17 - HUMAN ACTION RECOGNITION

Si analizza un video per identificare un'azione che sta accadendo nel video

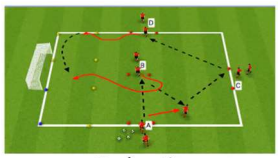
«running»

Si usa il video perché solo alcune azioni sono riconoscibili dalla singola immagine. L'evoluzione temporale svolge un ruolo fondamentale. (es: piegarsi vs cadere; camminata veloce vs camminata lenta)

t_{i-1} t_i t_{i+1} t_{i+2}

Il video lo si può pensare come un segnale 3D: (x,y) coordinate spaziali + t coordinata temporale.
L'intervallo temporale da andare a prendere in considerazione dipende dall'azione e dal contesto.

«squat»
(short interval)



«attack action»
(long interval)

Perché si è interessati a riconoscere le azioni?

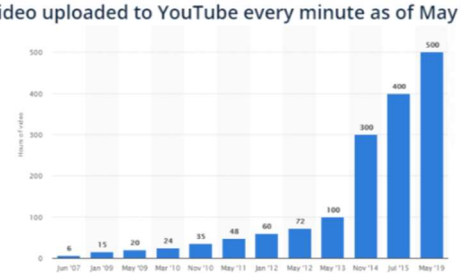
- Automatic video tagging/summarization
- Smart User-Interfaces (UI)
- Smart surveillance system
- Robotics (for human-robot interaction/robot learning)

Applications

Video indexing/retrieval

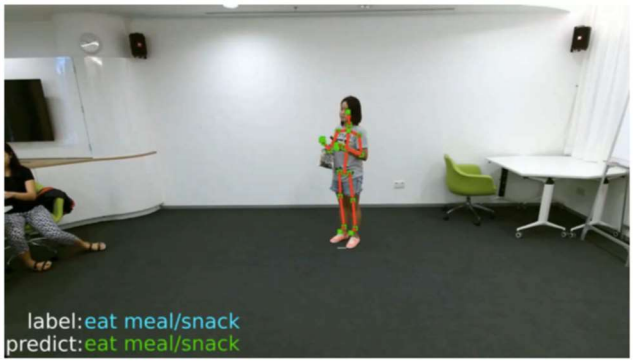
- ~500 hours of videos per minute
  - Need automatic tools for helping video indexing and retrieval

Hours of video uploaded to YouTube every minute as of May 2019

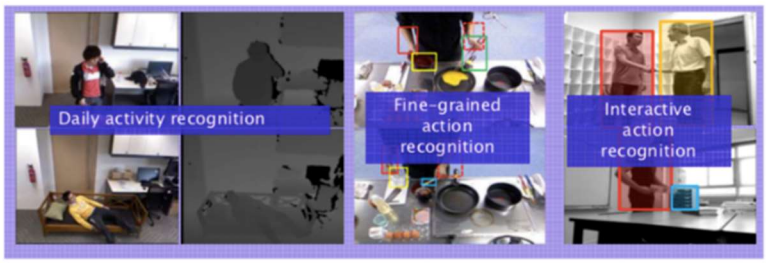


Crowd behaviour and event analysis

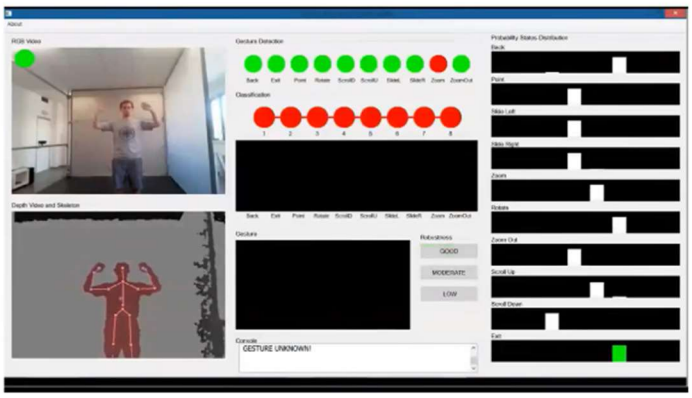
- Fine grained action labeling



- Intelligent assisted living and home monitoring
- Egocentring vision for tele-monitoring and assistance (e.g., for elders)



- Developing natural user interfaces by exploiting human motions and actions
  - Iterazione con il computer tramite gesti e movimenti



Challenges in Action Recognition

- Perople can appear at different scale in different videos, yet performing the same action


- Larger inter-class variability
  - Stessa persona può camminare in 100 modi diversi -> essere umano è in grado di capire ma pe run computer questa cosa qua diventa un incubo

- Occlusions
  - Actions may not be fully visible


➔ Perché non in ambiente totalmente controllato, persone possono anche scomparire in quanto occluse da altri oggetti o altre persone

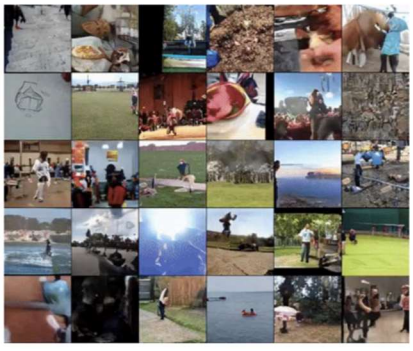
- Camera movements
  - Camera can be hand-held (or worn by the subject) or mounted on something moving causing shakes

- Background "clutter"
  - Other objects/humans present in the video frame.
- Human variation
  - Humans are of different sizes/shapes/clothes
- Trimmed vs untrimmed videos
  - Trimmed: video is cut focusing on the specific action
  - Untrimmed: there are no indication where in the timeline the action occurs

➔ Forme delle persone, diversi colori, adulti/bambini, persone magre/non magre/etc,

➔ Ci possono essere spezzoni video focalizzati sull'azione da riconoscere (trimmed) o sequenze di video delle quali non conosco la durata e non so quando inzia/finisce l'azione (untrimmed)

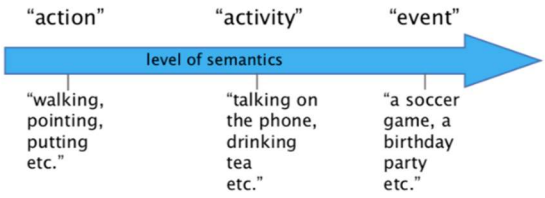
- Collecting trainign datasets is extremely challenging (many action classes, large inter-class variability, rare ouccurence)



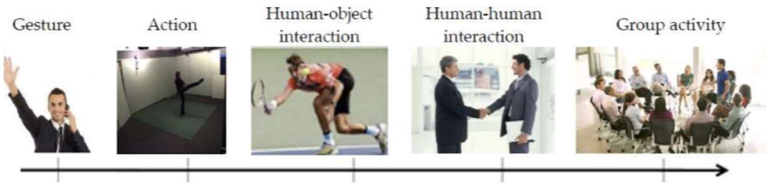
-> servono una quantità di dati enormi ma non c'è ancora un dataset con tanti video labelled.

what is action recognition?

- Input: video image
- Output: the «action» label



- Different types/levels of activities
  - The ultimate goal is to recognize all of them reliably



➔ Devo capire cosa mi interessa veramente catturare nella mia applicazione.

DATASET FOR ACTION RECOGNITION

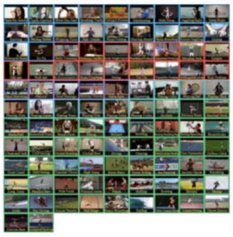
HDMB51 (Human Motion DB)

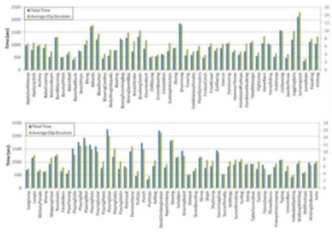
- 7K clips from 51 action categories (each with 101 samples) commcted from public repositories (YouTube, Prelinger archive)
  - Gesture, action, human-object interaction, human-human interaction

-> usato per valutazione

UCF-101

- 13K videos from 101 action categories. Large variations of camera motion, object apperance/scale/pose and so on
  - Gesture, action, human-object interaction, human-human interaction («sport» category as group activity)





-> video presi da youtube

Sports 1M

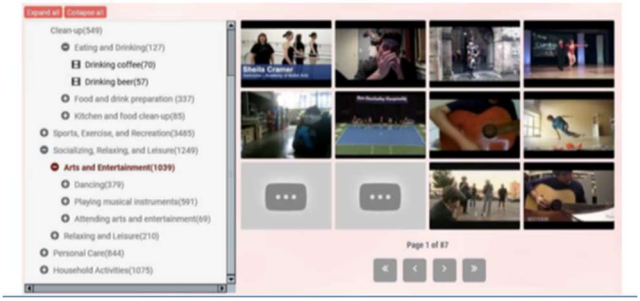
- 1M sport video from YouTube, 487 sport labels

-> maggiormente azioni individuali

ActivityNet

- 849 hours of video for human (daily-living) activity understanding 203 activity classes (an average of 137 untrimmed video per class, and 1.4 actions per video)
  - Three challenges: untrimmed video labeling, trimmed video labeling, action detection

-> ci sono video con più azioni, sia untrimmed ma si può fare trimmed



EpicKitchen

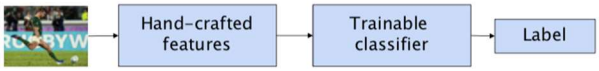
100h of video, shot in 45 kitchens (4 different cities). 97K action segments, 97 verb classes + 300 noun classes

Kinetics-700: tanti video trimmed da 10secondi

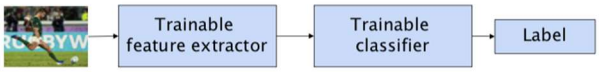
APPROACHES TO ACTION RECOGNITION

Recognition approaches

- Hand-crafted approaches



- Learning-based approaches



Hand-crafted approaches

- First, design «good» features for action representation


Space-time interest points

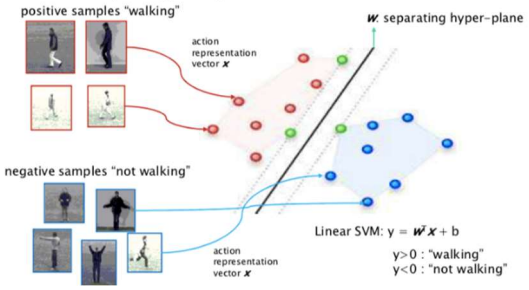

"Dense Trajectories


Motion History Images (MHI)


Body joints

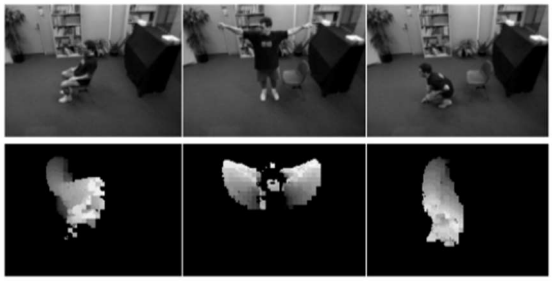
➔ Analizzano una serie di feature per andare a estrare informazioni dalle immagini. Feature che vengono poi combinate con una serie di classificatori.

- Then, design a «good» classifier



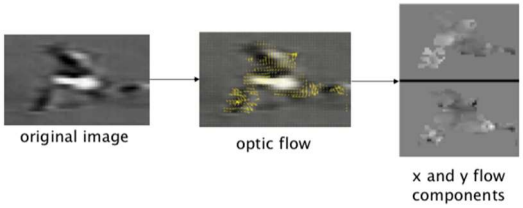
Motion History Images (MHI)

Sono delle immagini la cui rappresentazione segue un modello temporale. In ogni punto dell'immagine vado a definire un valore di intensità che mi indica se c'è stato del movimento e quanto è recente tale movimento in quello specifico pixel.



Optical flow

Mi permette di andare a decrivere lo spostamento di un pixel tra un frame e quello successivo

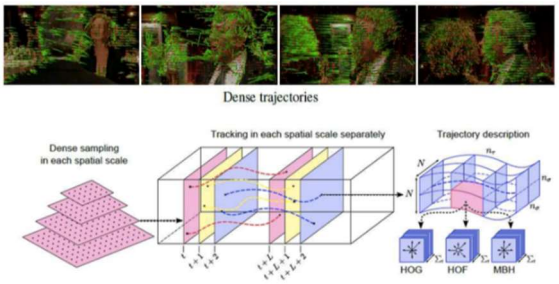


➔ Flusso ottico è generalmente rappresentato da un vettore a due dimensioni

Il flusso ottico non da un'informazione molto robusta. Sono stati sviluppati approcci che partendo dal flusso ottico fornissero delle feature più robuste.

Dense trajectories

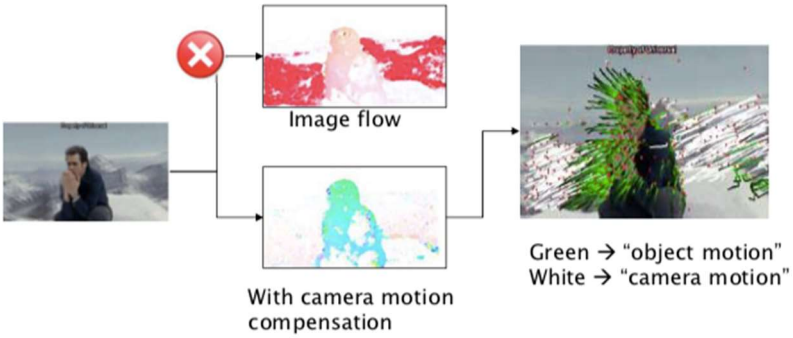
L'idea di base è quella di andare a prendere le informazioni del flusso ottico e metterle insieme a livello temporale per creare le traiettorie dense. Identifico i punti che hanno movimenti significativi e per arricchire la loro descrizione si può andare ad aggiungere un descrittore dell'immagine che mi rappresenti in qualche modo l'intorno de punto 2D della traiettoria nell'istante t → ho descrittori calcolati su volumi.



Dato che queste rappresentazioni sono abbastanza grandi, si prova ad usare dei metodi matematici (istogrammi su gradienti di vari biani, …) per ridurre.

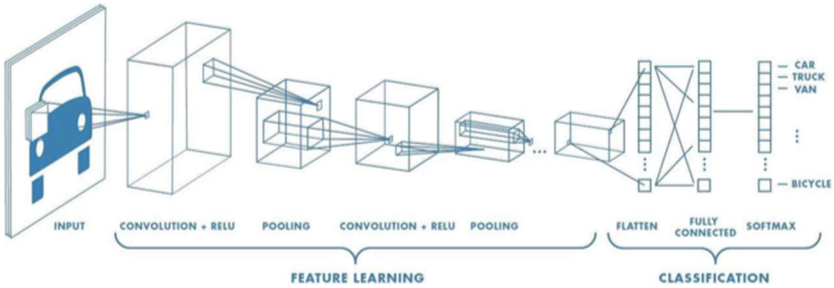
Improved dense trajectories

Si cerca di andare a compensare i movimenti di camera che non sono altro che rumore per l'analisi.
Lo si fa calcolando l'omografia di due immagini consecutive (metto in corrispondenza, allineo le due immagini, calcolo il flusso ottico) → come flusso ottico mi rimane solo il movimento degli oggetti. (riesco in questo modo a separare le traiettorie causate dal movimento della camera e quelle che sono quelle di interesse - movimento dell'oggetto)



Learning-based approaches

Le CNN riescono a portare allo stato dell'arte i task di image analysis.

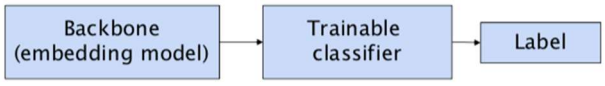


Backbone per estrarre le feature e poi classificazione. Nel nostro problema, dobbiamo estendere questa struttura per analizzare sequenze temporali

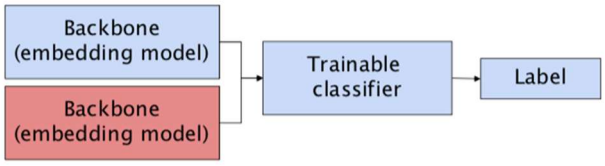
From image to sequences

Two basic approaches:

- Single stream architectures



- Two-stream architectures



Rami diversi per analizzare informazioni sul video di origine diversa.

SINGLE-STREAM NETWORKS

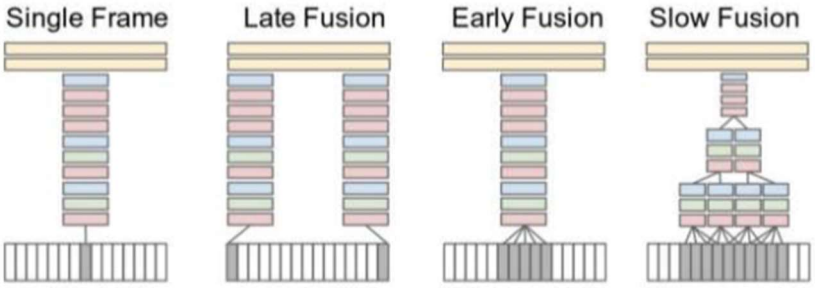
Single-stream network

L'idea fondamentale è quella di utilizzare una CNN standard per estrarre le features dei singoli frame del video, poi di andare a capire con approcci diversi quelle che erano le features temporali.

- Quale modello di fusione più ADATTO PER RISOLVERE il problema
- Come sfruttare al meglio le informazioni del dominio temporale

Il processo va a cercare di integrare in punti e modi diversi le feature combinate

- Partiti da una struttura base di una CNN su singolo frame
- Avanzato con vari Fusion

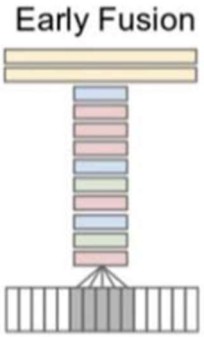


Early fusion

Combines information of a full time-window (10 frames)

Frames are stacked and the filter of the first levels are modified to operate on a T (T=10) temporal extent

- Size WxHx3xT



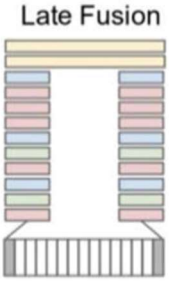
10 frame mandati insieme alla rete.

- Modifica fatta sui filtri del primo livello, combinando le informaizoni a basso livello, posso estrarre delle feature come direzione e velocità locale.
- I filtri sono addestrabili quindi cerco di ottimizzarlo.

Late fusion

Two branches (with shared weights) analyzyng frames at a fixed time ditance (15 frames)

The two (identical) streams are then mergedin the first FC layer of the classifier



Analizzo 2 frame ad una distanza temporale fissa.

- Mergiati nel mio layer FC
- Si analizzano concetti più strutturati

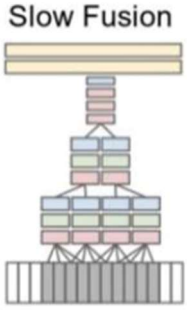
Slow fusion

Mix between the previous approaches

The fusion of temporal and saptial information is «slowly»distributed along the architecture

The number of branches are consecutively halved along the model

- The first four branches analyzes 4 frames each (two shared). The last conv layers combines info from all10 frames



Serie di fusioni progressive a livelli diversi per la rete in modo tale che i livelli più alti siano in grado di comprendere informazioni sempre più strutturate (sia spaziali che temporali).

- Ogno dei blocchi condivide 2 frame con il layer successivo
- Ad un certo punto si passa a 2 rami che elaborano ognuno la sequenza corrispondente a 6 frame (ci sono dei frame in comune per cercare di dare un raccordo tra le informazioni di tutti i rami)
- Infine c'è un restringimento finale e un unico ramo convoluzionale che va ad elaborare le informazioni che derivano dai 10 frames di ingresso. Le feature finali vengono mandate al classificatore.

Results

Quello che funziona meglio è lo slow fusion però questi approcci non è che diano chissà quale contributo significativo rispetto ad utilizzare il modello che elabora ogni frame e poi fa la media delle previsioni di ogni frame.

Sostanzialmente questo approccio non cattura in maniera efficace le informazioni spazio-temporali del video.

Better spatio-temporal features

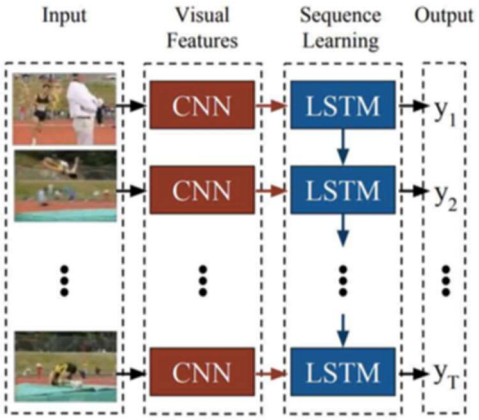
Esistono dei modelli per sfruttare meglio la dimensione temporale con RNN/GRU/LSTM e si possono combinare con i modelli che trattano le immagini quindi CNN.

- Andare ad analizzare le variabili latenti e combinare elaborazione spaziale e temporale

Il primo tentativo è stato quello di una LSTM sulle features estratte da una backbone pre-trainata, non ha portato vantaggi.

Long-term Recurrent CNN (LRCN)

Il vero passo in avanti è stato dato dall'aver integrato il blocco LSTM all'interno dell'architettura della rete.



- Prima parte: CNN che fa da encoder (trasforma immagine di ingresso in spazio delle feature)
- Seconda parte: LSTM che fa da decoder

La CNN funziona da encoder e poi c'è il blocco LSTM che fa da etichettatore.
Il vantaggio è che essendo la LSTM retropropagata, ne trae vantaggio anche la CNN perché collegata ad essa e quindi impara a definire delle feature anche le informazioni temporali derivanti dalla LSTM (oltre a quelle spaziali derivanti da input)

La predizione finale dell'azione è data dalla media delle predizioni per ogni singolo frame.

Come altra opzione si è cercato di capire quale potesse essere il contributo del flusso ottico ma si è scoperto che l'approccio migliore è di usare sia questo che l'RGB -> combinare.

Il vantaggio principale di questa rete è il **training end-to-end** dove quindi la rete impara in contemporanea la parte spaziale e quella temporale.

- Uno svantaggio è il problema del false labeling perché due frame successivi li posso catalogare come due azioni opposte che vengono rinforzate con l'allenamento della rete (es: clip di un salto, in un certo istante la persona è ferma – ma io do l'etichetta salto- -> questo crea rumore all'interno del labelling.)
- L'altro svantaggio riguarda l'intervallo temporale in cui la rete analizza il video (16 frame), la rete non prende le informazioni long-range perché prende tot frame alla volta.
    - Se devo analizzare un'azione più lunga, mi serve un numero di frame ppiù grande
        - Dovrei stretchare l'intervallo temporale ma a questo punto andrei a campionare die frame che non sono abbastanza significativi

una possibile idea è quella di considerare la sequenza come un volume in uno spazio tridimensionale -> input diventa un intero clip video.

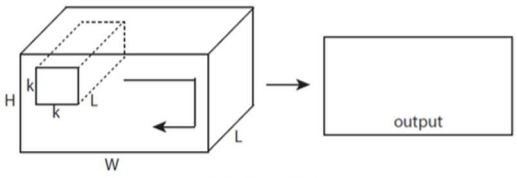
3D convolutions

Il video è un tensore 4D → RGB-t → perché non prendere questo come input?

Clips (k stacked RGB images)

Processing an image with a 2D ConvNet produces an image

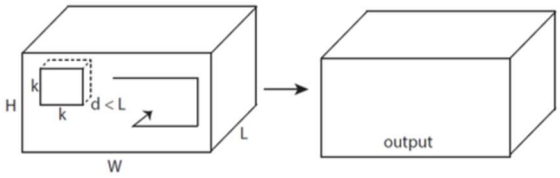
- Convolutions and pooling are only done spatially



2D ConvNet

Processing an image with a 3D ConvNet produces a volume

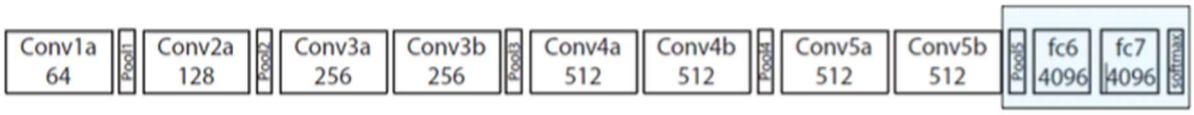
- Convolutions and pooling involve all spatio-temporal dimensions



3D ConvNet

C3D

Usa la convoluzione 3D per estrarre le features.



C3D features

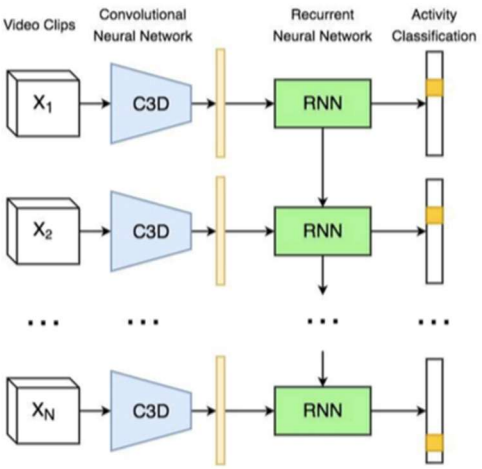
Una serie di blocchttini, poi fully-connected e poi la classificazione finale viene fatta con SVM lineare multi-classe.

L'allenamento viene fatto usando 5 clip casuali mentre nella fase di test la predizione su 10 clip casuali viene mediata per ottenere la predizione finale.

Limitazioni:

- Non riesce a catture le dipendenze temporali a lungo termine (clip di ingresso da 16 frame campionati ad una certa distanza, corrispondono a circa 2 secondi)
- Le convoluzioni 3D hanno un numero di parametri più grande, addestramento più complesso, ho bisogno di datasert più grossi.

Una possibile soluzione è unire questa architettura alle RNN



Invece di fare semplicemente la media, si fa un filtraggio per isolare fram, clip dove non è identificata alcuna azione, scartando frame e poi facendo media su quello che rimane.

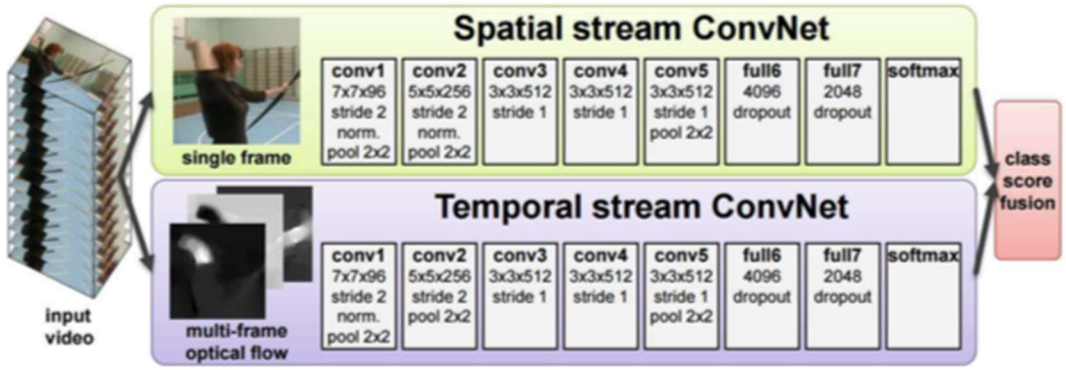
TWO-STREAM NETWORKS

Questi modelli hanno due rami di elaborazioni diversi e quindi lavorano su informazioni di natura diversa per riuscire poi a combinarle.

- Un ramo su spaziale
- Un ramo su temporale

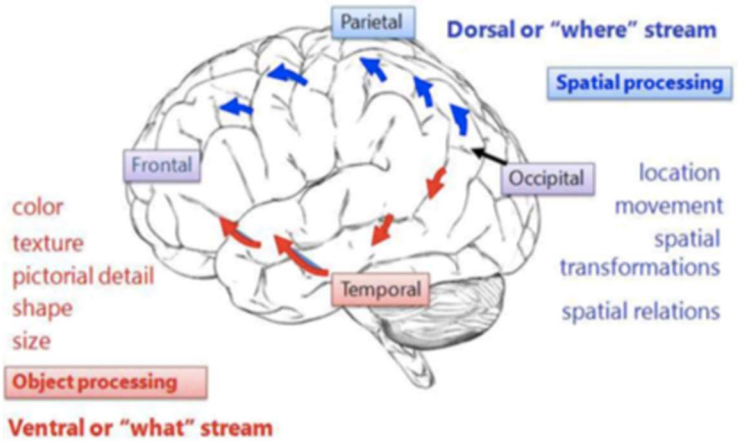
Two stream models

Informazioni di movimento estratto in maniera esplicita da un ramo della rete che si occupa esclusivamente del flusso ottico.
L'altro ramo elabora il flusso ottico basandosi anche su qualche frame precedente e qualche frame successivo.



Si ha un CNN standard che estrae il contesto spaziale di un frame RGB e una CNN che elabora il contesto temporale dal flusso ottico di più frame.

L'idea viene dalla corteccia visiva del cervello umano che trasmette due flussi diversi su due canali diversi.

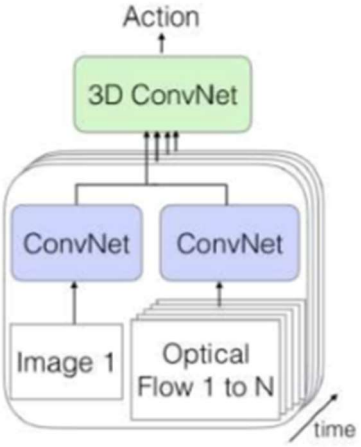


Nella rete la combinazione finale dei flussi viene fatta andando a concatenare le features estratte dei due rami dandole in pasto ad un SVM ottenendo l'etichetta associata al frame RGB in ingresso.

Questa architettura presenta performance migliori delle architetture single-stream però rimangono i problemi delle informazioni long-range temporali, dei false-label assignement e inoltre non può essere trainato end-to-end -> non può beneficiare dell'unione di informazioni spazialie temporali duranti il training (3 parti addestrate separatamente: spaziale, temporale, svm).

3D-fused two streams

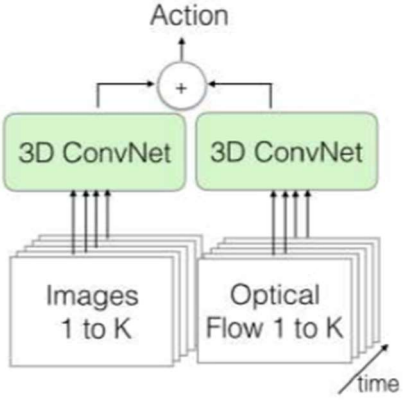
Si usano diversi blocchi two-stream in modo tale da analizzare più frame e andare a classificarli insieme tramite una rete convoluzionale 3D.



L'analisi finale, quindi, viene fatte sulle features di più immagini RGB tramite una 3D conv.
Lo scopo è quello di ridurre i problemi dovuti alla visione temporale ridotta dell'architettura precedente.
In effetti si ottengono risultati migliori.

Two (3D) stream models (i3D)

Estendo i singoli rami in 3D piuttosto che la singola rete



Sfrutto I vantaggi dell'elaborazione 3D, ottenendo informazioni di più alto livello.

Questi modelli hanno tanti parametri che devono essere addestrati con una serie di dati opportuni.

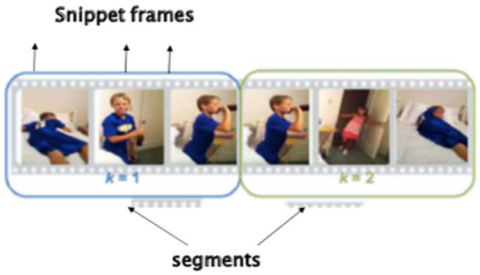
- **Come aum**entare ulteriormente intervallo temporale?
- **Come** estratte le informazioni rilevanti nei domini spaziali e temporali (Evidando fakse labelling)?

Temporal Segment Network (TSN)

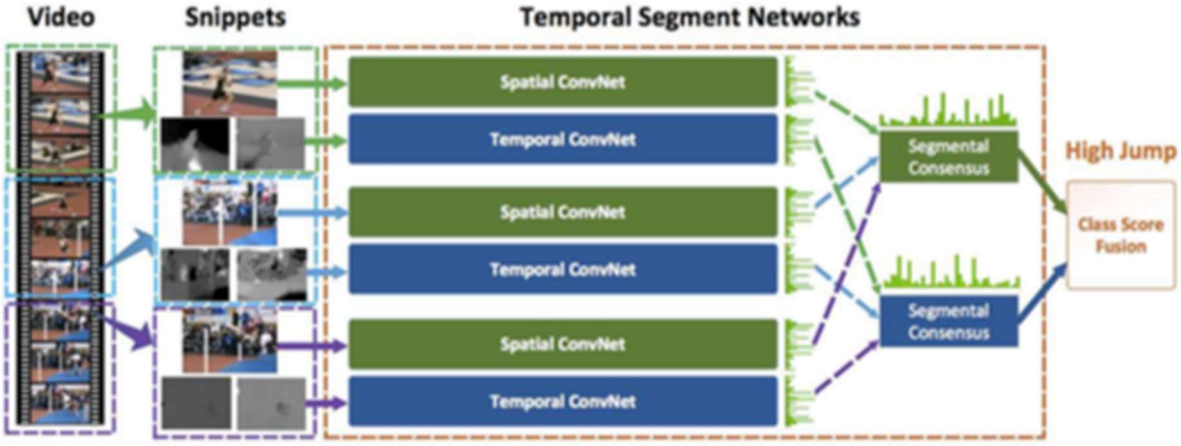
Un'osservazione importante da fare sulla questione temporale è che i frame di un video molto spesso hanno informazioni ridondanti; quindi, fare un campionamento denso è un'operazione inutile (a volte dannosa).

Nelle TSN si usa un campionamento temporale più sparso delle clips (snippets), si usa uno schema robusto per aggregare le predizioni degli snippets e inoltre il training è end-to-end.

- Spezzo video in un certo numero di segmenti
- Da ogni segmento estraggo n frames in maniera casuale (se uso lo stesso video 2 volte, estraggo frames diversi, quindi dovrebbe aiutare ad evitare false labelling).
- Elaboro informazioni di ogni singolo frammento
- Combino le informazioni per valutare output



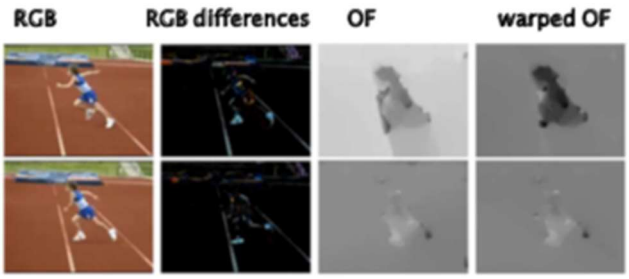
Ogni snippet è processato da un modello two stream le cui uscite vengono elaborate in maniera separata per la parte spaziale e quella temporale da delle funzioni di consenso le cui uscite vengono combinate per ottenere la decisione finale.



Com'è la funzione di consenso?

L'importante è che sia una funzione in qualche modo differenziabile per garantire la backpropagation e si è visto che la migliore opzione è la media non pesata.

Performance



Modality	Performance
RGB Image	84.5%
RGB Difference	83.8%
RGB Image + RGB Difference	87.3%
Optical Flow	87.2%
Warped Flow	86.9%
Optical Flow + Warped Flow	87.8%
Optical Flow + Warped Flow + RGB	**92.3%**
All Modalities	91.7%

Accuracies on UCF-101

La miglior combinazione si ottiene con dati RGB con la combinazione dei flussi ottici.

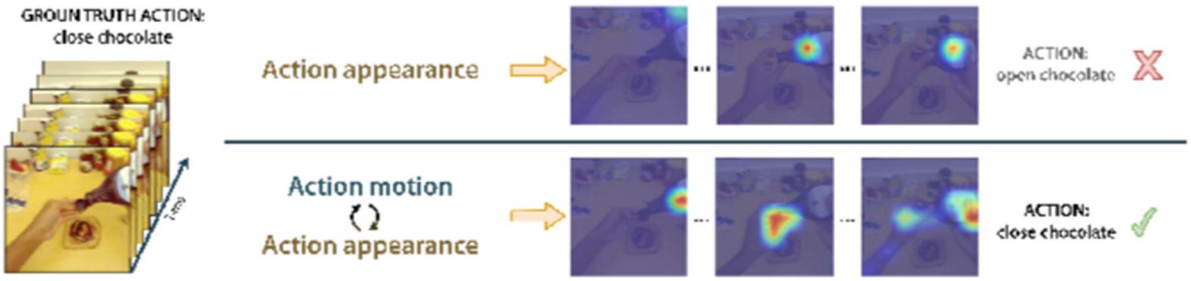
---- NON FATTO

Egocentric vision

Per questo tipo di video non si possono applicare le metodologie standard per le caratteristiche della visione egocentrica.
Bisogna cercare di estrarre più informazioni possibili sugli oggetti interessati, la loro posizione, il movimento nel video.

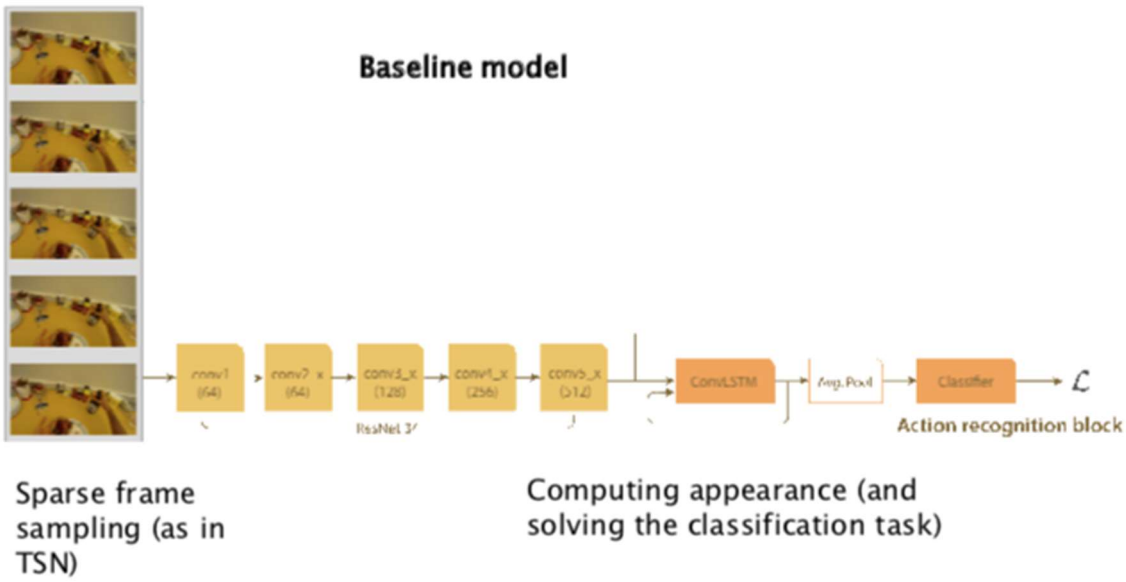
SparNet

L'idea principale è quella di utilizzare in fase di test un singolo flusso RGB. Il flusso ottico viene usato solo nella fase di addestramento però in una maniera diversa rispetto ai precedenti modelli two-stream.



Non viene usato come informazione aggiuntiva all'informazione RGB ma per fargli risolvere un task alternativo che non c'entra col problema principale della classificazione. La rete deve classificare l'azione e nello stesso tempo deve risolvere un problema legato a quelle che sono le nidificazioni del movimento all'interno delle immagini.
La cosa importante è che questo task è self supervised perché le etichette sono già contenute all'interno dei dati.
Questo dovrebbe aiutare le features RGB con le informazioni spazio temporali.

SparNet Architecture



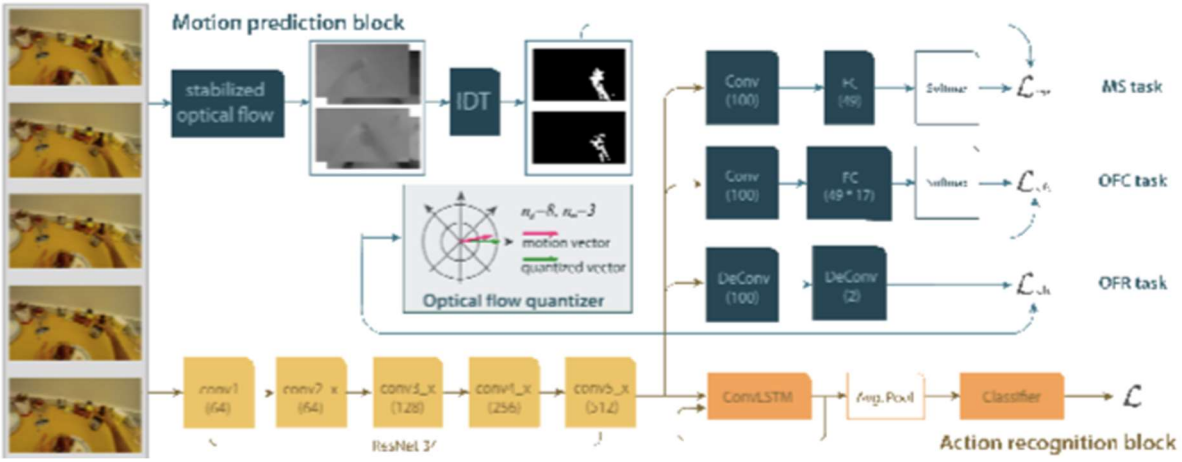
L’architettura ha una versione base che ha soltanto il blocco che si occupa di riconoscere l’azione. Questo blocco è composto da una backbone standard che è una CNN che va a generare una serie di feature.
Questo sono passate ad una Conv LSTM e infine in un classificatore.
Questo prende un numero molto piccolo di snippets del video.

Manca ancora la parte dei task ausiliari.
Dalla singola immagine RGB si fa una motion prediction cercando di rispondere ad uno o entrambi i quesiti:

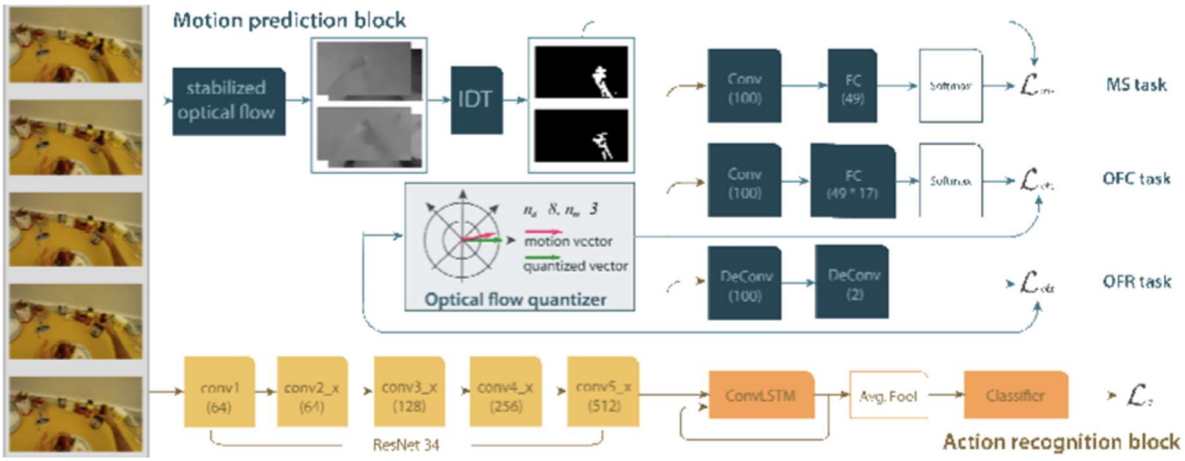
- Quali sono le parti dell’immagine che si muoveranno nel prossimo frame?
- In che direzione si muoveranno?

Questo lo si può fare partendo dall’imagine RGB usando come ground truth il warp flow in 3 modi:

- Motion segmentation (MS) → etichetto le parti in movimento
- Optical Flow Regression (OFR) → quantizzo il flusso ottico
- Optical Flow Classification (OFC) → ricostruisco il flusso ottico quantizzato



Ogni task viene gestito da un apposito ramo che da una parte prende le features estratte dall’ultimo layer convoluzionale della ResNet e poi va a fare l’elaborazione specifica.



SparNet results

La backbone riesce a generare delle features che hanno delle caratteristiche spazio temporali. Tra i vari tasks quello che da risultati migliori è quello che combina il motion segmentation con la classificazione del flusso ottico. La spiegazione intuitiva è che entrambe le parti capiscono la direzione di movimento. Inoltre

le operazioni al secondo richieste rispetto alle altre architetture sono molto inferiori rendendo più veloce le fasi.

Method	Acc (%)	Param (M)	GFLOPS
baseline	65.46	24.34	41.52
SparNet-MS	68.43	24.87	41.55
SparNet-OFR	65.73	24.80	43.01
SparNet-OFC ($n_d = 8$)	69.32	30.39	41.61
SparNet-OFC ($n_d = 16$)	69.49	36.16	41.68
SparNet-OFC ($n_d = 20$)	68.96	39.04	42.21
SparNet-OFR+OFC ($n_d = 16$)	69.57	36.62	43.17
SparNet-MS+OFC ($n_d = 16$)	**69.80**	36.70	41.71
Ego-RNN RGB [3]	-	24.34	94.36
Ego-RNN [3]	-	45.71	98.31
LSTA RGB [2]	57.96	41.22	114.92
LSTA [2]	61.86	62.59	118.86
Two-stream I3D + STAM [19]	68.60	-	-
3DConv MTL [17]	68.99	-	-