INTERAZIONE MULTIMODALE

Introduzione

Il campo di ricerca dell'interfaccia uomo-macchina si focalizza sul rendere l'interazione più agevole, più sicura, più efficace e senza interruzione per l'utente.

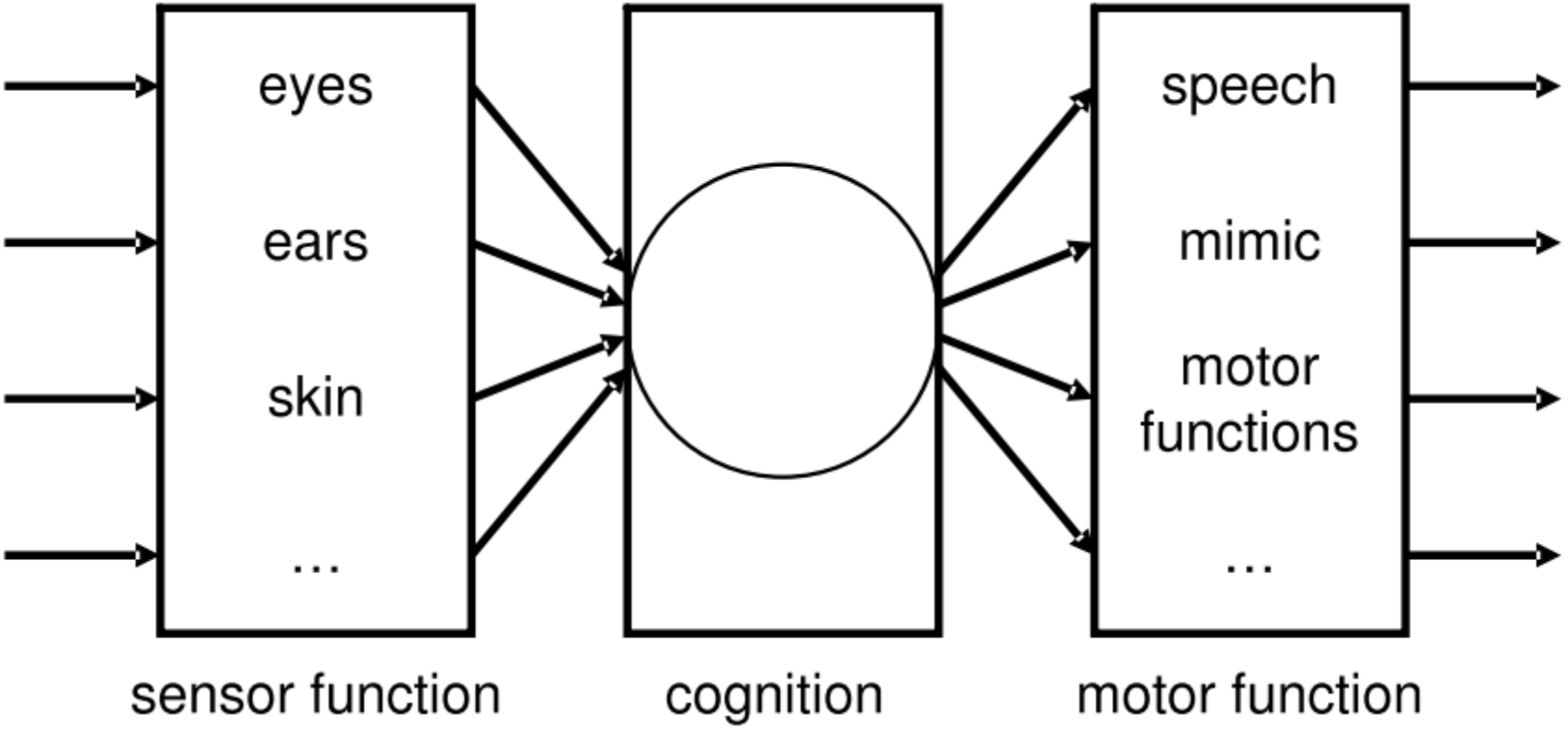
Generalmente, il termine interazione descrive l'influenza reciproca di diversi partecipanti nello scambio di informazioni. Tali informazioni vengono veicolate tramite mezzi diversi in modo bilaterale.

Un obiettivo importante del campo di ricerca dell'interfaccia uomo-macchina è quello di far convergere le modalità di interazione verso modalità di interazione familiari e ordinarie.

Nella comunicazione umana naturale, diversi canali di ingresso e di uscita sono combinati in modo multimodale.

INTERAZIONE MULTIMODALE
Canali di ingresso e di uscita

L'essere umano è capace di raccogliere, elaborare e esprimere l'informazione tramite canali differenti. Il canale di ingresso può essere descritto come *sensor function* o percezione, l'elaborazione delle informazioni può essere descritta come *cognition* e il canale di uscita può essere descritto come *motor function* (si veda la figura).



Canali di ingresso e di uscita riguardanti la raccolta, l'elaborazione e l'espressione delle informazioni per gli umani.

INTERAZIONE MULTIMODALE
Canali di ingresso e di uscita

Gli umani sono dotati di sei sensi e dei rispettivi organi di senso per raccogliere stimoli. Questi sensi sono definiti dalla fisiologia:

sense of sight	visual channel
sense of hearing	auditive channel
sense of smell	olfactory channel
sense of taste	gustatory channel
sense of balance	vestibular channel
sense of touch	tactile channel

L'essere umano possiede anche un ampio range di abilità associate al canale di uscita. Il canale di uscita può processare meno informazioni rispetto al canale di ingresso. L'informazione in uscita viene trasmessa dai canali uditivo, visivo e aptico (*haptic*). Il tattile come modalità di percezione viene distinto dall'aptico che invece è una modalità di output. In questo modo l'umano comunica con gli altri umani in modo semplice, efficace e robusto agli errori. Con un utilizzo integrato e sincronizzato di canali differenti egli si può adattare in modo flessibile alle abilità

INTERAZIONE MULTIMODALE
Canali di ingresso e di uscita

specifiche degli altri soggetti coinvolti nella conversazione e al contesto.

Tali modalità intuitive e naturali non sono state ancora del tutto trasmesse all'interfaccia uomo-macchina a causa di alcune limitazioni delle macchine riguardo al numero e alle prestazioni delle singole modalità di ingresso e di uscita.
Il termine modalità indica il tipo di canale di comunicazione utilizzato per trasmettere o acquisire informazioni.
Ad esempio, in alcuni ambiti, l'utente può trasmettere i suoi comandi ad una macchina solo mediante dispositivi di input standard, come il mouse o la tastiera; il feedback della macchina viene attuato ad esempio solo mediante il canale visivo o uditivo con monitor e altoparlanti. Quindi, in generale, le macchine utilizzano solo alcune modalità di interazione degli umani.

Quando vengono coinvolte due o più modalità, si parla di multimodalità.

INTERAZIONE MULTIMODALE
Principi

I sistemi multimodali hanno due caratteristiche principali:

- L'utente può comunicare con la macchina mediante diverse modalità di input e di output.
- I differenti canali possono interagire in modo appropriato.

Una possibile definizione di interfacce multimodali è: le interfacce multimodali combinano modalità di input naturali (esempi: parlato, scrittura, tatto, gesti manuali, sguardo, movimenti della testa e del corpo) in modo coordinato con l'uscita del sistema multimediale. Esse sono una classe di interfacce che mira a riconoscere le forme del linguaggio e del comportamento umano che si presentano e che includono una o più tecnologie basate sul riconoscimento (esempi: parlato, scrittura, visione).

Un'altra definizione è un'estensione della precedente definizione a sistemi che rappresentano e manipolano l'informazione da differenti canali di comunicazione umani che si trovano a differenti livelli di astrazione. Tali sistemi sono in grado di estrarre automaticamente il significato dei dati multimodali in input e viceversa producono informazioni percepibili a partire da rappresentazioni simboliche astratte.

INTERAZIONE MULTIMODALE
Principi

I sistemi multimodali beneficiano dei progressi che sono stati ottenuti e che si stanno ottenendo nel campo delle tecnologie di riconoscimento. Tali tecnologie permettono di comprendere forme del linguaggio e del comportamento umano.

Uno dei temi dominanti nell'organizzazione naturale degli input multimodali da parte degli utenti è la complementarità dei contenuti: ogni input contribuisce in modo significativo a diverse informazioni semantiche.
Le sequenze di informazione parziale devono essere fuse e possono essere interpretate solo congiuntamente.
La ridondanza di informazione è molto poco comune nella comunicazione umana. Alcune volte differenti modalità possono fornire in input contenuti che devono essere processati indipendentemente gli uni dagli altri.

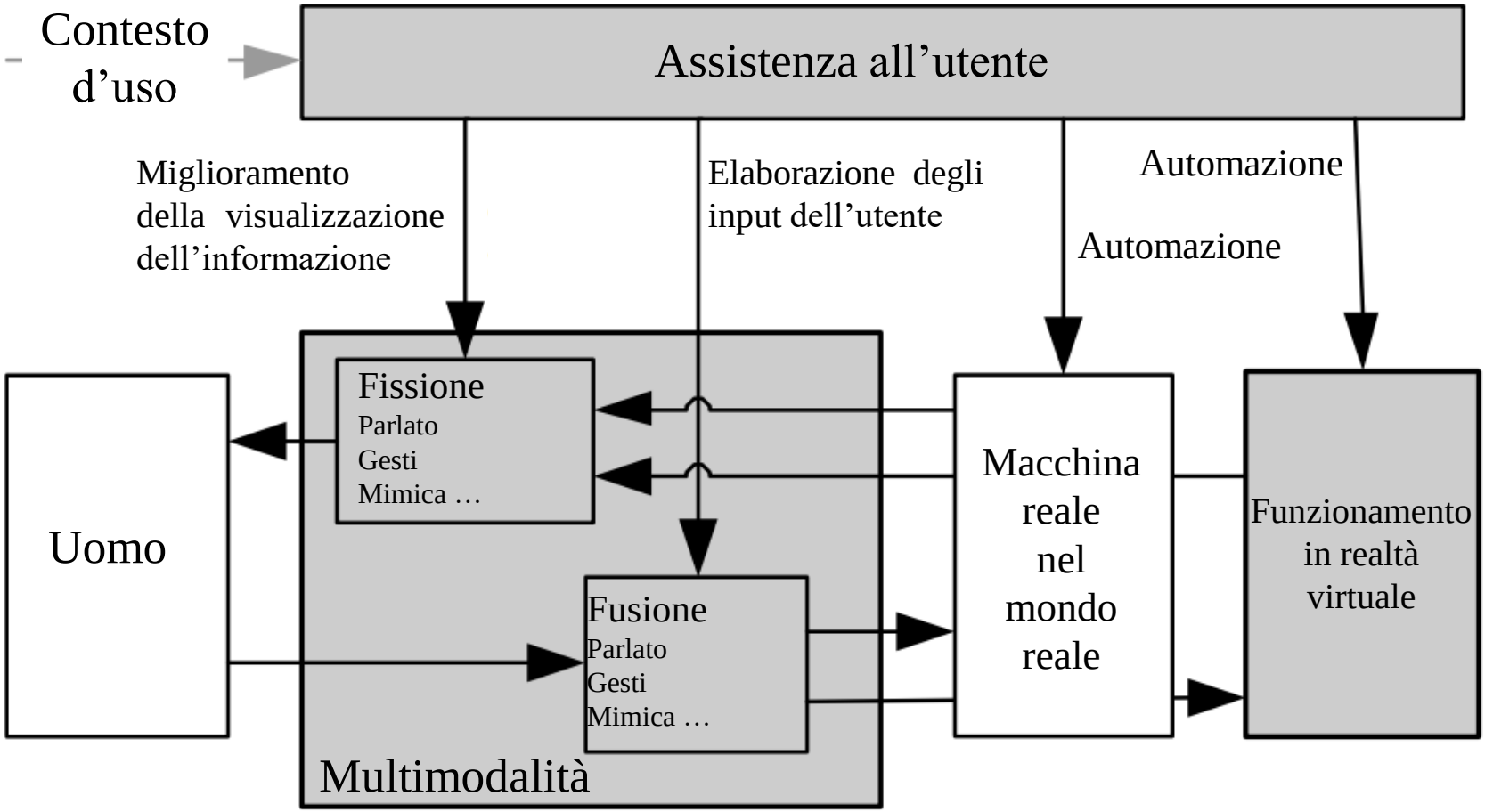
INTERAZIONE MULTIMODALE
Principi

Esempi di applicazioni multimodali sono le applicazioni *map-based* (esempio: informazioni per i turisti), i sistemi di realtà virtuale, i sistemi di identificazione e di riconoscimento delle persone, i sistemi medici e i sistemi che eseguono transazioni mediante il web.

Alcuni sistemi integrano due o più tecnologie di riconoscimento, come ad esempio il riconoscimento del parlato e il riconoscimento del movimento delle labbra.

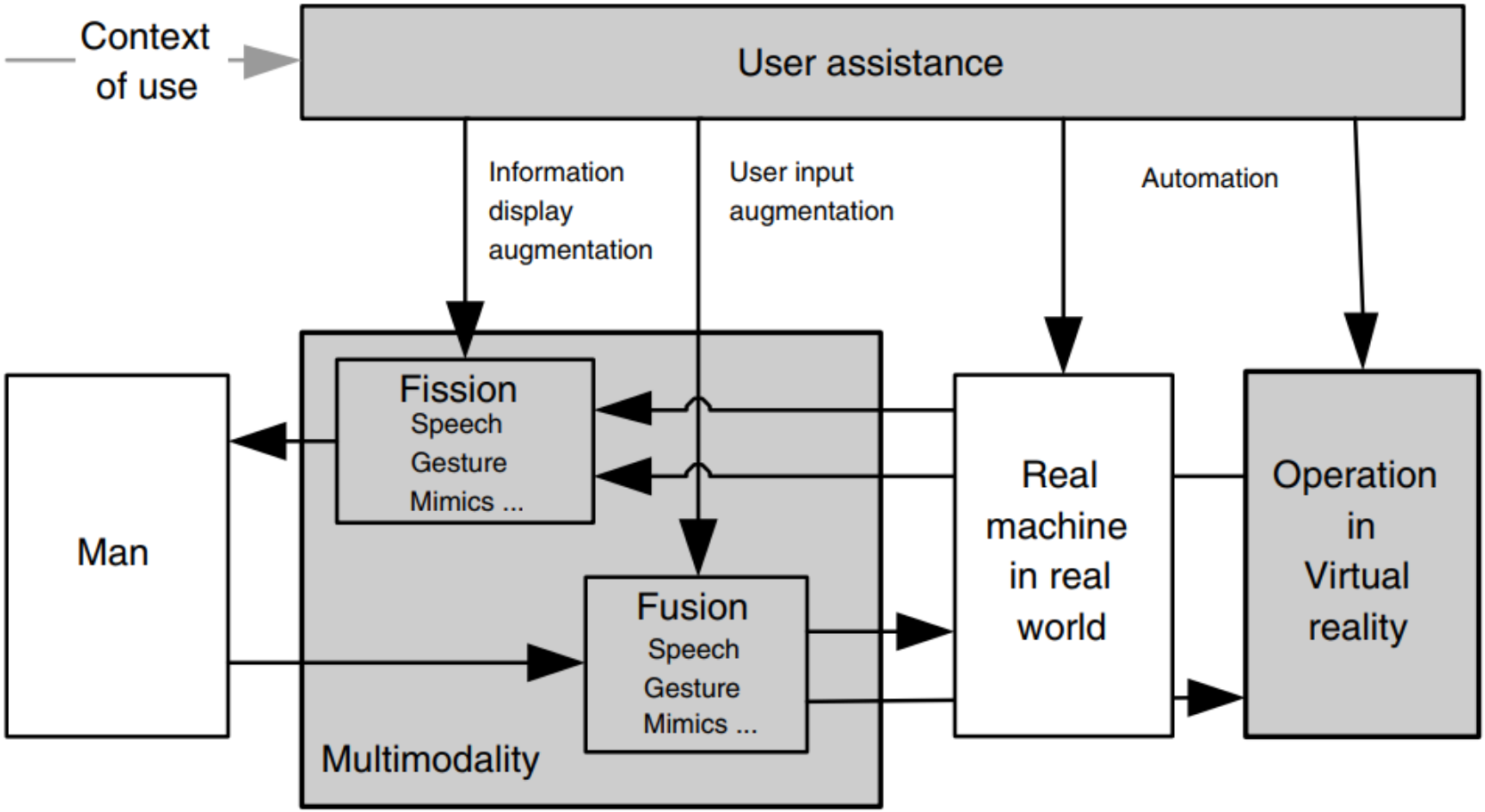
Un aspetto importante è rappresentato dall'integrazione e dalla sincronizzazione di questi differenti flussi di informazioni.

INTERAZIONE MULTIMODALE
Principi



Blocchi bianchi: schema base della MMI (Man-Machine Interaction)

INTERAZIONE MULTIMODALE
Principi



Blocchi bianchi: schema base della MMI (Man-Machine Interaction)

INTERAZIONE MULTIMODALE
Vantaggi

Le interfacce multimodali hanno l'obiettivo di favorire un interfacciamento più trasparente, flessibile, efficace, efficiente e robusto.

La flessibilità delle modalità di input rappresenta un importante aspetto della progettazione. Ciò include la scelta della modalità appropriata per diversi tipi di informazione, la combinazione di diverse modalità di input, o l'uso alternato di modalità diverse. Le modalità di input possono essere selezionate dall'utente o dal sistema in base al contesto e al compito. Per compiti e ambienti complessi, i sistemi multimodali permettono all'utente un'interazione più efficace.

Poiché ci possono essere numerose differenze riguardanti le capacità e le preferenze dei singoli utenti, è cruciale supportare la selezione e il controllo delle modalità di input per diversi gruppi di utenti. Per tale ragione, ci si aspetta che le interfacce multimodali permettano un apprendimento e un utilizzo agevole.

I continui cambiamenti nelle applicazioni permettono all'utente di alternare diverse modalità, come ad esempio in applicazioni relative ai veicoli.

INTERAZIONE MULTIMODALE
Vantaggi

Molti studi dimostrano che le interfacce multimodali soddisfano ampiamente gli utenti.

Il vantaggio principale è il guadagno nell'efficienza che deriva dall'abilità umana di processare modalità di input in parallelo.

La struttura del cervello umano è sviluppata in modo da raccogliere ogni tipo di informazione con input sensoriali specifici.

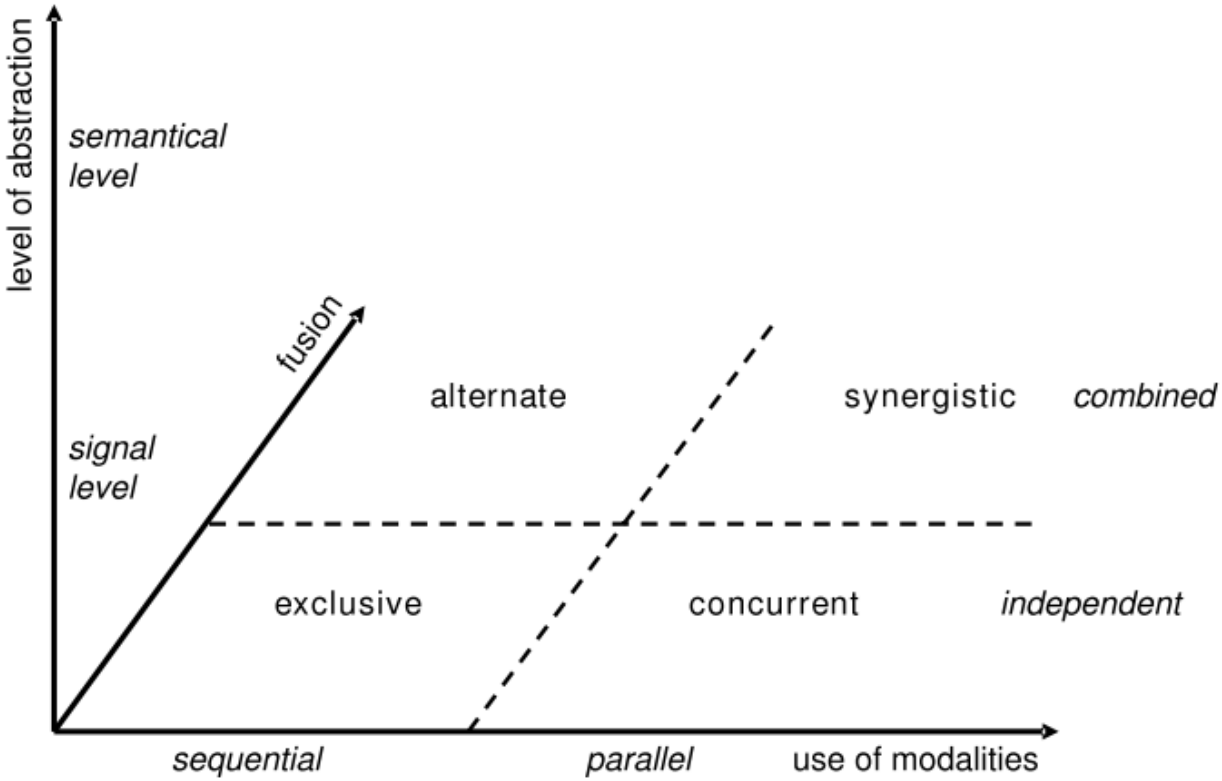
Il progetto di interfacce multimodali permette una migliore gestione degli errori in modo da evitare gli errori o da risolvere gli errori. Tali errori possono avere cause *user-centered* o *system-centered*.

Un'ambizione importante delle interfacce multimodali è quella di poter interpretare input continui associati ai canali visivo, uditivo e tattile per i sistemi di uso quotidiano al fine di supportare l'adattamento intelligente all'utente, al compito e all'ambiente di utilizzo.

INTERAZIONE MULTIMODALE
Tassonomia

Nell'interazione multimodale possono essere distinti tre gradi di libertà:

- il grado di astrazione
- la modalità temporale di applicazione
- la fusione delle differenti modalità

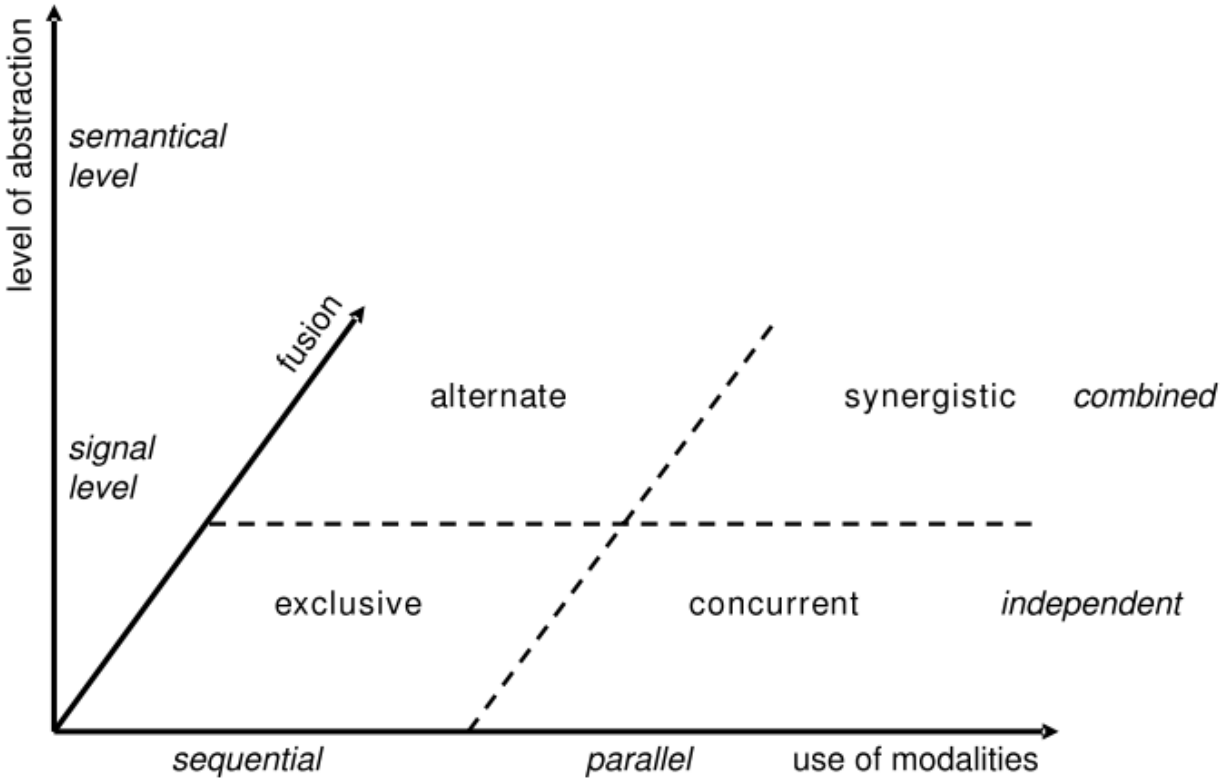


Spazio di classificazione per la classificazione di sistemi multimodali.

INTERAZIONE MULTIMODALE

Tassonomia

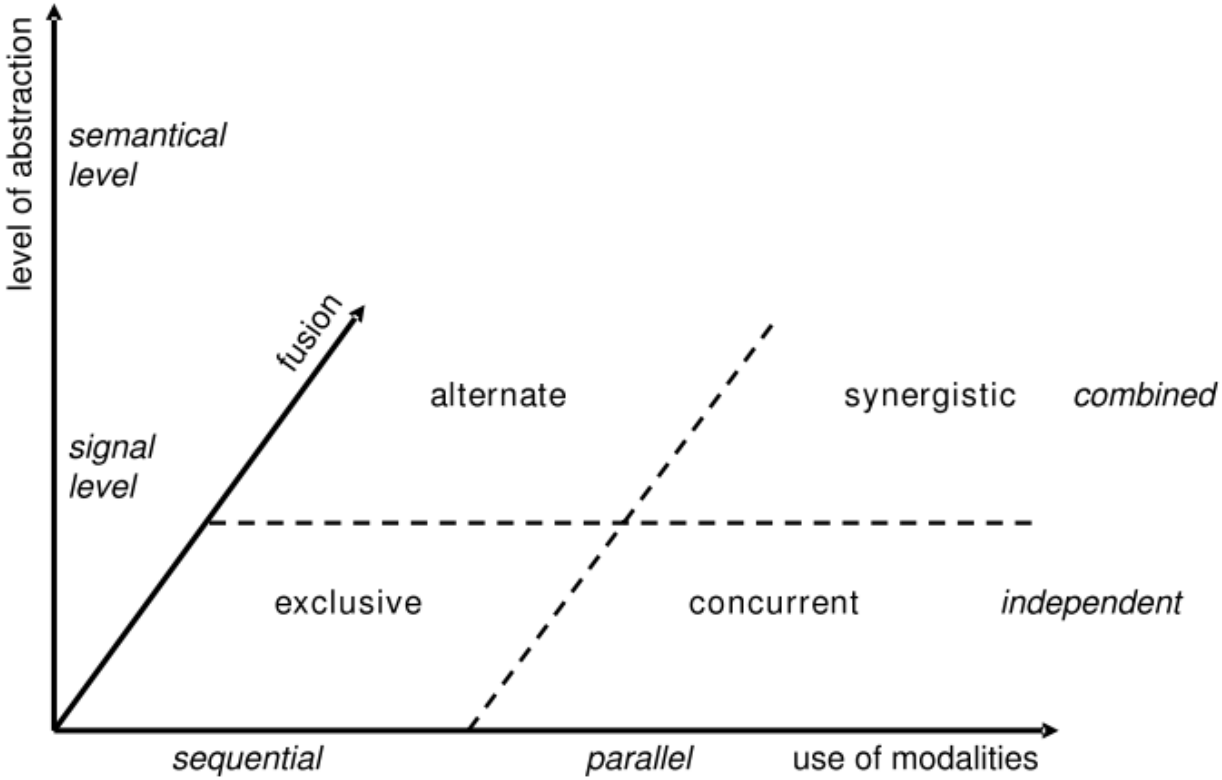
La figura mostra lo spazio di classificazione e le quattro categorie di base delle applicazioni multimodali. La definizione della categoria dipende dal valore dei parametri associati alla fusione dei dati (*combined/independent*) e all'utilizzo temporale (*sequential/parallel*).



Spazio di classificazione per la classificazione di sistemi multimodali.

INTERAZIONE MULTIMODALE
Tassonomia

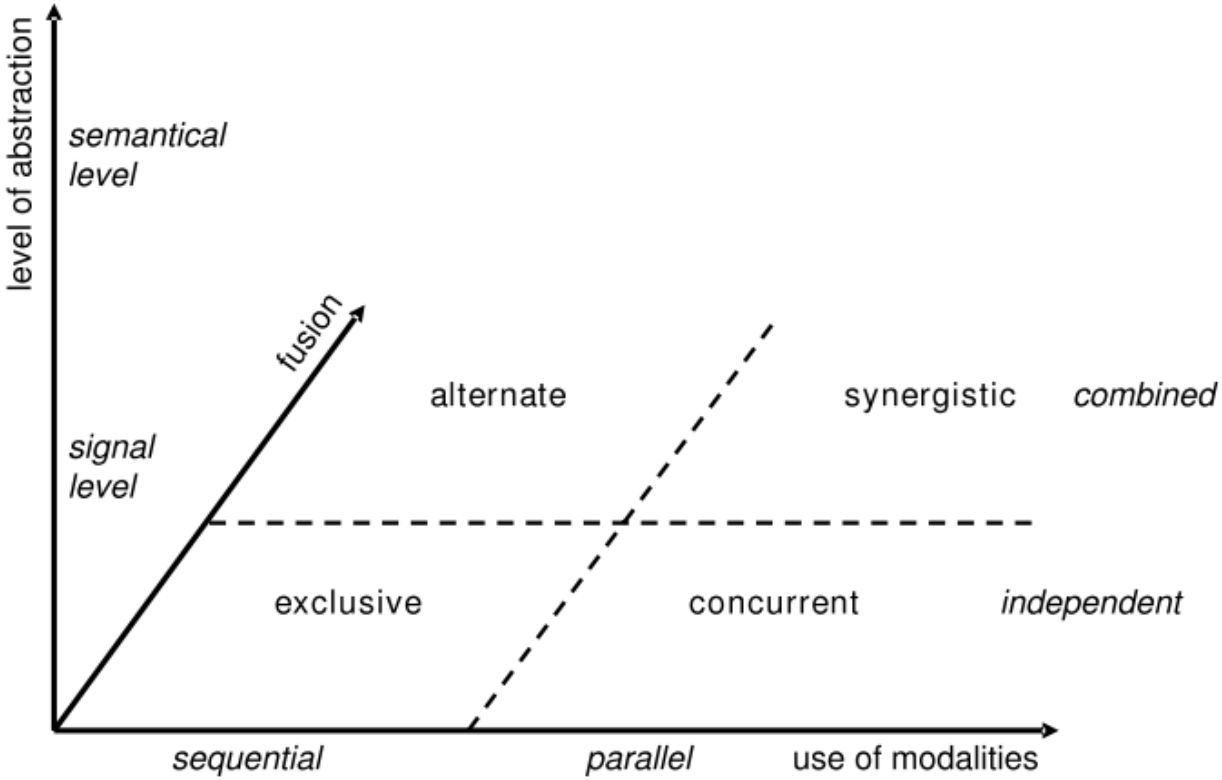
La categoria *exclusive* è la variante più semplice di un sistema multimodale. Tale sistema supporta due o più canali di interazione ma non c'è alcuna connessione temporale o associata al contenuto. Un'applicazione sequenziale delle modalità con coesione funzionale appartiene alla categoria *alternate*.



Spazio di classificazione per la classificazione di sistemi multimodali.

INTERAZIONE MULTIMODALE
Tassonomia

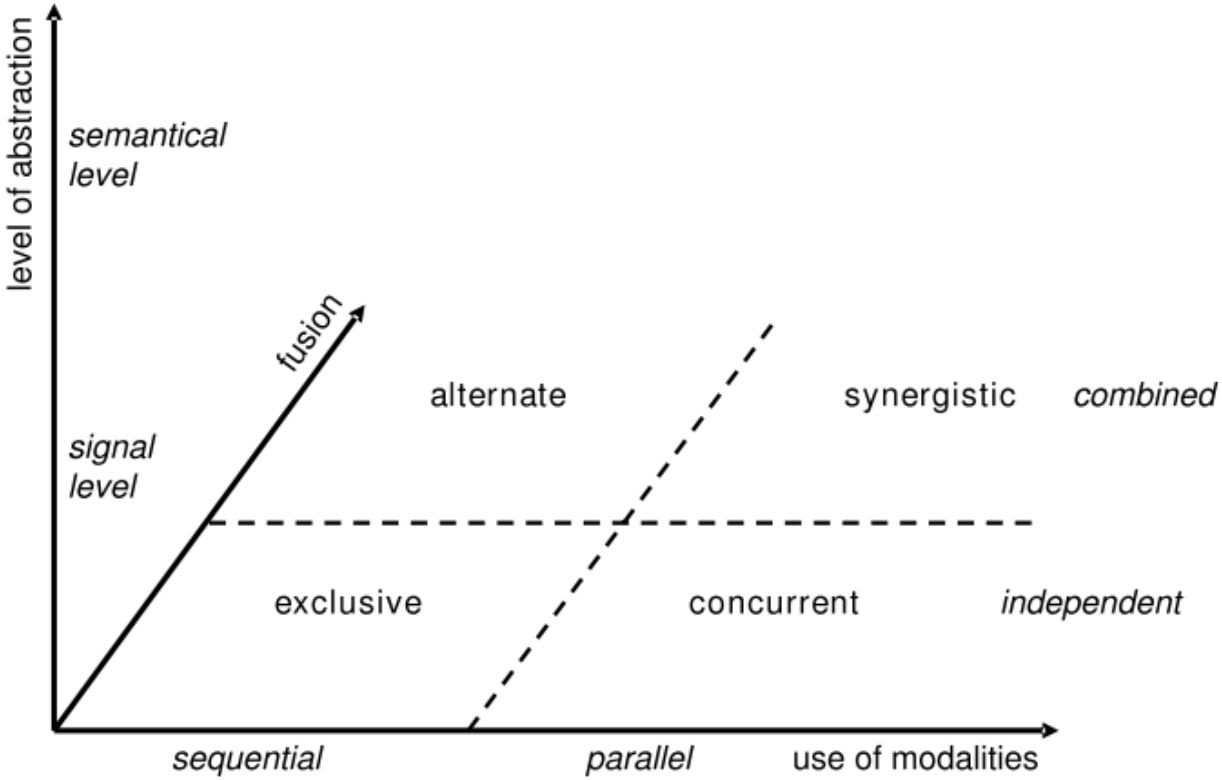
Oltre alle modalità sequenziali c'è anche la possibilità di operazioni in parallelo tra le differenti modalità. In caso di operazioni in parallelo, la distinzione viene effettuata sulla base della modalità di fusione dei canali di interazione in multimodalità simultanea (*concurrent*) e sinergica (*synergistic*).



Spazio di classificazione per la classificazione di sistemi multimodali.

INTERAZIONE MULTIMODALE
Tassonomia

Il terzo grado di libertà è il livello di astrazione (*level of abstraction*); esso si riferisce al livello tecnico (*technical level*), nel quale vengono elaborati i segnali. Tali segnali possono variare da semplici sequenze binarie (*signal level*) a termini semantici molto complessi (*semantical level*).



Spazio di classificazione per la classificazione di sistemi multimodali.

INTERAZIONE MULTIMODALE
Fusione multimodale

Come descritto in precedenza, nei sistemi multimodali il flusso delle informazioni può essere trasmesso dagli umani alle macchine mediante modalità differenti. Per poter utilizzare l'informazione trasmessa dai differenti sensi, è necessario integrare i canali di ingresso al fine di creare un comando appropriato che sia equivalente all'intenzione dell'utente.

I dati di input raccolti mediante le singole modalità sono generati tramite sistemi di riconoscimento associati a modalità singole.

Esistono tre approcci di base per la combinazione dei risultati di ogni sistema di riconoscimento in un unico flusso di informazioni:

- *Early (signal) fusion*
- *Late (semantic) fusion*
- *Soft decision fusion*

INTERAZIONE MULTIMODALE
Fusione multimodale

Early (signal) fusion
Tale approccio per la fusione dei dati si basa sulla combinazione dei dati grezzi specifici di ogni modalità di input. La classificazione dei dati può essere eseguita ad esempio mediante *Hidden Markov Models (HMMs)* o reti neurali.

Tale tipologia di fusione è adeguata per input sincronizzati. Tale approccio ha successo solo se i dati (che provengono da sorgenti differenti) sono della stessa tipologia e se esiste una forte correlazione tra le modalità. Un esempio è rappresentato dalla fusione di immagini generate con una camera normale e con una camera a infrarossi nei sistemi di visione adoperati durante la notte.

Inoltre tale tipologia di fusione viene applicata nei sistemi di riconoscimento del parlato supportati da tecnologie per la lettura labiale, nei quali la progressione dei fonemi e dei visemi può essere registrata in un *HMM*.

Uno svantaggio di tale approccio consiste nella grande quantità di dati necessaria per la fase di training dei *HMMs* utilizzati.

INTERAZIONE MULTIMODALE
Fusione multimodale

Late (semantic) fusion
I sistemi multimodali che utilizzano tale tipologia di fusione sono costituiti da sistemi di riconoscimento associati alle singole modalità e da un sistema di fusione dei dati. Tale approccio include, per ogni singola modalità, fasi di preelaborazione, di estrazione delle features e di decisione. I risultati dei livelli di decisione distinti vengono poi fusi in modo da ottenere un risultato globale.

Per ogni processo di classificazione, ogni livello di decisione calcola un risultato di probabilità e un risultato di confidenza per la scelta di una classe. Tali risultati di confidenza vengono poi fusi mediante opportune tecniche.

Il vantaggio di tale approccio è che i differenti sistemi di riconoscimento possono essere realizzati indipendentemente gli uni dagli altri. Quindi non è necessaria l'acquisizione di insiemi di dati multimodali. I differenti sistemi di riconoscimento vengono allenati (fase di training) con insiemi di dati monomodali.

INTERAZIONE MULTIMODALE
Fusione multimodale

Grazie alla facilità di integrazione di sistemi di riconoscimento addizionali, i sistemi che utilizzano tale approccio sono più scalabili rispetto a sistemi che utilizzano l'approccio descritto in precedenza (*early (signal) fusion*) sia per quanto riguarda il numero di modalità sia per la dimensione dell'insieme dei comandi.

Soft decision fusion

Tale approccio è un compromesso tra i due approcci precedentemente descritti.

INTERAZIONE MULTIMODALE
Fusione multimodale

In generale, i sistemi multimodali sono costituiti da vari moduli (ad esempio i sistemi di riconoscimento delle singole modalità, l'integrazione multimodale, l'interfaccia utente). Tipicamente, tali componenti software vengono sviluppati e implementati indipendentemente gli uni dagli altri. Pertanto, per quanto riguarda l'architettura software di un sistema multimodale, occorre tenere conto di differenti specifiche. Un'infrastruttura che viene comunemente adottata si basa su architetture multi-agente, dove ogni processo software rappresenta un agente. In tali architetture, i molti moduli necessari per supportare il sistema multimodale possono essere scritti in differenti linguaggi di programmazione e possono essere eseguiti su differenti macchine.

Un esempio di architettura è un'architettura caratterizzata da tre livelli di elaborazione: *input level*, *integration level* e *output level*. L'*input level* contiene ogni tipo di interfaccia in grado di riconoscere gli input dell'utente (esempi: mouse, tasti, sistema di riconoscimento del parlato). Nell'*integration level*, le uscite dei sistemi di riconoscimento e le informazioni aggiuntive sul contesto (esempi: informazioni sull'ambiente di applicazione e sullo stato dell'utente) vengono combinate con un approccio *late (semantic) fusion*. L'*output level* si basa su dispositivi in grado di inviare un feedback adeguato del sistema multimodale.

Riferimenti Bibliografici

[1] Kraiss, K. -F. (2006). Advanced Man-Machine Interaction: Fundamentals and Implementation. Springer-Verlag Berlin Heidelberg. ISBN-10: 3-540-30618-8