DEEP LEARNING IN CUCINA: SVILUPPO E VALIDAZIONE DI UN SISTEMA DI RICONOSCIMENTO DI AZIONI BASATO SU SENSORI RGBD

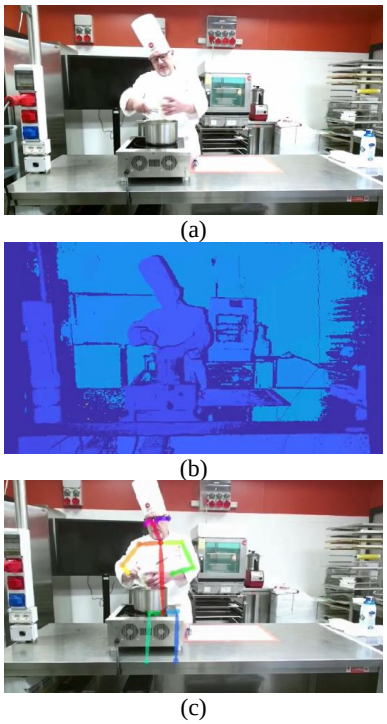
S. Pasinetti[(1)], C. Nuzzi[(1)], E. Picinardi[(1)], G. Sansoni[(1)]
[(1)] Dip. di Ingegneria Meccanica e Industriale, Università di Brescia, Via Branze, 38- 25123 Brescia
mail autore di riferimento: simone.pasinetti@unibs.it

1. INTRODUZIONE

La presente memoria descrive i primi risultati raggiunti nell'ambito di un progetto di ricerca con la scuola bresciana di cucina Cast Alimenti. L'obiettivo del lavoro di ricerca è lo sviluppo di un sistema intelligente per il riconoscimento delle azioni svolte da un cuoco durante la preparazione di una ricetta. Cast Alimenti mira ad ottenere un prodotto da utilizzare durante la didattica che abbia un duplice scopo: da una parte si vuole riconoscere che operazione il cuoco docente sta effettuando, con che ingredienti e con quali utensili; dall'altra parte il sistema sarà in grado di effettuare la stessa operazione di riconoscimento con gli alunni della lezione, controllando se l'operazione pratica di cucina viene svolta nel modo migliore.
In questa memoria vengono descritti i primi risultati raggiunti relativi al riconoscimento delle azioni del cuoco. Il riconoscimento delle azioni è stato effettuato e valutato confrontando due tra i migliori algoritmi di riconoscimento azioni basati su reti neurali ricorsive [1]: il primo, denominato Human Pose Model and Temporal Modelling (HPM+TM) [2], basato sull'analisi di immagini di profondità e il secondo, denominato Indipendetly Recurrent Neural Network (IndRNN) [3], basato sulla misura di diversi keypoint individuati a partire da una skeletonization del soggetto ripreso.

2. METODI

Il setup sperimentale è composto da due telecamere Kinect One (Microsoft) poste all'interno di una cucina attrezzata, poste frontalmente al banco di lavoro del cuoco. La ricetta utilizzata per i test è stata scelta in modo da coprire il più possibile la casistica di azioni che un cuoco può effettuare durante il proprio lavoro. Per questo motivo, i test di riconoscimento sono stati effettuati durante la preparazione di lasagne alla bolognese che comprendono quindi preparazione della sfoglia, preparazione del ragù, preparazione della besciamella e assemblaggio finale del tutto. Le prove di misura sono state svolte in due giornate diverse, e hanno previsto due diverse sessioni di cottura in modo da avere un dataset di immagini adeguato al training dei due algoritmi di misura.
Una volta acquisite tutte le immagini necessarie è stata svolta una operazione di labelling manuale in modo da identificare per ogni acquisizione i frame relativi a determinate azioni. Tramite questa operazione è stato possibile individuare una serie di azioni da riconoscere. Il dataset finale analizzato comprendeva 21 diverse categorie azioni di cucina. Ogni categoria comprendeva un numero di ripetizioni della singola azione variabile da un minimo di 4 ripetizioni (azione "imburrare"), ad un massimo di 369 ripetizioni (azione "mescolare"), per un totale di 1131 ripetizioni. Tra le varie categorie solo 14 comprendevano più di 10 ripetizioni. Si è scelto di allenare e testare i due algoritmi solamente su queste ultime in quanto per le categorie che presentavano meno di 10 azioni il training sarebbe stato inefficace.



(a) (b) (c)

Fig. 1 – Funzionamento dei sue algoritmi: frame originale (a), l'immagine di profondità per algoritmo HPM+TM (b), e skeletonization per l'algoritmo IndRNN (c).

Una volta suddivise le varie azioni i due algoritmi sono stati valutati tramite 2 diverse analisi:

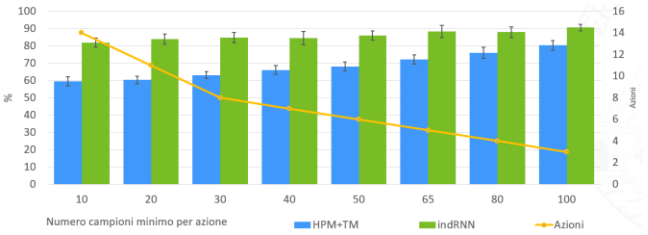
1. valore minimo di ripetizioni per categoria (Vmin): in questa analisi è analizzato il comportamento dei due algoritmi in base alla numerosità di ripetizioni di ogni categoria usate per il training in modo da identificare come varia l'accuratezza di riconoscimento in funzione del numero di ripetizioni usate nel training. Per svolgere questa analisi sono state svolte 240 sessioni di training per ogni algoritmo e altrettanti test di accuratezza.
2. Numero di categorie costante (Ccost): in questa analisi è stato valutato il comportamento dei due algoritmi analizzando solamente quelle azioni aventi un numero maggiore di ripetizioni e variando il numero di ripetizioni utilizzate per il training. Questa analisi è stata effettuata per analizzare le performance degli algoritmi solamente in base al numero di ripetizioni utilizzate per il training, eliminando il fattore numero di categorie. Questa analisi si è resa necessaria perché tramite l'analisi Vmin, mano a mano che aumentavano le ripetizioni utilizzate per il training, il numero di categorie che l'algoritmo doveva classificare diminuiva. Quindi tramite l'analisi Vmin era impossibile stabilire se l'accuratezza migliore dipendeva da una migliore risposta dell'algoritmo o da una maggiore semplicità di riconoscimento (dovuta al minor numeri di categorie da riconoscere presenti).

In figura 1 è rappresentato un esempio di funzionamento dei due algoritmi: la figura rappresenta il frame originale acquisito da una delle due Kinect (a), l'immagine di profondità utilizzata dall'algoritmo HPM+TM (b), e la skeletonization utilizzata dall'algoritmo IndRNN (c).

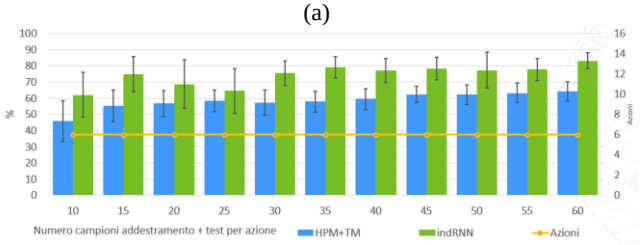
3. RISULTATI

In figura 2 sono riportati i risultati dell'accuratezza di riconoscimento ottenuta tramite le analisi descritte precedentemente.

Per l'analisi Vmin possiamo affermare che in entrambi gli algoritmi esiste un miglioramento dell'accuratezza in funzione di un maggior numeri di ripetizioni utilizzate per il training. Per l'algoritmo IndRNN si nota che, anche in presenza di categorie con un numero scarso di campioni, non è presente una diminuzione significativa di accuratezza. A causa della logica dell'analisi nei casi con più ripetizioni l'accuratezza potrebbe migliorare anche a causa di un numero minore di categorie di azioni da riconoscere.



(a)



(b)

Fig. 2 – Risultati di accuratezza ottenuti per il riconoscimento tramite l'analisi Vmin (a), e l'analisi Ccost (b).

Per l'analisi Ccost si nota come l'aumento della dimensione del dataset di training migliori le performance di entrambi gli algoritmi: l'algoritmo HPM+TM necessita di un dataset ampio e di un basso numero di categorie mentre indRNN offre ottime prestazioni anche in presenza di molte categorie e poche ripetizioni. Questo dimostra come la qualità dei dati di training incida fortemente sui risultati. Si rende quindi necessario creare un dataset con un numero uniforme di ripetizioni per categoria, ed è necessario indagare questa inconsistenza riscontrata nella qualità dei dati, per capire se effettivamente esistano azioni più difficoltose da riconoscere per l'algoritmo e le relative cause.

RIFERIMENTI BIBLIOGRAFICI

[1] L. Wang, D. Q. Huynh and P. Koniusz, "A Comparative Review of Recent Kinect-Based Action Recognition Algorithms," in *IEEE Transactions on Image Processing*, vol. 29, pp. 15-28, 2020

[2] H. Rahmani and A. Mian, "3D Action Recognition from Novel Viewpoints," *CVPR*, pp. 1–12, 2016.

[3] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN," *CVPR*, pp. 5457–5466, 2018.