Un robot cognitivo, basato su percezione visuale e uditiva , cooperante con umani in ambienti produttivi reali: Esperienze del nodo AI3S dell'Universita' di Salerno

Stefano Bini, Antonio Greco, Vincenzo Claudio Pierro, Antonio Roberto, Alessia Saggese, Mario Vento

Dipartimento DIEM, Università degli Studi di Salerno
{sbini, agreco, vpierro, aroberto, asaggese, mvento}@unisa.it

Abstract

Un robot cognitivo è un robot dotato di capacità percettive, di ragionamento e di interazione intelligente con l'ambiente e le persone che lo circondano.In tale contesto si colloca l'assistente robotico proposto dal MIVIA Lab, in grado di percepire l'ambiente mediante tecniche avanzate di analisi video e audio, di interpretare svariati intenti e richieste dell'umano tramite sofisticati algoritmi di intelligenza artificiale e di muoversi in sicurezza su una linea di produzione industriale, intraprendendo eventualmente un dialogo intelligente con l'operatore, al fine di assistere con un atteggiamento proattivo uno o più operatori umani nel proprio lavoro.

1 Introduzione

Le prime applicazioni della robotica in ambito industriale, che risalgono agli anni settanta, si ritrovano nel campo automobilistico, in cui i primi robot (dotati di semplici pinze) furono inseriti nelle linee di produzione delle fabbriche al fine di saldare in modo automatico i pezzi dell'automobile. Oggigiorno, l'utilizzo di robot antropomorfi sulle linee di produzione è uno standard: tali robot sono infatti utilizzati per effettuare le più svariate operazioni: forature, smerigliatura, fresature, verniciature, smaltature etc.

Si possono identificare tre differenti tipologie di robot, sulla base del loro livello di autonomia: robot di 1° livello, programmati per svolgere operazioni ripetitive ma che richiedono una elevata precisione; il software di controllo specifica quindi staticamente accelerazione, velocità etc. robot di 2° livello, capaci di adattarsi automaticamente a eventuali variazioni dell'ambiente circostante, verificando ad esempio la presenza e la posizione di un pezzo su cui operare sulla linea di produzione e modificando dinamicamente la traiettoria. robot di 3° livello, capaci di prendere decisioni in modo autonomo avvalendosi di avanzati algoritmi di intelligenza artificiale e machine learning.

I robot di 3° livello, non utilizzati oggi nelle linee di produzione, attengono a quella che viene definita "Cognitive Robotics". Questa disciplina mira alla definizione di robot intelligenti, capaci di: (i) acquisire informazioni sull'ambiente attraverso l'utilizzo di sensori (ad esempio telecamere e microfoni); (ii) ragionare, attraverso l'impiego di avanzate tecniche di intelligenza artificiale; (iii) agire, interagendo con l'ambiente e con le persone che popolano l'ambiente, sulla base della conoscenza che si è appresa. Perception, Reasoning and Action diventano pertanto gli elementi cardine per la transizione da robot di 1° livello a *robot cognitivo*.

In tale contesto, le attività del MIVIA Lab mirano alla progettazione e alla realizzazione di un robot cognitivo che possa fungere da *assistente* per l'operatore umano addetto alla linea di produzione. Tale assistente robotico ha tre compiti principali da svolgere: i) risposta a comandi o richieste dell'operatore umano, vocali o impartiti tramite gesti, localizzando e riconoscendo l'identità dell'operatore; ii) interazione proattiva e personalizzata con gli operatori umani sulla linea di produzione, al fine di migliorare le condizioni lavorative; iii) controllo ambientale, riconoscendo suoni e/o rumori di interesse, così da intervenire prontamente in caso di necessità.

Tali compiti, già tutt'altro che banali, sono resi particolarmente complicati dall'ambiente in cui l'assistente robotico deve operare. Infatti, il movimento del robot e quindi della telecamera, può causare la presenza di rumore e sfocature sulle immagini; ciò comporta la necessità di progettare e realizzare algoritmi di video-analisi robusti a tali perturbazioni. Inoltre, la linea di produzione e il movimento stesso del robot possono aggiungere rumore di entità non trascurabile sul segnale audio acquisito dal microfono; pertanto, anche gli algoritmi di audio-analisi per la localizzazione e il riconoscimento degli operatori e per il rilevamento di suoni di interesse devono essere accurati ed efficaci anche in presenza di tali disturbi. Infine, gli operatori potrebbero non pronunciare i comandi sempre allo stesso modo, rendendo quindi necessario dotare l'assistente robotico della capacità di interpretare richieste espresse in maniera diversa ma riferite allo stesso intento.

E' necessario inoltre considerare che l'assistente robotico ha a bordo l'hardware per effettuare l'elaborazione di audio e immagini in tempo reale. L'esigenza di massimizzare l'autonomia energetica del robot e di minimizzare il carico da trasportare implica la necessità di scegliere come dispositivo di elaborazione un sistema embedded a basso consumo. Pertanto, gli algoritmi sopra citati dovranno anche essere progettati e ottimizzati in modo da effettuare l'elaborazione in tempo reale con risorse di calcolo (CPU, RAM, GPU, disco) limitate.

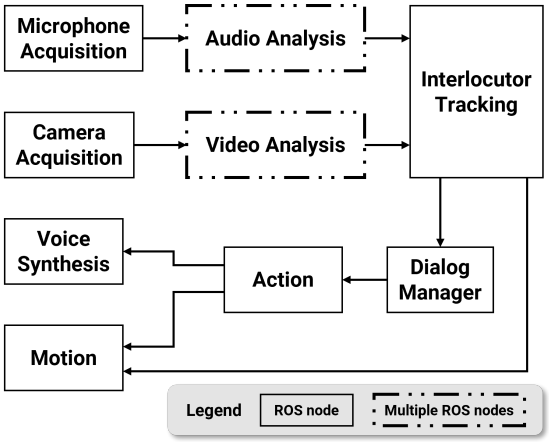

Figura 1: Architettura software basata su ROS dell'assistente robotico. Immagini e audio sono acquisiti in tempo reale mediante la telecamera e il microfono e vengono utilizzati per localizzare e profilare l'interlocutore e interpretarne i comandi e riconoscere suoni di interesse. Per semplificare la visualizzazione, tutti i moduli relativi all'analisi audio e video sono stati aggregati ad alto livello in un singolo modulo. Il modulo di tracking dell'interlocutore filtra i comandi pronunciati da altri operatori e permette al robot di avvicinarsi all'operatore. Il gestore del dialogo predice la prossima azione da fare (movimento o dialogo) concorde con la richiesta (intento) dell'operatore. L'ultimo nodo (Action) è, infine, incaricato di svolgere l'azione (Voice Synthesis e Motion).

2 L'assistente robotico del MIVIA Lab

L'architettura software dell'assistente robotico del MIVIA Lab è stata progettata e sviluppata attraverso l'utilizzo del framework ROS (Robotic Operating System) ed è stata ottimizzata per eseguire la pipeline di elaborazione in tempo reale a bordo del sistema embedded NVIDIA Jetson Xavier NX. Ogni modulo software, rappresentato in Figura 1, è stato progettato come un nodo ROS.

I nodi Microphone Acquisition e Camera Acquisition sono dedicati all'acquisizione in tempo reale di audio e immagini da microfono e telecamera. I due flussi di dati vengono analizzati da pipeline di elaborazione indipendenti, ovvero Audio Analysis e Video Analysis, al fine di estrarre informazioni di alto livello sull'interlocutore e sull'ambiente circostante. In particolare, i moduli di Audio Analysis si occupano della localizzazione (speaker localization) e del riconoscimento dell'interlocutore (speaker recognition) a partire dalla sua voce, del riconoscimento delle parole (speech recognition) e dei comandi pronunciati (speech command recognition) ed il rilevamento di suoni di interesse (sound event recognition). Analogamente, i moduli di Video Analysis utilizzano reti neurali profonde per il rilevamento (face detection), la profilazione (gender recognition, age estimation, emotion recognition) e il tracking dell'interlocutore (face tracking) e il riconoscimento di comandi impartiti tramite gesti.

I risultati dell'analisi audio e video sono utilizzati dal nodo Interlocutor Tracking per filtrare tutti comandi pronunciati da altri operatori nell'ambiente di lavoro e per seguire l'operatore durante lo svolgimento di particolari mansioni. Il Dialog Manager, invece, sfrutta tutte le informazioni estratte dai moduli precedenti per comprendere i comandi espressi in linguaggio naturale e i suoni ambientali e selezionare la prossima azione da effettuare. Inoltre, tale modulo si avvale degli attributi facciali dell'interlocutore, quali età, genere ed emozioni, al fine di personalizzare il modo con cui l'assistente robotico si pone quando interagisce in maniera proattiva con l'operatore umano. Infine, il nodo Action è responsabile di far eseguire le azioni motorie e/o di far pronunciare vocalmente all'assistente robotico le frasi più adatte.

2.1 Video Analysis

In un contesto industriale, è importante per l'operatore umano poter interagire con l'assistente robotico durante lo svolgimento delle proprie mansioni senza essere vincolato negli spostamenti. Le interazioni uomo macchina che solitamente vincolano l'interlocutore a posizionarsi di fronte al robot rischiano di diventare un impedimento in specifici campi applicativi come quello in esame [Saggese *et al.*, 2019b]. Al fine di evitare ciò, l'assistente robotico è dotato di un modulo di tracking dei volti [Carletti *et al.*, 2018] [Di Lascio *et al.*, 2013] che, insieme con il modulo di detection del volto [Greco *et al.*, 2021d], consente di seguire l'operatore umano nelle sue mansioni e di semplificare l'interazione. Individuato l'operatore, il robot può analizzare le sue caratteristiche facciali per riconoscere genere [Carletti *et al.*, 2019b] [Foggia *et al.*, 2019] ed età [Carletti *et al.*, 2019a] e personalizzare l'interazione di conseguenza. Inoltre, può raccogliere in tempo reale feedback sullo stato emozionale dell'operatore [Greco *et al.*, 2019a], distogliendolo per esempio da momenti di rabbia e noia. E' possibile, inoltre, riconoscere i gesti, in modo da permettere all'operatore di impartire comandi immediati all'assistente robotico mediante il solo movimento delle mani; o ancora analizzare le pose [Brun *et al.*, 2016][Saggese *et al.*, 2019a], al fine ad esempio di riconoscere posture errate che potrebbero causare problemi di salute a lungo termine.

Tutti gli algoritmi sopra citati sono stati valutati ed ottimizzati anche in condizioni critiche di funzionamento. Ad esempio, l'algoritmo di rilevamento dei volti è stato addestrato anche con volti parzialmente occlusi, al fine di garantire un'elevata accuratezza (oltre 90%) anche con volti non perfettamente visibili e in presenza di mascherine [Greco *et al.*, 2021d]. Le reti neurali profonde per il riconoscimento del genere [Greco *et al.*, 2020c], dell'età [Greco *et al.*, 2021c] e delle emozioni, che ottengono performance allo stato dell'arte in condizioni ideali (99% di accuratezza sul genere, MAE pari a 2 sull'età, quasi 90% di accuratezza sulle emozioni), sono state valutate anche in presenza di 15 diversi tipi di corruzioni delle immagini e con occlusioni dovute alle mascherine [Greco *et al.*, 2021d], dimostrando un'ottima resilienza a tali disturbi.

Come anticipato in precedenza, gli algoritmi utilizzati dall'assistente robotico devono essere non solo efficaci, ma anche efficienti. A tale proposito, l'algoritmo per il riconoscimento del genere è stato selezionato dopo un'attenta analisi del compromesso tra accuratezza e complessità computazionale [Greco *et al.*, 2020b], mentre la rete neurale per la stima dell'età è stata ottenuta con una tecnica di ottimizzazione nota come knowledge distillation [Greco *et al.*, 2021c]. Tale ottimizzazione consente di eseguire tutti gli algoritmi con-

temporaneamente ed in tempo reale sul sistema embedded disponibile a bordo dell'assistente robotico. Per ridurre ulteriormente il carico computazionale e l'utilizzo di risorse, è stata inoltre progettata e realizzata un'unica rete multi-task in grado di effettuare tutti i compiti di face analysis; l'accuratezza nei vari task rimane pressochè invariata, riducendo di un fattore tra 2.5 e 4 il tempo di elaborazione e la memoria occupata.

2.2 Audio Analysis

Insieme alla capacità di percepire l'ambiente circostante tramite l'utilizzo dell'apparato sensoristico visuale, l'assistente robotico è dotato di un array di microfoni che gli permette di acquisire il flusso audio e, di conseguenza, di comprendere le parole dell'operatore e di riconoscere i suoni di interesse.

Come anticipato, il primo compito dell'assistente robotico è di riconoscere le parole pronunciate dall'operatore umano. Questo modulo è di fondamentale interesse perché rende possibile l'interazione tra uomo e robot senza la necessità che l'operatore umano si trovi nel campo visuale di quest'ultimo. In combinazione con il modulo di riconoscimento delle parole, l'assistente robotico adopera un sistema di localizzazione della sorgente vocale che gli consente di voltarsi e dirigersi nella direzione dell'operatore umano che lo ha interpellato. L'algoritmo di localizzazione [Saggese *et al.*, 2017] consente di stimare la direzione di provenienza del comando vocale con un errore medio di circa 4° in condizioni particolarmente sfavorevoli (0 dB, ovvero quando l'energia del rumore ambientale è pari a quella della voce dell'operatore).

Inoltre, è importante per un assistente robotico identificare l'operatore umano con cui sta interagendo al fine di permettere o meno l'esecuzione di particolari comandi. Infatti, data la larga portata dei microfoni potrebbe accadere che eventuali conversazioni nell'ambiente di lavoro vengano interpretate come comandi. Al fine di evitare tale inconveniente, è possibile utilizzare l'identità della persona che il robot ha di fronte per filtrare tutte le frasi pronunciate da altre persone. Tale sistema può essere inoltre utilizzato per motivi di sicurezza al fine di gestire una gerarchia del personale in grado di dare specifici tipi di comandi all'assistente robotico. In aggiunta ai requisiti di robustezza dei sistemi di analisi audio che lavorano nell'ambiente industriale, nel caso dell'identificazione da segnali vocali è importante tener conto dell'assenza di dati per l'addestramento del sistema; è difficile infatti raccogliere un gran numero di campioni audio per ognuno degli operai della fabbrica. Il modulo di Speaker Recognition a bordo dell'assistente robotico è stato quindi progettato per adattarsi facilmente a nuovi utenti, ovvero operatori umani. Inoltre, per valutare l'accuratezza e la robustezza dei nostri algoritmi, abbiamo acquisito e pubblicato un benchmark per Speaker Recognition [Roberto *et al.*, 2019] che tenesse in considerazione le suddette problematiche. Il sistema equipaggiato a bordo dell'assistente robotico è in grado di riconoscere la persona che ha pronunciato il comando con un errore medio del 2% avendone sentito la voce una sola volta.

Nell'ottica di aiutare l'essere umano nello svolgimento di lavori monotoni e noiosi, abbiamo dotato l'assistente robotico di un sistema per il controllo ambientale, in particolare, il rilevamento di suoni in ambito sorveglianza. Tale sistema permette al robot di rilevare suoni quali urla, vetri rotti e spari [Roberto *et al.*, 2020],[Greco *et al.*, 2019b], [Greco *et al.*, 2021b] e di lanciare prontamente un segnale di allarme all'operatore umano. In questo modo, il nostro assistente robotico può sorvegliare aree molto vaste anche in condizioni di luminosità scarse o assenti. Il punto di forza di tale componente è la sua alta accuratezza nel rilevamento dei suoni di interesse anche in condizioni rumorose particolarmente svantaggiose [Greco *et al.*, 2021a] [Greco *et al.*, 2020a], raggiungendo un tasso di rilevamento pari al 99.42% in condizioni ambientali caratterizzate da un basso rapporto segnale-rumore (fino a 5 dB). Allo stesso tempo, il tasso di falsi positivi è infierore all' 1%.

Tutti gli algoritmi di analisi audio sono stati progettati tenendo conto degli stringenti vincoli computazionali sopra descritti e possono essere eseguiti in real-time a bordo dell'operatore robotico insieme a tutti gli altri moduli.

2.3 Dialog Manager

L'assistente robotico è dotato di un sistema di comprensione del linguaggio naturale che consente agli operatori umani di rivolgersi a lui attraverso comandi formulati in liguaggio comune. Ciò significa che i comandi non sono predefiniti ma possono essere pronunciati in diversi modi senza alcun vincolo, come ad esempio, "portami il cacciavite" o "prendimi il giravite". Nonostante la complessità del problema, l'operatore robotico è in grado di riconoscere gli intenti dell'operatore umano con un'accuratezza del 97%.

Oltre ad un algoritmo di comprensione del linguaggio naturale, l'assistente robotico è dotato di un gestore del dialogo che gli permette di decidere se eseguire il comando impartito dall'operatore umano o formulare delle domande di disambiguazione (per esempio "Devo portarti il cacciavite, giusto?") in caso di bassa confidenza nel riconoscimento dell'intento.

Tale modulo presenta inoltre una componente sociale al fine di migliorare le condizioni lavorative dell'operatore. Infatti, sulla base delle emozioni espresse da quest'ultimo il robot può decidere di intrapprerendere in maniera proattiva una conversazione finalizzata a fare da supporto o passatempo in momenti di rabbia e noia. Inoltre, informazioni quali età e genere vengono sfruttate per personalizzare i modi con cui il robot si interfaccia con l'operatore. Inoltre, al fine di rendere l'operatore robotico capace di apprendere come e quali contenuti presentare all'operatore umano, sono state introdotte delle tecniche basate su *life-long learning*, ovvero tecniche di apprendimento che continuano a tempo di esecuzione. In particolare, il cambio o meno dello stato emotivo dell'operatore umano a valle della conversazione avviata può essere utilizzato come segnale di retroazione per capire se le azioni intraprese siano state efficaci o meno. Infine, l'assistente robotico è caratterizzato da una bassa latenza di elaborazione. Il tempo che intercorre tra l'acquisizione del campione vocale fino all'avvio della corrispondente azione è mediamente di 465 millisecondi. Tale tempo è inferiore al tempo di risposta medio umano (maggiore di un secondo in media per conversazioni quotidiane) ed è un risultato non trascurabile se si considera che tutti gli algoritmi di cui prima vengono eseguiti a bordo del sistema embedded NVIDIA Jetson Xavier NX.

3 Ringraziamenti

Questo lavoro scientifico è stato parzialmente finanziato dal progetto europeo HH2020-EU.2.1.1 Flexible Assembly Manufacturing with human-robot collaboration and digital twin models (FELICE) ID: 101017151 e dal progetto PRIN 20172BH297 002 CUP D44I17000200005.

Riferimenti bibliografici

[Brun *et al.*, 2016] Luc Brun, Gennaro Percannella, Alessia Saggese, e Mario Vento. Action recognition by using kernels on aclets sequences. *Computer Vision and Image Understanding*, 144:3–13, 2016. Individual and Group Activities in Video Event Analysis.

[Carletti *et al.*, 2018] Vincenzo Carletti, Antonio Greco, Alessia Saggese, e Mario Vento. Multi-object tracking by flying cameras based on a forward-backward interaction. *IEEE Access*, 6:43905–43919, 2018.

[Carletti *et al.*, 2019a] Vincenzo Carletti, Antonio Greco, Gennaro Percannella, e Mario Vento. Age from faces in the deep learning revolution. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2113–2132, 2019.

[Carletti *et al.*, 2019b] Vincenzo Carletti, Antonio Greco, Alessia Saggese, e Mario Vento. An effective real time gender recognition system for smart cameras. *Journal of Ambient Intelligence and Humanized Computing*, 11(6):2407–2419, March 2019.

[Di Lascio *et al.*, 2013] Rosario Di Lascio, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, e Mario Vento. A real time algorithm for people tracking using contextual reasoning. *Computer Vision and Image Understanding*, 117(8):892–908, 2013.

[Foggia *et al.*, 2019] Pasquale Foggia, Antonio Greco, Gennaro Percannella, Mario Vento, e Vincenzo Vigilante. A system for gender recognition on mobile robots. In *Proceedings of the 2nd international conference on applications of intelligent systems*, pages 1–6, 2019.

[Greco *et al.*, 2019a] Antonio Greco, Antonio Roberto, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Emotion analysis from faces for social robotics. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 358–364, 2019.

[Greco *et al.*, 2019b] Antonio Greco, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Sorenet: A novel deep network for audio surveillance applications. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*, pages 546–551. IEEE, 2019.

[Greco *et al.*, 2020a] Antonio Greco, Nicolai Petkov, Alessia Saggese, e Mario Vento. Aren: A deep learning approach for sound event recognition using a brain inspired representation. *IEEE Transactions on Information Forensics and Security*, 15:3610–3624, 2020.

[Greco *et al.*, 2020b] Antonio Greco, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff. *IEEE Access*, 8:130771–130781, 2020.

[Greco *et al.*, 2020c] Antonio Greco, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Gender recognition in the wild: a robustness evaluation over corrupted images. *Journal of Ambient Intelligence and Humanized Computing*, 12(12):10461–10472, December 2020.

[Greco *et al.*, 2021a] Antonio Greco, Antonio Roberto, Alessia Saggese, e Mario Vento. DENet: a deep architecture for audio surveillance applications. *Neural Computing and Applications*, January 2021.

[Greco *et al.*, 2021b] Antonio Greco, Antonio Roberto, Alessia Saggese, e Mario Vento. Which are the factors affecting the performance of audio surveillance systems? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7876–7883, 2021.

[Greco *et al.*, 2021c] Antonio Greco, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Effective training of convolutional neural networks for age estimation based on knowledge distillation. *Neural Computing and Applications*, April 2021.

[Greco *et al.*, 2021d] Antonio Greco, Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Performance assessment of face analysis algorithms with occluded faces. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 472–486. Springer International Publishing, 2021.

[Roberto *et al.*, 2019] Antonio Roberto, Alessia Saggese, e Mario Vento. A challenging voice dataset for robotic applications in noisy environments. In *Computer Analysis of Images and Patterns*, pages 354–364. Springer International Publishing, 2019.

[Roberto *et al.*, 2020] Antonio Roberto, Alessia Saggese, e Mario Vento. A deep convolutionary network for automatic detection of audio events. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, pages 1–6, 2020.

[Saggese *et al.*, 2017] Alessia Saggese, Nicola Strisciuglio, Mario Vento, e Nicolai Petkov. A real-time system for audio source localization with cheap sensor device. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7, 2017.

[Saggese *et al.*, 2019a] Alessia Saggese, Nicola Strisciuglio, Mario Vento, e Nicolai Petkov. Learning skeleton representations for human action recognition. *Pattern Recognition Letters*, 118:23–31, 2019. Cooperative and Social Robots: Understanding Human Activities and Intentions.

[Saggese *et al.*, 2019b] Alessia Saggese, Mario Vento, e Vincenzo Vigilante. Miviabot: A cognitive robot for smart museum. In *International Conference on Computer Analysis of Images and Patterns*, pages 15–25. Springer, 2019.