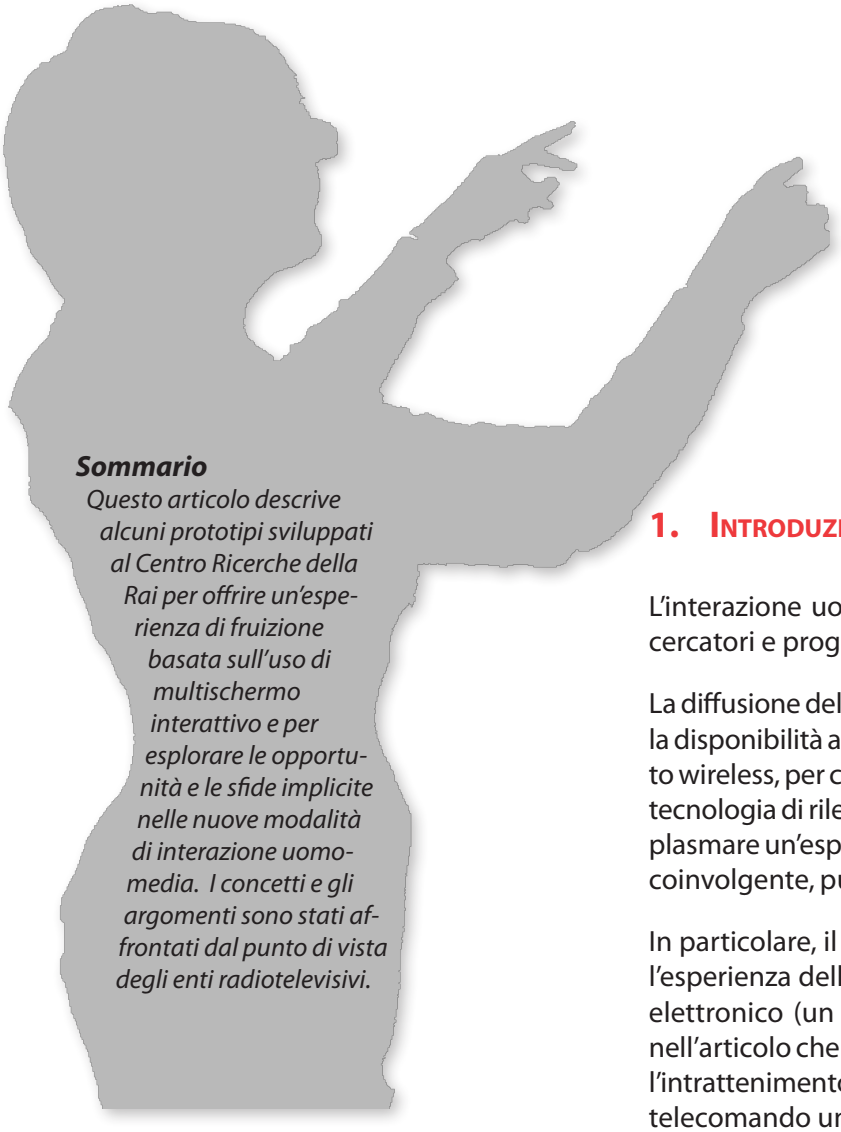
Nuovi paradigmi per l'interazione uomo-media:

la "TV aumentata" e "Senza telecomando"

Roberto **Iacoviello** e Paola **Sunna,**
Rai - Centro Ricerche e Innovazione Tecnologica



Sommario

Questo articolo descrive alcuni prototipi sviluppati al Centro Ricerche della Rai per offrire un'esperienza di fruizione basata sull'uso di multischermo interattivo e per esplorare le opportunità e le sfide implicite nelle nuove modalità di interazione uomo-media. I concetti e gli argomenti sono stati affrontati dal punto di vista degli enti radiotelevisivi.

1. Introduzione

L'interazione uomo-media è un argomento caldo tra ricercatori e progettisti.

La diffusione delle nuove tecnologie "abilitanti" (a esempio la disponibilità a buon mercato di dispositivi di puntamento wireless, per comandi vocali, telecamere 3D, basati sulla tecnologia di rilevamento del movimento, ...) può aiutare a plasmare un'esperienza "arricchita", molto più interattiva e coinvolgente, pur mantenendo semplicità e accessibilità.

In particolare, il controllo gestuale potrebbe rimodellare l'esperienza della visione nel soggiorno al di là del gioco elettronico (un esempio è il sistema Kinect [1] trattato nell'articolo che segue) permettendo agli utenti di gestire l'intrattenimento TV a mani nude, rendendo il tradizionale telecomando un dispositivo elettronico del passato.

L'attenzione è posta sulla descrizione dei prototipi realizzati al Centro Ricerche della Rai allo scopo di:

- ✓ permettere la confluenza dei servizi primari di diffusione con quelli aggiuntivi identificati come "contenuti e interattività", per offrire una presentazione più attraente e "aumentata" su terminali mobili via IP (Internet Protocol);
- ✓ abilitare il controllo gestuale nel dominio TV.

Una particolare attenzione è rivolta ai requisiti, dal punto di vista degli enti radiotelevisivi, necessari per fornire agli utenti questi nuovi tipi di esperienza arricchita e coinvolgente.

La Rai ha ricevuto il Premio ICMT 2011 per la categoria "convergenza Media-driven" per il progetto "Rai+" che comprende i *concept* "Chi sei" e "Gioca con me", oggetto di questo articolo.

2. Gli strumenti tecnologici

Incominciano ad apparire nelle case set-top-box ibridi e televisori in grado di ricevere sia i programmi televisivi convenzionali sia i contenuti a banda larga e nuovi comportamenti di fruizione sono influenzati dalla disponibilità di tariffe *flat* per la connessione, che consentono una nuova realtà multi-schermo, fatta di televisione, ma anche PC e terminali mobili.

E' iniziata l'era dello stile di vita *always-on*, grazie ad una connessione onnipresente e ininterrotta a Internet (anche tra gli stessi dispositivi): le generazioni nate dopo il 1990 non possono rinunciare alle attività di social network e all'uso dei dispositivi mobili che accompagnano le loro attività quotidiane.

E' la tecnologia a rendere possibile questa esperienza visiva senza interruzioni, ed una forte creatività è il mezzo che vincola gli spettatori in questa navigazione continua tra questi mondi (media radiodiffusi, media on-line, altri spettatori "sociali", ...) diversi, ma non ancora "in sincrono": attraverso un viaggio trans-mediale.

Di conseguenza le società radiotelevisive cercano di sfruttare sempre più queste stimolanti opportunità per promuovere il loro marchio attraverso un'offerta arricchita ed interattiva, utilizzando la distribuzione mediante più canali e dispositivi sofisticati. Si sta rivelando sempre più importante, per rendere più profonda la relazione tra i proprietari dei contenuti ed il loro pubblico, la possibilità di fornire all'utente, mentre guarda il programma in tempo reale o in differita, funzionalità quali indice di gradimento, quiz, risultati di sondaggi, voti, "condivisione" sociale, informazioni supplementari su dispositivi mobili (tablet e smart phone) .

Ma la sola interpretazione creativa di questi comportamenti individuali e sociali non sarebbe sufficiente se non si ottenesse come risultato anche un adeguato controllo dell'*entertainment*!

Nell'era "Internet delle cose"[Nota 1], un dispositivo digitale (set-top-box, console per videogiochi, televisore ...) è in grado di vedere, di capire e interagire con il mondo circostante aprendo così la strada a nuovi paradigmi di comunicazione uomo-macchina. Queste caratteristiche sono tipiche delle nuove telecamere ("vedere" attraverso i sensori) e schermi secondari ("vedere" attraverso la connettività IP) utilizzati per costruire i nostri *concept*.

Acronimi e sigle

API	Application Programming Interface
IoT	Internet of Things
MIT	Massachusetts Institute of Technology
MPEG	Movig Picture Experts Group (ISO/IEC JTC 1/SC 29/WG 11)
OCR	Optical character recognition
RSS	RDF Site Summary Really Simple Syndication
SIFT	Scale-Invariant Feature Transform
ToF	Time of Flight

Nota 1 - Internet delle cose (o Internet degli oggetti), è un neologismo che trae origine dall'inglese Internet of Things riferito all'estensione di Internet al mondo degli oggetti e dei luoghi concreti. Il concetto dell'internet delle cose è attribuito all'Auto-ID Center, fondato nel 1999 e da allora con sede al MIT.

E' possibile, ad esempio, catturare i movimenti degli utenti in tempo reale con un' unica telecamera 3D in grado di rilevare la profondità. Incorporare tali dispositivi nei prodotti elettronici di consumo e adottare accurati algoritmi di riconoscimento delle forme permette ai creativi di immaginare una incredibile serie di nuove possibilità e di sviluppare nuove applicazioni interattive, di tipo immersivo e intuitivo, basate sui gesti dell'utente.

Gli "schermi secondari" sono i terminali portatili avanzati e collegati a internet (*smartphone*, *tablet*, ...) che stanno diventando i compagni di ogni giorno degli utenti, anche quando guardano i tradizionali programmi televisivi.

Dal punto di vista del *broadcaster*, gli schermi secondari favoriranno la creazione di *mini-format* associati al *brand*, non a detrimento dell'*audience* per il programma televisivo primario, ma, al contrario, in grado di creare opportunità di profitto, consentendo di focalizzare l'attenzione del pubblico sul *brand*, di analizzare ciò che vede l'utente, di offrire contenuti *premium*, ...

3. Dal concetto allo sviluppo

Si esaminano le interfacce multimodali per il controllo dell'esperienza *entertainment* sullo schermo del televisore e la presentazione in forma interattiva delle informazioni su schermi secondari (realizzata per le piattaforme Android e iOS).

3.1 Senza telecomando

I terminali mobili (*tablet*, *smartphone*, *notebook* ...) si sono evoluti dando origine a dispositivi potenti, utilizzati anche per controllare in remoto lo schermo del televisore principale, ma le loro capacità di interfacciamento risultano in realtà, per questo scopo, piuttosto limitate. Sono possibili scenari totalmente nuovi, grazie alle interfacce gestuali.

Si è studiato e realizzato un sistema di riconoscimento gesti per fornire in modo naturale l'interazione con una interfaccia utente di tipo grafico. I gesti della mano, eseguiti nello spazio libero, sono riconosciuti da algoritmi specifici, e danno origine a nuove metafore per l'interazione tra utenti e dispositivi di *home entertainment*, possibilmente senza richiedere la calibrazione del sistema ed un periodo di *training* per l'utente.

I componenti del sistema sviluppato per il riconoscimento dei gesti basato sulla visione sono rappresentati in figura 1.

L'acquisizione avviene tramite una telecamera ToF che fornisce la profondità d'immagine della scena (la profondità è la distanza tra la telecamera e il punto dell'oggetto corrispondente a quello dei *pixel*). Tale immagine è l'*input* di una fase di rilevamento che, dopo il filtraggio del rumore e alcune elaborazioni per eliminare eventuali imperfezioni presenti, fornisce le coordinate 2D della posizione e la distanza della mano dalla telecamera (coordinata Z). L'ultima fase prevede l'inseguimento (*tracking*) della mano

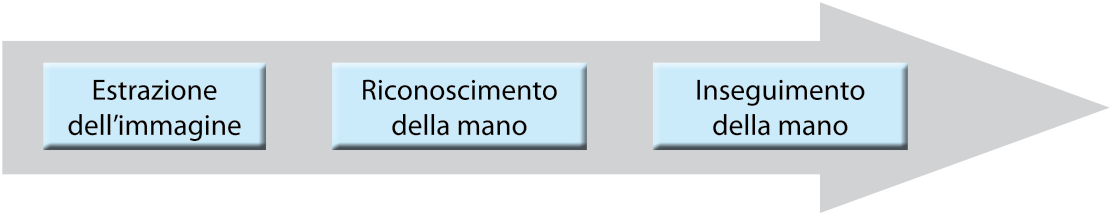


Fig.1 - Componenti del software sviluppato per il riconoscimento sulla base della visione dei gesti.

Fig. 2 - Interfaccia grafica d'utente basata sulla guida gestuale.

ottenuto con una fase di rilevamento continuo. E così possibile individuare due tipi di eventi: il puntamento ottenuto mediante le coordinate 2D e l'operazione di *click* in base alla distanza (coordinata Z).

L'interfaccia grafica sviluppata presso il Centro Ricerche comprende un menu di quattro elementi: Catch-up TV, Photo, News, TV (figura 2) selezionabili dall'utente con un *click*, muovendo una mano nell'ambito dello schermo. All'interno di ogni sezione l'utente può poi navigare puntando la mano e fare click per leggere un *feed RSS*, avviare un video, una sequenza di foto, richiedere l'avanzamento veloce o il riavvolgimento di un video.

3.2 Chi sei?

Le nostre applicazioni mobili a realtà aumentata mostrano informazioni aggiuntive, contestuali ai contenuti video sul televisore, su schermi secondari, che diventano una estensione "fisica" dello schermo principale.

Nell'uso ipotizzato, l'applicazione a realtà aumentata cerca di soddisfare la curiosità dell'utente di conoscere l'identità delle persone presenti sullo schermo mentre guarda il programma televisivo.

L'applicazione identifica automaticamente il personaggio ripreso con la fotocamera in dotazione al telefono la cui immagine è inviata a un *server* remoto. Il nucleo dell'algoritmo per questa applicazione è la SIFT [2] ampiamente utilizzata in *computer vision* per descrivere un'immagine come un insieme di caratteristiche rilevabili anche in caso di variazioni di scala dell'immagine, di rumore e di illuminazione.

Il sistema confronta specifici lineamenti del viso estratti dall'immagine con quelli memorizzati in un "database di caratteristiche facciali" e restituisce il nome del personaggio e una breve biografia (figure 3 e 4).

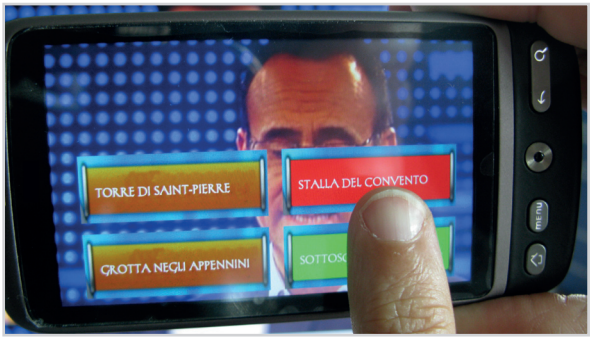
In questa fase, è unitilizzata una API in dotazione al dispositivo per realizzare la conversione da testo a parlato applicato alle didascalie sullo schermo.

Fig. 3 - L'applicazione "Chi sei?".

Fig. 4 - Nome del personaggio e relativa biografia.

Fig. 5 - L'applicazione "Gioca con me".

Fig. 6 - Risposte.



3,3 Gioca con me

Questo è il prototipo di un'applicazione che permette agli utenti di interagire nell'ambito di un programma a quiz durante la trasmissione televisiva, partecipando al gioco insieme ai concorrenti reali.

In questo scenario, l'utente scatta una foto (con il suo terminale mobile) della domanda in sovrapposizione sullo schermo del televisore, l'applicazione la invia al *server* remoto della Rai per l'elaborazione e riceve le possibili risposte che vengono poi presentate sul terminale mobile, in modalità realtà aumentata (figure 5-6).

Il riconoscimento ottico dei caratteri (OCR) è utilizzato per interpretare il testo della domanda a partire da una versione opportunamente ridimensionata della foto originale. Tale ridimensionamento è necessario per ridurre le dimensioni del file, e di conseguenza il tempo di elaborazione OCR.

Il testo estratto è confrontato con il database Rai che contiene il testo delle domande possibili, al fine di identificare le relative risposte.

4. Scenari futuri

Segue una sintesi delle principali problematiche emerse nel corso dello sviluppo dei prototipi precedentemente descritti e le implicazioni ad esse associate

4,1 Interfacce gestuali

Attualmente, Kinect è il solo dispositivo elettronico consumer disponibile sul mercato che permette di giocare con Xbox 360 utilizzando gesti e comandi vocali, senza la necessità di un controller.

Questo è solo l'inizio e, probabilmente lungo è il cammino prima di vedere questo tipo di interazioni adottate negli ambienti PC, SmartTV e Set Top Box. Tuttavia stanno emergendo in gran numero software proprietari e *open source* di tipo *middleware* in grado di rilevare caratteristiche facciali, gestualità, identificare l'utente, rilevare movimenti di parti del corpo dell'utente, interpretare la scena per consentire il controllo individuando la presenza di più utenti.

Gli elementi chiave per il successo di questi nuovi paradigmi di interazione sono: rilevamento e inseguimento privo di errori e senza interruzioni dei gesti, anche in presenza di occlusioni, rapidità di risposta. Possono apparire come dettagli agli enti radiotelevisivi interessati a fornire questo tipo di esperienza immersiva, ma in realtà è necessario sviluppare un set completo di strumenti e API agnostico nei confronti di hardware e software di base al fine di ridurre drasticamente i cicli di sviluppo. In particolare, queste API (ancora mancanti) permetterebbero agli sviluppatori di applicazioni di utilizzare una varietà di sensori e disposivi di controllo per rendere possibile all'utente l'impiego dei coman-

di gestuali per la navigazione (ad esempio nella scena televisiva, per i menu di giochi, nell'ambito delle interfacce Web); comandi come puntamento, selezione mediante click, spostamento di oggetti, onde, cerchi, ecc...

4,2 Riconoscimento facciale

La ricerca visuale implica prendere una foto di qualcosa o qualcuno, in base alla quale effettuare una *query* di ricerca, attendere tempi pari a secondi dovuti ai tempi di elaborazione, e ulteriori ritardi per il trasferimento di immagini di grandi dimensioni. Quindi una buona precisione, minimizzando la probabilità di falsi risultati ed una breve latenza sono le chiavi per ottenere l'adozione e la diffusione di questa tecnologia.

Molto impegnativo è l'investimento sulle strutture a supporto dei servizi: sono necessari miliardi di immagini e di metadati ad esse relative; queste informazioni devono essere acquisite, memorizzate e indicizzate, e diventa fondamentale la ricerca e il recupero di queste informazioni in modo efficiente dai database.

E' possibile raggiungere un più alto livello di prestazioni, ridurre il carico sulle reti *wireless* e migliorare l'interoperabilità mediante la definizione, su basi standard, di descrittori visuali compatti (utilizzando eventualmente supporti hardware per l'estrazione dei descrittori) e del corrispondente processo di estrazione, così come previsto da MPEG.

4,3 Sincronizzazione tra TV e schermi secondari

Il *concept* realizzato si basa sul servizio principale di tipo televisivo, i cui contenuti sono preregistrati, in modo che la diffusione del contenuto IP non pregiudichi la correttezza della competizione. L'approccio seguito è basato su OCR, potrebbe essere potenziato in futuro grazie ad un ridotto periodo di latenza e ad una maggiore precisione del motore dell'OCR, non richiede una stretta sincronizzazione temporale tra i contenuti principali e quelli secondari ed il formato quiz consente la gestione del ritardo variabile *end-to-end* tipico del percorso IP, grazie al fatto che l'utente necessita di un tempo significativo per selezionare la risposta giusta.

L'identificazione automatica del canale TV e dei relativi contenuti è fondamentale per permettere la perfetta sincronizzazione temporale tra i contenuti televisivi ed i servizi interattivi supplementari destinati ai terminali mobili, è ciò è ancora più vero nel caso di programmazione di tipo *Live TV*.

In alternativa all'uso dell'OCR, si può basare la sincronizzazione con il televisore principale sull'attivazione automatica di richieste di informazioni, ma ciò richiederebbe che la generazione dei metadati sia precisa e dettagliata a lato *back-end*. Occorre ulteriormente approfondire come effettuare la risincronizzazione tra le richieste generate e i dati fatti pervenire via IP al fine di convalidare le risposte fornite dagli utenti da casa.

5. Conclusioni

Sono in gran numero le possibilità e gli scenari che si possono aprire per gli enti televisivi grazie alla diffusione di prodotti multischermo e alle nuove forme di interazione uomo-media.

I *concept* illustrati permettono, in questo contesto, di aggiungere immersività e pervasività all'esperienza di intrattenimento domestico.

I sistemi qui descritti hanno scopo dimostrativo e richiedono ulteriori miglioramenti prima che sia possibile effettuare prove sul campo, ma rendono evidente come sia la tecnologia sia la creatività siano fondamentali per rendere possibile scenari più ricchi per l'intrattenimento televisivo.

Al fine di evitare una Babele causata della proliferazione di realizzazioni non compatibili fra loro, è necessario un processo di armonizzazione per molti degli aspetti presenti in questo scenario, quali il formato dei metadati, i set di comandi e di messaggi.

Inoltre, la cooperazione di applicazioni realizzate per più dispositivi non interoperabili potrebbe richiedere una quantità enorme di lavoro di sviluppo nel caso di proliferazione delle applicazioni, rendendo quindi cruciale la disponibilità di strumenti per la conversione automatica tra piattaforme diverse.

Ringraziamenti

Gli autori ringraziano Christian Culeddu (Gruppo Eurix) che ha attivamente contribuito a sviluppare i concetti e a Sabino Mantovano (del Centro Ricerche Rai) per il suo supporto tecnico.

Bibliografia

1. www.xbox.com/en-US/kinect
2. David G. Lowe: "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60, 2 (2004), pp. 91-110
3. mpeg.chiariglione.org/working_documents/explorations/cdvs/cdvscfp.zip