Analisi visuale di azioni e interazioni in ambienti educativi

(dal tracking al video tagging)

Calderara Simone
DIEF Università di Modena e Reggio Emilia
Modena, Italia
Simone.calderara@unimore.it

Vezzani Roberto
DIEF Università di Modena e Reggio Emilia
Modena, Italia
Roberto.vezzani@unimore.it

***Abstract*— Le interazioni in ambienti educative abbracciano diversi livelli di astrazione e dipendono fortemente dal contesto. Nello specifico si distinguono due tipi di interazioni principali, quella osservata e di natura spantanea in cui i bambini sono liberi da vincoli di natura tecnologica, e quella stimolata in cui il sistema guida il bambino in interazioni attive quali le fasi di gioco. In questi contesti le tecnologie visuali permettono di anlizzare le interazioni e forniscono un importante strumento con il duplice scopo di documentazione dell'attività e analisi delle stesse. In questo articolo si descrivono le tecnologie visuali che Imagelab UNIMORE ha applicato all'interno del progetto e si delineano le possibilità che gli sviluppi futuri metteranno a disposizione degli educatori nei diversi contesti.**

I. INTRODUZIONE

L'obiettivo principale del gruppo di ricerca Imagelab dell'Università di Modena e Reggio Emilia all'interno del progetto è quello di studiare e realizzare strumenti tecnologici avanzati per l'educazione con particolare riferimento agli ambiti della visione artificiale, della pattern recognition e del multimedia. E' infatti in questi ambiti che Imagelab ha dedicato la maggior parte delle attività di ricerca dalla sua fondazione.

L'importanza delle tecnologie proposte è anche supportata dai recenti trend tecnologici e soprattutto commerciali legati ad altri campi, in primis quello dell'entertainment. L'evoluzione dei sistemi di interazione uomo-macchina (dagli "innaturali" dispositivi quali mouse e tastiera alle moderne interfacce basate su schermi touch o controllate da sistemi di visione quali Kinect), l'integrazione di dati multisensoriali, il paradigma Internet-of-Everything (dove ogni oggetto dell'ambiente può diventare un potenziale attore della scena) sono solo alcuni esempi applicativi.

Fig.1

Imagelab si occupa di fornire nuove tecnologie software e hardware basate su elaborazione di sensori, in particolare ottici.

Le interazioni dei bambini con i sensori di tipo ottico possono variare a seconda dell'ambiente occupato e dell'attività svolta.

In particolare si individuano due differenti livelli di interazione:

- Interazione non collaborativa / osservata
- Interazione collaborativa / stimolata

Nel primo caso l'utente (bambino) interagisce con lo spazio e le nuove tecnologie in modo attivo e consapevole, in linea con le attività pianificate. In questo scenario, gli strumenti tecnologici rappresentano nuove **modalità di input** sostituendosi agli strumenti tradizionali.

Nel secondo caso l'utente (bambino) interagisce con lo spazio e le nuove tecnologie in modo inconsapevole. **L'osservazione** delle attività e la

documentazione delle stesse permette di fornire agli educatori una analisi più generale e accessibile a posteriori di aspetti non direttamente legati alla attività, come ad esempio le interazioni sociali.

II. Interazione Osservata e Ambienti Educativi

L'interazione osservata è riservata agli ambienti educativi ed ai contesti in cui il bambino gode di libertà di movimento ed interazione.

Fig.2

Tali contesti possono essere le zone di gioco comuni indoor e outdoor come ad esempio mostrato in Fig. 1.
In questi ambienti le interazioni si caratterizzano per la loro naturalezza e l'invasività dello strumento di rilievo/osservazione deve essere minimizzata per non influenzare la spontaneità dei bambini.

Lo studio e l'osservazione delle attività in tali contesti necessità quindi di strumenti di analisi e sintesi che possano coadiuvare l'educatore e fornire informazioni importanti per la cartterizzazione dei comportamenti. Le tecnologie abilitanti nell'ambito della visione artificiale consentono l'estrazione di informazioni sia di basso livello che di tipo semantico quali ad esempio (aumentando il livello diastrazione):

1. Individuazione e tracking dei bambini
2. Individuazione e caratterizzazione dei volti e delle espressioni
3. Individuazione delle attività complesse per video tagging e documentazione.

Le tecnologie sviluppate presso Imagelab sono state applicate in questo contesto per

- Tracciare e documentare l'attività e le features del bambino in scenari complessi Fig. 2-3
- Caratterizzare le traiettorie di più individui in momenti di interazione comune. Fig. 4.

.

Fig. 3

Questa prima analisi permette al sistema automatico di raccogliere informazioni importanti per poter effettuare l'analisi automatica di dati visuali da differenti sorgenti video e in contesti non strutturati

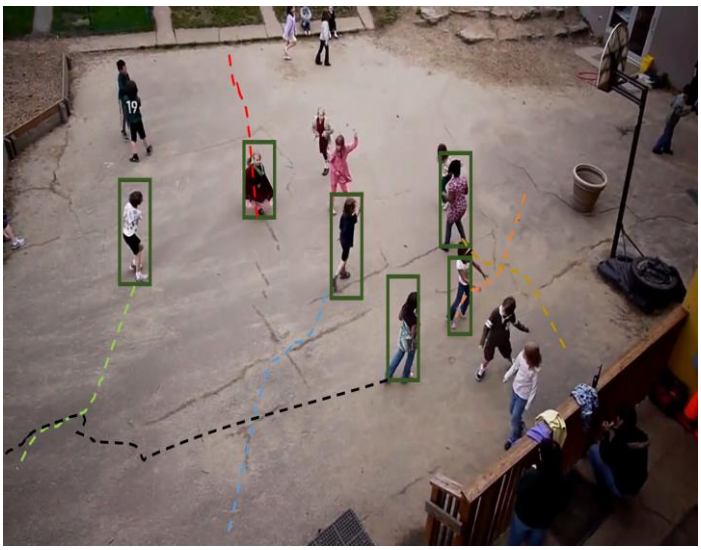


Fig. 4

A. *Scenari Tridimensionali*

All'interno di ambienti strutturati come stanze di gioco o digital playgrounds è possibile utilizzare sensori tridimensionali che permettono una acquisizione di dettagli anatomici dl corpo quali rappresentazioni sintetiche a scheletro Fig. 5.

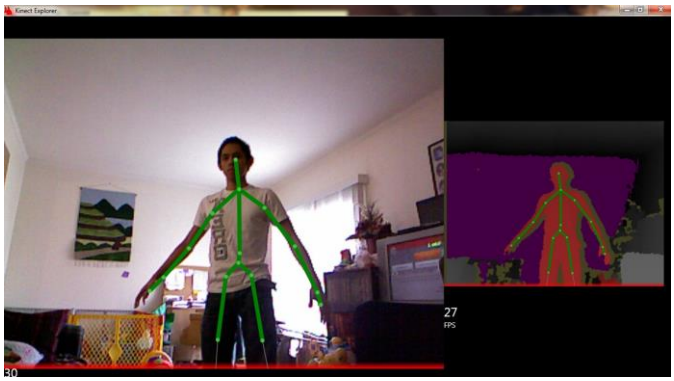


Fig. 5

Tali rappresentazioni forniscono informazioni importanti sulle attività dei bambini e possono rappresentare le azioni a diversi livelli di astrazione aggiungendo semantica e caratterizzando le interazioni di più soggetti.

III. Video Tagging di attività

Gli elementi elaborati ed estratti da sensori visuali possono essere utilizzati per caretterizzare le attività dei bambini.
In ogni caso è necessario considerare un set di attività definito che consenta di poter applicare modelli computazionali discreti e strumenti apprendimento per la categorizzazione dei dati visuali.
L'attività di tagging automatico del video consiste nell'estrarre in modo automatico le attività di alto livello svolte dai soggetti.
L'utilità di tale fase risulta duplice:

- Utilità documentale: consente di mantenere una traccia sintetica e testuale delle attività dei bambini evitando la fase manuale di "etichettatura" dei dati.
- Utilità di analisi: consente di effettuare analisi e correlazioni sulle attività svolte in relazione alla situazione l'ambiente e la natura dei soggetti coinvolti. Ad. Es la valutazione della partecipazione ad un gioco interattivo o l'engagement in una attività di classe.

I livelli di riconoscimento dell'attività possono essere differenti a seconda dei livelli di astrazione considerati.
In generale si fa riferimento a una gerarchia a tre livelli:

1. Dati di moto: azioni elementari del singolo soggetto quali in moto/fermo..
2. Azioni: azioni strutturate del singolo soggetto quali corre/saluta/osserva/sorride….
3. Interazioni: azioni strutturate di più soggetti che prevedono forme di engagement quali: parlare/giocare…..

La proposta all'interno del progetto è quella di avere differenti sorgenti video dalle quali si possono estrarre informazioni di tagging a diversi livelli di astrazione.

Per quanto concerne i video generici (ad es. video esistenti e precedentemenți acquisiti all'interno della normale attività scolastica) le informazioni che verranno estratte sono relative ai punti 1-2 per via dell'estrema eterogeneità dei dati estratti da sensori generici e della mancanza di informazioni tridimensionali.

All'interno invece di video provenienti da sensori tridimensionali, grazie alla presenza di modelli complessi e ricchi di informazioni è possibile individuare le Interazioni strutturate al punto 3 fornendo la possibilità di introdurre tag semantici all'interno del video per la documentazione delle attività e la navigazione dei contenbuti multimediali Fig. 6.

IV. Interazione Stimolata in ambienti educativi

In questo contesto le tecnologie ICT sperimentate e studiate devono diventare strumenti di interazione tra l'utente (bambino) e la macchina (ambiente).

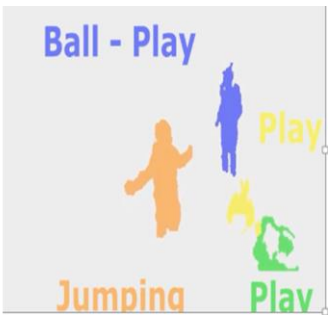


Fig. 6

In passato l'interazione classica uomo-macchina era rappresentata dalla coppia tastiera-mouse, sostituita solo in campo entertainment da console e device più user-friendly.

Nel campo educational, invece, l'inerzia all'uso degli strumenti tradizionali ha enormemente rallentato l'evoluzione e l'introduzione di nuovi strumenti di interazione. Imagelab, forte della sua esperienza in campo di visione artificiale e di pattern recognition, sta attualmente studiando nuovi sistemi di interazione, con specifiche applicazioni nel campo educational nel contesto del progetto. In particolare in questo articolo verranno descritti tre diversi scenari basati su device di nuova concezione quali Microsoft Kinect ® e il pavimento sensorizzato Florimage ®.

A. Ricostruzioni tridimensionali e ambienti virtuali

Questo scenario è stato ideato in collaborazione con il Centro Loris Malaguzzi. E' stata attrezzata una stanza con un sensore di acquisizione Microsoft Kinect 2.0 ed un proiettore in grado di riprodurre contro una intera parete le immagini prodotte dal sistema. All'interno della stanza sono stati introdotti alcuni bambini di diverse età (dai 4 ai 12 anni), liberi di interagire tra loro, con la scena proiettata e con oggetti lasciati all'interno della stanza.

Mediante il sensore Microsoft Kinect è stato possibile proiettare sullo schermo diversi scenari, abilitando gli educatori ed osservatori a studiare la risposta e il comportamento dei discenti. Tra i diversi scenari menzioniamo:

- Eliminazione di background e sovra-imposizione del foreground su sfondo variabile (mare, spazio, prato, deserto), creando ambienti immersivi (vedi Figura 5.a).
- Individuazione dello scheletro e visualizzazione tridimensionale mediante linea spezzata, per sottolineare la possibilità di movimentare una identità diversa da quella umana mediante il proprio corpo.
- Visualizzazione della nuvola di punti 3D acquisita, sia con che senza informazione di colore, con possibile modifica del punto di vista del rendering, per simulare l'acquisizione dall'alto, da dietro o da posizione diversa da quella del sensore stesso (vedi Figura 5.b).

Fig. 7.: setup di studio installato presso il Centro Loris Malaguzzi

B. Riconoscimento automatico di gesti da scheletri 3D

Nel caso di interazione diretta singolo utente mediante sensore Kinect, un ambito ancora aperto dal punto di vista della ricerca è il riconoscimento online automatico di gesti. In particolare, grazie alla possibilità di individuazione della posizione 3D

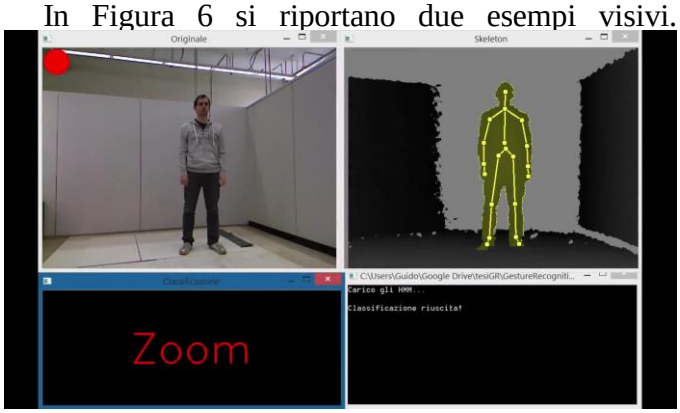
degli arti [2,3], il sistema si baserà su tali informazioni per il riconoscimento dei gesti.

Ipotizziamo di voler controllare un sistema mediante gesti svolti dall'utente di fronte al sensore di acquisizione. L'approccio classico è quello di dividere la sequenza temporale di frame in finestre di lunghezza fissa (esempio 2 secondi) e classificare indipendentemente ogni sotto-sequenza, valutando la presenza di un gesto al suo interno. Tale approccio, usato anche dalla stessa Microsoft nel software dimostrativo del sensore Kinect, ha un difetto di fondo che si presenta quando i gesti da riconoscere hanno una durata temporale diversa tra loro. In tali casi, la selezione della finestra di segmentazione risulta difficile. Pochi sono gli approcci che cercano di segmentare automaticamente le sequenze di input in finestre di lunghezza variabile [5] o che, al contrario, cercano di effettuare contemporaneamente la segmentazione e la classificazione dei gesti.

L'approccio in fase di studio appartiene a questa categoria ed è una evoluzione del metodo presentato in [1]. Il sistema si basa su una batteria di classificatori HMM [4], addestrati ciascuno al riconoscimento di una singola azione.

Il sistema è in grado di riconoscere i gesti inseriti in fase di training, individuarne inizio e fine (temporale), identificare le fasi in cui l'utente non sta svolgendo gesti, evidenziare la presenza di gesti interrotti e non completati.

In Figura 6 si riportano due esempi visivi.



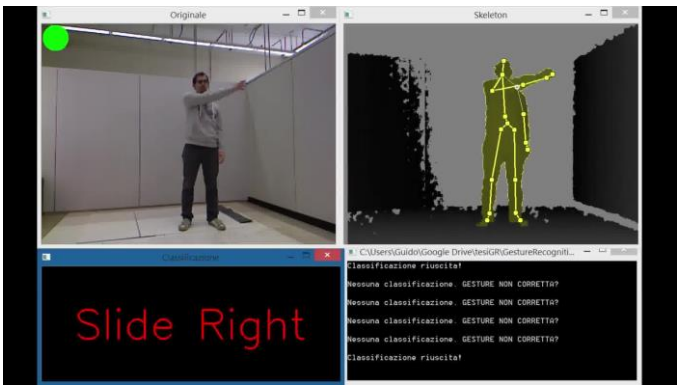


Fig. 6.: output del sistema di classificazione di azioni avuto in caso di gesto Zoom e Slide Right.

C. *Interazione uomo-macchina mediante pavimento sensorizzato*

In questo terzo scenario l'interazione uomo-macchina avviene mediante un pavimento sensorizzato [6,7].

Il dispositivo di input è composta da 4 m^2. L'unita di processing di basso livello raccoglie i dati provenienti dal sensore, traccia la posizione di eventuali utenti presenti sul pavimento e ne riconosce alcuni comportamenti sfruttando delle tecniche di analisi e processamento delle immagini. Le informazioni così ricavate vengono inviate e messe a disposizione delle applicazioni di alto di livello sviluppate secondo necessità ed obiettivi differenti.

In tale contesto è stata messa a punto una applicazione di gioco. L'utente sarà invitato a muoversi sul pavimento per "scoprire" una immagine misteriosa. Mediante un apposito telecomando sarà in grado di rispondere alla domanda posta sullo schermo, non prima però di aver rilevato l'immagine o la porzione di immagine necessaria. Le immagini possono essere di tipo diverso. In figura 7 sono riportate alcune immagini prese durante una sessione di gioco il cui obiettivo era la risoluzione di semplici operazioni matematiche.

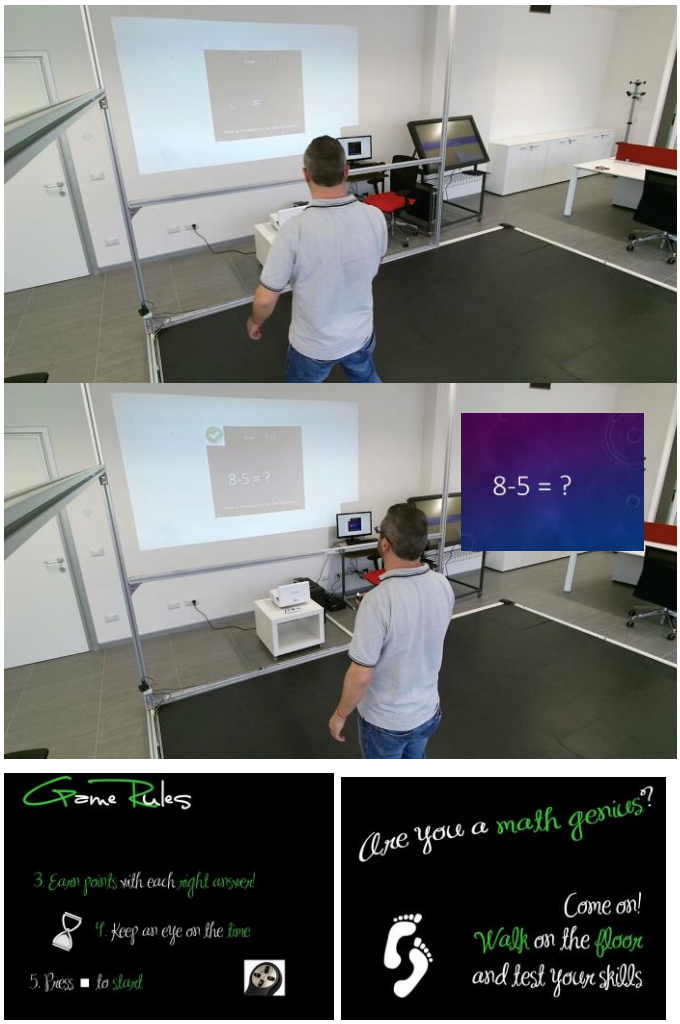


Fig. 7: sessione di gioco su pavimento interattivo e corrispondente schermate proiettate

ACKNOWLEDGMENT

Parte delle attività sono state svolte all'interno del laboratorio Florim del centro Softech-ICT, finanziato da Florim Ceramiche SpA.

REFERENCES

[1] R. Vezzani, M. Piccardi, R. Cucchiara, "An efficient Bayesian framework for on-line action recognition" in Proceedings of the 16th International Conference on Image Processing (ICIP 2009), Cairo, Egypt, Nov. 7-11, 2009

[2] Lulu Chen, Hong Wei, James Ferryman, A survey of human motion analysis using depth imagery, Pattern Recognition Letters, Volume 34, Issue 15, 1 November 2013, Pages 1995-2006, ISSN 0167-8655, http://dx.doi.org/10.1016/j.patrec.2013.02.006.

[3] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, A. Blake Efficient human pose estimation from single depth images IEEE Trans. Pattern Anal. Machine Intell. (2012), p. 1

[4] Rabiner, L., "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE , vol.77, no.2, pp.257,286, Feb 1989

[5] Dian Gong, Gérard Medioni, Sikai Zhu, and Xuemei Zhao. 2012. Kernelized temporal cut for online temporal segmentation and recognition. In Proceedings of the 12th European conference on Computer Vision - Volume Part III (ECCV'12), Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 229-243.

[6] R. Vezzani, M. Lombardi, A. Pieracci, P. Santinelli, R. Cucchiara, "A General-Purpose Sensing Floor Architecture for Human-Environment interaction" in press on ACM Transactions on Interactive Intelligent Systems, 2015

[7] International patent WO 2014141166-A1. Title: "Substrate for a sensitive floor and method for displaying loads on the substrate. Publication date: 18/09/2014. Depository: Claudio Lucchese. Inventors: Rita Cucchiara, Martino Lombardi, Augusto Pieracci, Paolo Santinelli, Roberto Vezzani.