UNIVERSITÀ DEGLI STUDI DI FIRENZE
Scuola di Scienze Matematiche Fisiche e Naturali
Corso di Laurea in Informatica

Reti neurali per il riconoscimento dei gesti

Andrea Venuta, matr. 4738148
Giovanni Pinto, matr. 5577276

A.A. 2013/2014

Indice

1. Introduzione e motivazioni

Negli ultimi anni lo sviluppo e la ricerca nell'ambito della User Experience hanno portato ad un rinnovamento radicale dei paradigmi di interazione. Un esempio eccellente di questo fenomeno è rappresentato dall'introduzione degli smartphone e dei tablet touch-screen, la cui semplicità d'uso ed accessibilità hanno notevolmente abbassato le barriere d'entrata al mondo dell'informatica. Collateralmente, gli ambienti ingegneristici di aziende piccole e grandi si muovono sulla sperimentazione di paradigmi inusuali e controintuitivi: Oculus VR produce un head-mounted display che amplia il campo visivo permettendo applicazioni di realtà virtuale, Google conduce esperimenti su dispositivi portatili ad alte prestazioni per il mapping ambientale tridimensionale (Project Tango), le comunità di sviluppatori studiano algoritmi per la rilevazione del movimento con strumenti quali Microsoft Kinect.

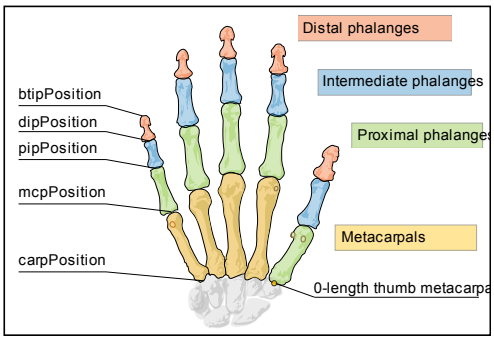
Tra questi progetti, spicca quello della statunitense Leap Motion: si tratta di un dispositivo di rilevazione del movimento delle mani originariamente concepito per semplificare la modellazione 3D, rendendola intuitiva come la modellazione della creta nella realtà. Il Leap rileva, in un'area approssimativamente semisferica, ostacoli entro una distanza di circa un metro dalla sua superficie. Due telecamere monocromatiche raccolgono la luce infrarossa riflessa da tre LED ed algoritmi di computer vision, non resi pubblici dalla compagnia produttrice, rilevano la posizione della mano o dell'ostacolo e ne offrono una rappresentazione in uno spazio euclideo tridimensionale standardizzato.

I dati forniti dal Leap Motion sono adatti in contesti in cui la posizione spaziale istante per istante è un elemento importante dell'interazione, ad es. la manipolazione di oggetti nello spazio euclideo; un'applicazione che però non rientra tra le feature dello strumento è la semantica delle

gestualità. Tramite gli argomenti trattati nella presente relazione, si vuole proporre un sistema che permetta di classificare un generico alfabeto di simboli, le cui poi applicazioni dipenderanno dal contesto in cui il sistema viene impiegato. L'ambito che si è voluto studiare nella presente relazione riguarda il campionamento e la classificazione dei simboli della lingua dei segni americana (ASL). Applicazioni su questo dominio potrebbero riguardare la trascrizione di conversazioni tra sordomuti, oppure la videoscrittura in contesti in cui l'utilizzo di una tastiera potrebbe risultare poco pratico (es. su dispositivi molto piccoli od in ambienti angusti che non permettono di utilizzare con efficacia una tastiera, come ad esempio una miniera o sott'acqua). Vedremo che le reti neurali, unitamente al Leap Motion, sono uno strumento utile ad affrontare questa tipologia di problema.

2. Analisi del problema e delle tecnologie impiegate

2.1. Acquisizione e struttura delle informazioni



Una recente release beta del firmware del Leap ha reso possibile codificare la mano come una struttura gerarchica affine a quella dell'ossatura naturale: il palmo è considerato la radice dell'albero, i cui figli di primo livello sono cinque metacarpi (al pollice, che non ha un metacarpo nella realtà, si codifica per coerenza strutturale un metacarpo di lunghezza nulla). Ciascuna falange metacarpale avrà uno ed un solo nodo figlio, la falange prossimale, che a sua volta si estenderà in un nodo rappresentante la falange intermedia, che a sua volta terminerà con una foglia, la falange distale (vedi foto). Questa struttura è affine, tanto nel concetto quanto nella codifica, a quella utilizzata nell'ambito dell'animazione grafica tridimensionale. Ogni falange, infatti, espone varie informazioni spaziali che permettono di posizionarla efficacemente nello spazio euclideo del Leap: un vettore che ne indica la traslazione rispetto al palmo, un vettore normale che ne indica la direzione, uno scalare che ne indica la lunghezza, una base ortonormale che ne indica la rotazione ed altre ancora; il palmo, a sua volta, espone informazioni simili, tra cui la velocità, la direzione, il vettore normale alla sua superficie e così via.

2.2. Scelta del sottoinsieme utile di informazioni

Il progetto si prefigge di addestrare una rete neurale al riconoscimento dei gesti, intesi come posizione assunta dalla articolazioni dalla mano in un dato istante. Non tutte le informazioni trasmesse dal Leap sono utili a questo scopo, ed anzi potrebbero rivelarsi controproducenti. Se, ad

esempio, considerassimo la posizione spaziale assoluta di ogni articolazione della mano, dovremmo istruire la rete neurale affinché associ ad un dato gesto un larghissimo numero di posizioni della mano nello spazio, ed un larghissimo numero di mani di dimensioni diverse in ognuna delle suddette posizioni, perché la posizione di ogni articolazione nello spazio potrebbe variare significativamente tra due mani differenti. Nel modo in cui un essere umano descrive un gesto, inoltre, la posizione spaziale è un parametro irrilevante alla definizione concettuale di "gesto": se il saggio indica la luna, lo stolto distingue il gesto fatto quando il saggio gli sta accanto dal gesto fatto quando il saggio è dal lato opposto della strada. È d'uopo quindi ricercare un sottoinsieme dei parametri forniti dal Leap che descriva accuratamente ed in maniera più indipendente ed astratta possibile il concetto di gesto che vogliamo riconoscere.

Un parametro valido a rappresentare il gesto è il modo in cui le articolazioni sono ruotate nello spazio. Ogni rotazione è rappresentabile attraverso un vettore normale che indica semplicemente la direzione della falange. Geometricamente è calcolato traslando l'articolazione al centro dello spazio e normalizzando il vettore che parte da esso e che arriva alla posizione dell'articolazione successiva, oppure normalizzando il vettore differenza tra la posizione dell'articolazione successiva e la posizione dell'articolazione considerata. Per quanto riguarda il palmo, invece, registreremo la normale alla sua superficie. Nella realtà ovviamente il palmo non ha una superficie puramente piana, ma il Leap ne fornisce un'approssimazione semanticamente sensata. La normale di superficie del palmo ha, nell'àmbito del progetto, la stessa funzione concettuale delle direzioni delle falangi, ovvero quella di rappresentare sinteticamente una rotazione.

Si decide dunque di rappresentare il layout delle feature sull'input della rete neurale come la concatenazione del vettore normale di superficie del palmo e di tutti i vettori normali direzione di ogni articolazione delle dita,

secondo il seguente schema:

$$
\begin{array}{lll}
[palmNormal_x, & palmNormal_y, & palmNormal_z, \\
thumb_x^0, & thumb_y^0, & thumb_z^0, \\
thumb_x^1, & thumb_y^1, & thumb_z^1, \\
thumb_x^2, & thumb_y^2, & thumb_z^2, \\
 & \dots & \\
little_x^3, & little_y^3, & little_z^3]
\end{array}
$$

laddove il pedice indica l'asse a cui la componente del vettore fa riferimento, e l'indice numerico rappresenta la falange, secondo la seguente tabella:

0	*metacarpale*
1	*prossimale*
2	*intermedia*
3	*distale*

Avremo dunque un totale di 63 feature, di valore compreso tra 0 e 1, dal momento che ogni vettore è di lunghezza unitaria e quindi ogni sua componente ha valori compresi tra 0 e 1.

2.3. Design ed algoritmi per il software di riconoscimento

È stata realizzata un'interfaccia web based per l'acquisizione ed il riconoscimento delle gesture. Si ha la possibilità di aggiungere un numero arbitrario di gesture durante tutto il ciclo di vita dell'applicazione. Ogni gesture è rappresentata in una riga di una tabella, la quale è munita di un pulsante di registrazione. Quando si desidera campionare dati per una gesture dal Leap, si deve premere il bottone corrispondente; il software inizierà dunque ad ascoltare gli eventi provenienti dal Leap. Alla ricezione di un evento "frame" (ovvero una "fotografia" dello stato corrente delle rilevazioni del Leap, contenente tutte le informazioni strutturate relative alle mani), i dati utili vengono estratti e sintetizzati nella forma di un vettore unidimensionale, come sopra discusso. Dopo

ogni registrazione, la rete neurale viene ricreata ed addestrata con tutti i dati di tutte le gesture registrati fino a quel momento.

In ogni istante in cui non avviene il campionamento attivo dei frame per la registrazione, la ricezione dell'evento "frame" innescherà la procedura di correlazione della rete neurale con i dati sintetizzati in quel frame, l'ultimo cronologicamente ricevuto. A fronte del risultato riportato dalla rete neurale, i valori di correlazione per ogni gesture saranno riportati sull'interfaccia.

Usando il paradigma prototype-based/object oriented, si è definita la classe `Gesture` il cui scopo è sostanzialmente quello di mantenere la lista dei campioni per una data gesture, e di controllare la vista ad essa relativa nell'interfaccia utente, gestendo gli eventi di registrazione ed aggiornando l'interfaccia in seguito al cambiamento nel tempo del valore di correlazione della gesture con i dati ricevuti in ogni frame. Si è definita quindi la classe `GesturePool` che si occupa di tenere traccia dell'insieme di gesture create, di addestrare la rete ad ogni cambiamento dell'insieme dei dati e di invocare la correlazione della rete ogni volta che è disponibile un nuovo frame.

3. Reti neurali per il riconoscimento di gesture

In questa sezione andremo a descrivere il lavoro effettuato sulla progettazione e configurazione di una rete neurale adatta al riconoscimento di gesture. Dopo aver definito che tipologia di gesture utilizzeremo e come viene codificata una gesture, stabiliremo un alfabeto di gesti e vedremo come questi dati vengono usati per addestrare la rete. Infine, sarà mostrato il risultati di alcuni test che evidenziano la performance della rete.

3.1 Cosa è un gesto

I gesti fanno parte del linguaggio del corpo, l'aspetto più studiato e conosciuto della comunicazione non verbale. Il gesto spontaneo va distinto da quelli usati nelle lingue dei segni, che sono codificati. Parlando di gesti effettuati usando le mani, questi possono essere definiti come qualsiasi configurazione o movimento della mano che abbia un minimo di significato semantico. Quindi gesti della mano possono prevedere movimenti della stessa (ad esempio il gesto che indica il "no" o il rifiuto, effettuato tenendo il pugno chiuso con l'indice alzato e muovendolo da destra a sinistra ripetutamente), ma anche configurazioni statiche (ad esempio i gesti utilizzati per identificare i numeri). La differenza fra un gesto statico ed uno dinamico è notevole, e gli approcci possibili per identificare entrambi con una rete neurale sono molteplici. I gesti statici possono essere considerati gesti dinamici che non hanno movimento. Questa loro caratteristica li rende estremamente più semplici da identificare e, per i nostri test, abbiamo deciso di utilizzare esclusivamente questi ultimi a causa delle prblematiche esposte in seguito.

3.2 Problematiche relative al riconoscimento di gesti dinamici

Nei paragrafi precedenti è stato descritto come vengono codificati i gesti dal nostro sistema. Bisogna notare che ogni gesto può avere una durata nel tempo diversa e di conseguenza essere codificato con vettori di lunghezza diversa. Chiaramente la dimensione dell'input, per poter essere usato correttamente da una rete neurale, deve essere costante per ogni campione. Per ovviare a questo problema sono stati presi in considerazione diversi approcci:

1. Troncare le codifiche dei gesti più lunghi alla dimensione del gesto più corto;
2. Definire una durata standard per ogni gesto e normalizzare le codifiche a questa dimensione;
3. Usare la dimensione della codifica del gesto più lungo come lunghezza dell'input.

Il primo approccio è il più semplice da realizzare e permette di ottenere comunque risultati soddisfacenti. Se i gesti utilizzati hanno una durata limitata nel tempo, ad esempio meno di un secondo, e abbastanza omogenea fra loro, eliminare alcuni frame significa tagliare qualche centesimo di secondo alla durata complessiva del gesto, e probabilmente questo non influirebbe in maniera tale da impedirne il riconoscimento. Per gesti di durata maggiore invece il problema diventa rilevante: due gesti che hanno "prefissi" identici, cioè che iniziano con lo stesso movimento, rischiano di essere identificati come lo stesso gesto, in quanto la parte che li avrebbe discriminati è stata tagliata.

Il secondo approccio prevede di modificare la codifica di un gesto, normalizzandola ad una dimensione definita. Supponendo che la dimensione standard sia pari a n, un gesto di dimensione inferiore o superiore deve essere normalizzato a questa dimensione. Questo presupone che si debba interpolare il movimento ad intervalli di tempo predefiniti ed uniformi, da $t = 0$ fino a $t = n$. La semplice interpolazione sferica lineare sarebbe una buona approssimazione supponendo una

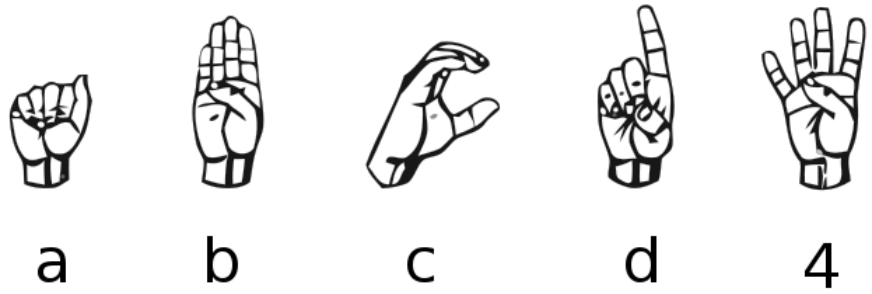
frequenza di rilevamento sufficientemente alta. Per quantificare la frequenza necessaria ad ottenere errori di interpolazione accettabili sarebbe necessario effettuare studi approfonditi sul movimento delle falangi, che non sono di nostra competenza.

Il terzo approccio può sembrare simile al primo ma ha un problema differente: in questo caso il gesto reale ha una durata potenzialmente inferiore al gesto registrato e analizzato dalla rete neurale. Questo implica che alla rete vengano passati dati in coda "sporchi", che potrebbero essere rumore o l'inizio di una successiva gesture. Dato che la rete analizza una "finestra" di frame inseriti in un buffer, e che ad ogni nuovo frame generato questa finestra si sposta, la rete si potrebbe trovare ad analizzare due gesti diversi contemporaneamente, degradando le prestazioni.

Una ulteriore problematica riguarda il riconoscimento dell'inizio e della fine del gesto. Un unico movimento può comprendere una moltitudine di gesti in sequenza (immaginiamo di poter controllare un computer tramite comandi gestuali). Potremmo definire l'inizio e la fine di ogni gesto come l'inizio di un movimento della mano e la fine dello stesso, rallentando l'input ma al contempo rendendo possibile ed efficace distinguere i gesti di una sequenza. In ogni caso bisognerebbe comunque analizzare una finestra di gesti e questo potrebbe rendere il riconoscimento non preciso e puntuale. Immaginiamo di muovere la mano facendo due gesti in sequenza e di analizzare il movimento in un intervallo di tempo di dimensione fissata (avendo quindi un buffer in cui si rimuove il primo frame e si aggiunge un nuovo frame ad ogni istante di tempo). Quando il tempo passa (e la mano si muove) il buffer conterrà inizialmente il primo gesto, successivamente la parte finale del primo gesto e la parte iniziale del secondo, successivamente solo il secondo e così via. Chiaramente questo rende il riconoscimento più complicato e meno conciso: ci si aspetta infatti una correlazione con il primo gesto dapprima molto alta (e

molto bassa per il secondo), poi decrescente (e quindi paragonabile a quella del secondo, che cresce), poi molto bassa (mentre si alza la correlazione con il secondo gesto).

3.3 Alfabeto dei gesti



Alla luce delle precedenti affermazioni, abbiamo deciso di utilizzare un alfabeto che sia esclusivamente composto di gesti statici. Un alfabeto del genere non va considerato troppo limitato, basti pensare che l'alfabeto manuale è costituito proprio da questo tipo di gesti.

Il sistema che abbiamo progettato è in grado di identificare la posizione della mano in qualsiasi configurazione, per un totale virtualmente infinito di gesti possibili. Riconfigurare la rete per apprendere un nuovo gesto è una procedura estremamente rapida, basta effettuare il gesto e premere il pulsante di registrazione. Per poter effettuare test sulle prestazioni della rete neurale, però, è utile definire un set di gesti che compongono un "alfabeto". I gesti di questo alfabeto verranno poi utilizzati effettivamente sia in fase di training che in fase di test. Questo approccio garantisce una standardizzazione utile ad ottenere risultati più accurati e di più facile interpretazione. Sarebbe impossibile infatti confrontare i risultati di diversi test in cui si sono utilizzati gesti totalmente differenti.

Abbiamo quindi deciso di utilizzare un alfabeto composto da 5 gesti, precisamente le prime 4 lettere dell'alfabeto ASL (A, B, C, D) ed il numero "4" secondo la tradizione italiana. L'inclusione del numero 4 è dovuta al

fatto che quest'ultimo è codificato da un gesto molto simile a quella della lettera "B", e può essere utile per identificare i limiti della rete.

3.4 Progettazione della rete

Sia n il numero di gesti previsti in un determinato set di gesture su cui è stata addestrata la rete. L'outupt della rete è un valore $r_i \in [0,1], i = 0, \dots, n-1$ che rappresenta la correlazione del gesto analizzato con ognuno dei gesti su cui è stata addestrata la rete. Per ricavare il gesto riconosciuto dalla rete come come candidato è necessario quindi vedere quale sia il grado di correlazione maggiore.

Nell'utilizzo pratico dello strumento si effettuano anche gesti che non sono di nessuna reale utilità (*gesti semanticamente nulli* o *gesti nulli*), ad esempio spostando la mano mentre ci si appresta ad effettuare un gesto semanticamente valido, ma che vengono comunque analizzati dalla rete. Chiaramente ci si aspettano valori di correlazione bassi per questo tipo di gesture rispetto ai gesti validi. Per distinguere i gesti nulli dai gesti validi è utile stabilire un valore s di soglia di correlazione, sotto la quale il sistema non afferma di aver riconosciuto un candidato. Questo valore deve essere il più alto possibile per non classificare gesti nulli, ma non abbastanza alto da annullare la classificazione corretta di gesti validi.

Questo implica che l'output della rete, sebbene sia comunque un valore $r_i \in [0,1]$, venga preso in considerazione solo se esiste un r_i per cui il valore di correlazione supera una soglia s. Altrimenti il gesto viene considerato nullo. Intervallando gesti validi a gesti nulli diventa possibile riconoscere sequenze di gesti, permettendo ad esempio di interpretare la lingua dei segni in tempo reale.

Nei paragrafi precedenti è stato descritto come viene generata la codifica di un gesto. Il training set della rete è proprio un insieme di queste codifiche. Per ogni componente del vettore normale che identifica la

direzione di ogni elemento della mano (quindi tre valori x, y, z) si avrà un neurone che avrà in input quella feature. In tutto si hanno 63 neuroni di input.
Intuitivamente possiamo vedere il training set della rete come un insieme di "fotografie" della mano in diverse posizioni in base al gesto effettuato. Ogni gesto avrà un training set definito da un numero di sample variabile a seconda delle necessità. In fase di test l'input è analogo: alla rete viene passata la codifica della configurazione della mano in tempo reale e la rete tenta di associare la posizione della mano con uno dei gesti dell'alfabeto di training.

4. Analisi delle prestazioni

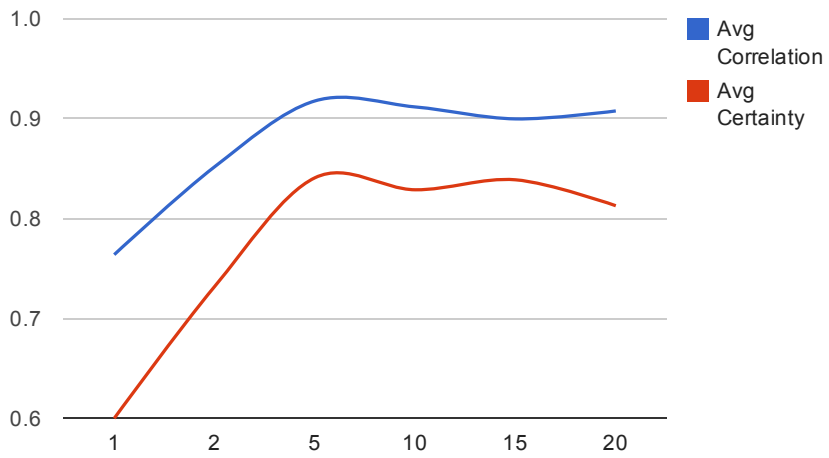
Chiaramente è necessario effettuare dei test per valutare le prestazioni della rete nel riconoscimento dei gesti. In tutti i test si è utilizzato l'alfabeto definito precedentemente. I gesti vengono rivolti verso il dispositivo (per intendersi, il gesto "numero 4" viene effettuato con il palmo volto verso il basso e non in avanti). Ogni gesto viene effettuato solo dopo aver rimosso la mano dal campo d'azione del dispositivo ed aver effettuato alcuni gesti neutri (ad esempio muovendo le dita vorticosamente o chiudendo e aprendo il pugno alcune volte); la mano viene quindi inserita nel campo d'azione in quella che definiamo *posizione neutra*, ovvero dischiusa, in maniera simile al gesto effettuato per codificare il numero 5; infine si effettua il gesto voluto.
A questo punto è importante che non sia lo stesso soggetto che effettua il gesto a registrare il risultato; ancor più importante è che il soggetto che effettua il gesto non sia al corrente dell'output della rete in quel momento. Questo serve chiaramente ad evitare di condizionare il test, in quanto il soggetto potrebbe registrare l'output solo nel momento più propizio, modificando leggermente la configurazione della mano per ottenere risultati ottimali. I soggetto deve quindi comunicare lo stato di "pronto" quando ha completato il gesto ed il risultato deve essere registrato senza che lui lo conosca.

A volte capita che il Leap Motion, essendo un dispositivo altamente sperimentale, non riconosca perfettamente la mano e costruisca una rappresentazione sbagliata di questa (ad esempio mettendo il pollice sul lato sbagliato del palmo o piegando le dita nel verso opposto a quello naturale, ottenendo risultati grotteschi e a tratti esilaranti). In questo caso il test viene semplicemente invalidato e quindi ripetuto.
Ogni test, salvo diversa indicazione, è stato eseguito seguendo le modalità descritte.

Come misura dell'efficacia del riconoscimento di un gesto abbiamo scelto quella che chiameremo *certezza*. Questo valore è uguale alla differenza tra il gesto che ha correlazione maggiore (quindi il candidato migliore secondo la rete) ed il gesto che ha correlazione immediatamente inferiore. Sia g_1 il candidato con correlazione massima c_1, e sia g_2 il candidato con correlazione c_2 tali che $c_1 < c_2 < c_3 < \ldots$. La *certezza* i è data da $c_1 - c_2$.
Questa variabile assumerà valori compresi tra 0 (non si può stabilire quale sia il gesto effettuato) e 1 (si ha la massima correlazione tra il gesto effettuato ed uno dei gesti con cui è stata addestrata la rete e correlazione 0 per gli altri).

4.1 Numero di sample di training

In questo test si studia la performance della rete al variare del numero di test sample utilizzati per addestrarla. L'ipotesi è che un numero maggiore di test sample garantisca risultati migliori (quindi livelli di correlazione più elevati ed incertezza minore).



Gesture	Correlation	Certainty	Gesture	Correlation	Certainty
A	0.579	0.420	**A**	0.946	0.899
B	0.742	0.477	**B**	0.746	0.516
C	0.923	0.846	**C**	0.878	0.818
D	0.737	0.628	**D**	0.914	0.871
Q	0.837	0.627	**Q**	0.777	0.558
Average	0.764	0.600	**Average**	0.852	0.732
Dev	0.115	0.148	**Dev**	0.078	0.162
Sample: 1			**Sample: 2**		

Gesture	Correlation	Certainty	Gesture	Correlation	Certainty
A	0.954	0.873	**A**	0.962	0.919
B	0.843	0.704	**B**	0.779	0.610
C	0.927	0.885	**C**	0.986	0.959
D	0.943	0.904	**D**	0.956	0.907
Q	0.926	0.840	**Q**	0.876	0.748
Average	0.918	0.841	**Average**	0.912	0.829
Dev	0.039	0.071	**Dev**	0.076	0.131
Sample: 5			**Sample: 10**		

Gesture	Correlation	Certainty	Gesture	Correlation	Certainty
A	0.960	0.916	**A**	0.959	0.913
B	0.794	0.642	**B**	0.717	0.456
C	0.823	0.796	**C**	0.976	0.945
D	0.981	0.957	**D**	0.946	0.903
Q	0.942	0.883	**Q**	0.941	0.847
Average	0.900	0.839	**Average**	0.908	0.813
Dev	0.076	0.112	**Dev**	0.096	0.181
Sample: 15			**Sample: 20**		

Si può notare come aumentando il numero di sample aumenta la precisione con cui la rete classifica il gesto, fino ad arrivare a 10 sample. Con un numero maggiore non sembrano esserci miglioramenti significativi. Non ci sono stati errori di classificazione ma i gesti B e Q sono quelli con sicurezza della classificazione più bassa, data la loro somiglianza.

Il minimo valore di correlazione, con 10 sample, è pari a 0.779 per il gesto B. Come valore di soglia per identificare i gesti validi da quelli nulli si dovrebbe utilizzare un valore inferiore a questo.

4.2 Riconoscimento con mani di soggetti differenti

Nel seguente test un soggetto S1 addestrerà la rete, successivamente un soggetto S2 eseguirà la routine di test per valutare la performance della rete. Indubbiamente soggetti differenti eseguono lo stesso gesto in maniera leggermente differente. Sapere il grado di generalizzazione della rete in questo caso può esserre utile per valutarne l'affidabilità.

Gesture	Correlation	Certainty
A	0.902	0.827
B	0.156	0.077
C	0.964	0.901
D	0.797	0.759
Q	0.874	0.744
Average	0.739	0.662
Dev	0.296	0.298
Training: S1; Test: S2		

Gesture	Correlation	Certainty
A	0.962	0.919
B	0.779	0.610
C	0.986	0.959
D	0.956	0.907
Q	0.876	0.748
Average	0.912	0.829
Dev	0.076	0.131
Training: S2; Test: S2		

Nel seguente test il training viene effettuato utilizzando 10 sample per ciascuno dei due soggetti S1 ed S2.

Gesture	Correlation	Certainty
A	0.969	0.935
Q	0.579	0.235
C	0.706	0.668
D	0.977	0.939
Q	0.961	0.910
Average	0.838	0.737
Dev	0.165	0.271
Training: S1 + S2; Test: S2		

In questo caso la rete ha sbagliato il riconoscimento del gesto B, classificandolo come il gesto Q, e restituendo un grado di correlazione particolarmente basso. In media però la rete sembra offrire risultati migliori rispetto al training effettuato con un solo soggetto. Chiaramente i risultati migliori si ottengono quando il soggetto che addestra la rete è lo stesso che ha effettuato il training.

Conclusioni

Abbiamo visto come l'approccio basato su reti neurali permetta di ottenere risultati soddisfacenti anche senza particolari elaborazioni dei dati forniti dal Leap Motion. Chiaramente si potrebbero ottenere risultati migliori introducendo nuove feature, ottenute combinando quelle già utilizzate, ma questa operazione va al di là dello scopo della relazione. Il limite più forte è stato rappresentato dalla classificazione di gesti dinamici. Riteniamo che utilizzare le reti neurali per affrontare questo tipo di problema non sia l'approccio migliore data la necessità di pre-processare, e quindi condizionare, i dati per renderli compatibili all'elaborazione da parte di una rete.

Abbiamo notato come la rete ottenga prestazioni peggiori nel classificare i gesti se questi sono effettuati da un soggetto diverso da colui che la ha addestrata. Sarebbe quindi utile studiare questa problematica in futuro, per individuare tecniche in grado di addestrare reti che generalizzino la classificazione in modo universale, e che classifichino gesti in modo ottimale indipendentemente dal soggetto che li effettua.