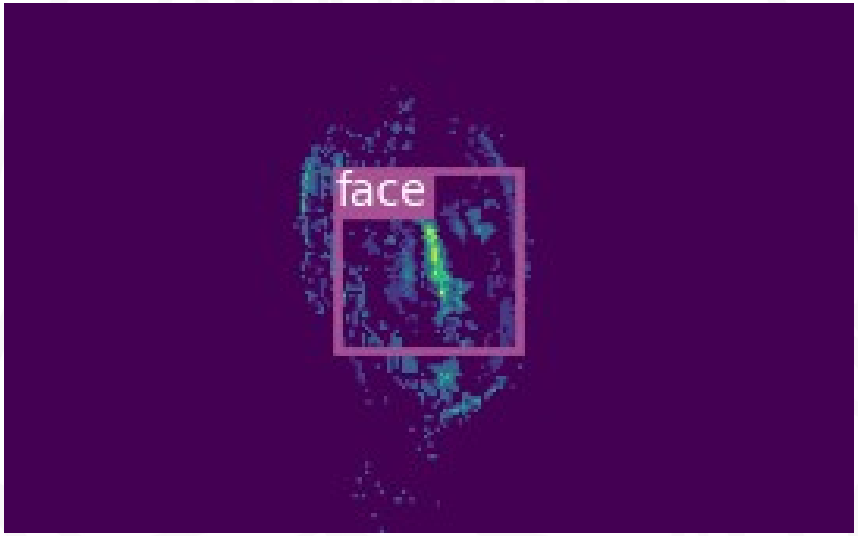
Durante l'addestramento compiuto in questo lavoro è stato ottenuto un mAP massimo del **98%** sul validation set.

Rappresenta un risultato ottimo, superiore anche alle aspettative, che è avvalorato dai risultati qualitativi mostrati nelle prossime slides.

UNIVERSITÀ DEGLI STUDI FIRENZE

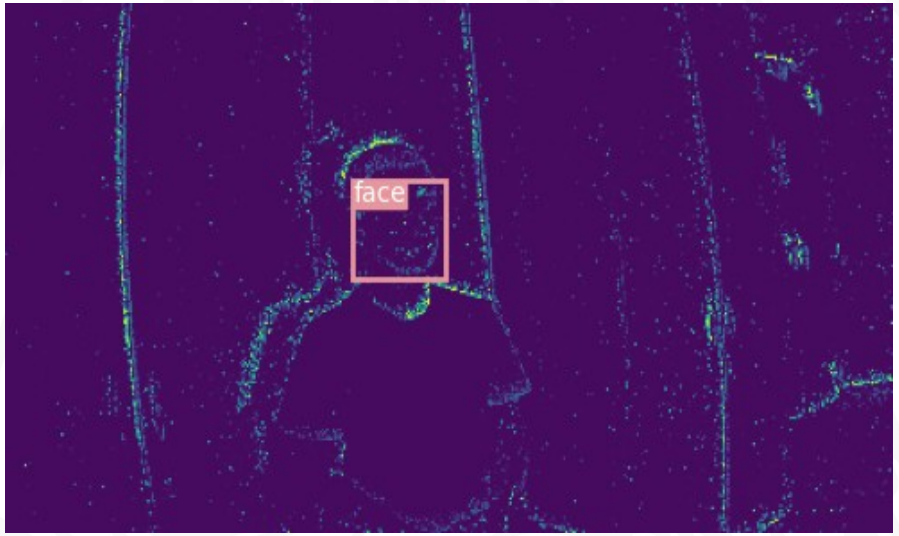
Esperimenti e Risultati – Test

Il face detector ricavato dall'addestramento è stato testato prima su video non appartenenti al dataset e sintetici.



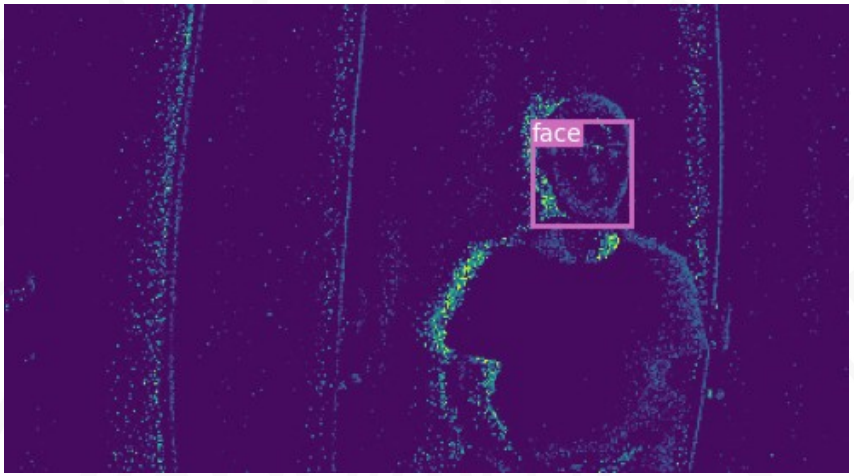
Esperimenti e Risultati – Test

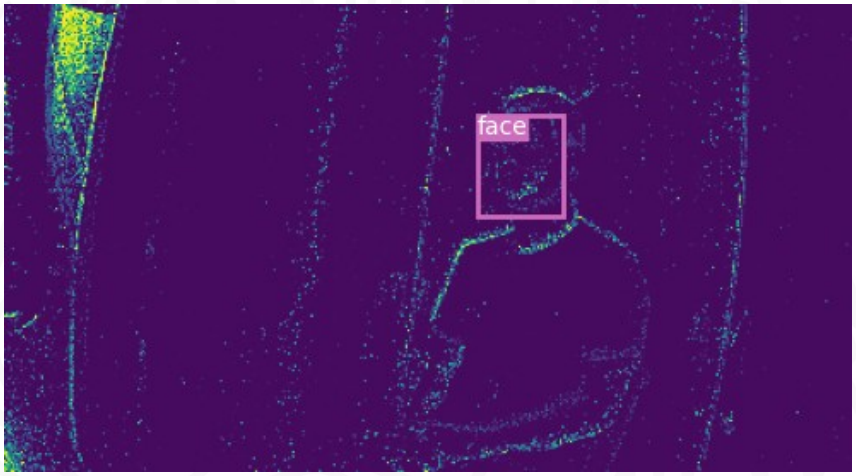
Infine sono stati fatti test su video a eventi realizzati utilizzando una vera event camera, che era fin dall'inizio l'obiettivo di questo progetto



UNIVERSITÀ DEGLI STUDI FIRENZE

Esperimenti e Risultati – Test





UNIVERSITÀ DEGLI STUDI FIRENZE

Conclusioni e possibili sviluppi

Al termine dei test si può affermare che l'obiettivo del progetto è stato raggiunto con successo.

Per il futuro:

- Sarebbe interessante ripetere lo stesso lavoro con un dataset di partenza più variegato rispetto a quello impiegato, in modo da ottenere un detector capace di lavorare in maniera ottimale in qualsiasi contesto.

- Si potrebbe testare la stessa metodologia adoperata in questo lavoro anche per task diversi dalla sola face detection.

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Grazie per l’attenzione!