WeAreClouds@Lucca

progetto finanziato da
Fondazione Cassa di Risparmio di Lucca

e dal
Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

e in collaborazione con il
Comune di Lucca

D1.2 STATO DELL'ARTE SCIENTIFICO

Autori:

Fabio Valerio Massoli

Andrea Carboni

Davide Moroni

Fabrizio Falchi

1 Introduzione

Nell'ultima decade, approcci basati su tecniche di apprendimento automatico comunemente note come Machine Learning (ML) hanno giocato un ruolo chiave nello sviluppo di nuove metodologie scientifiche per l'analisi dati e di nuove tecnologie per applicazioni industriali.

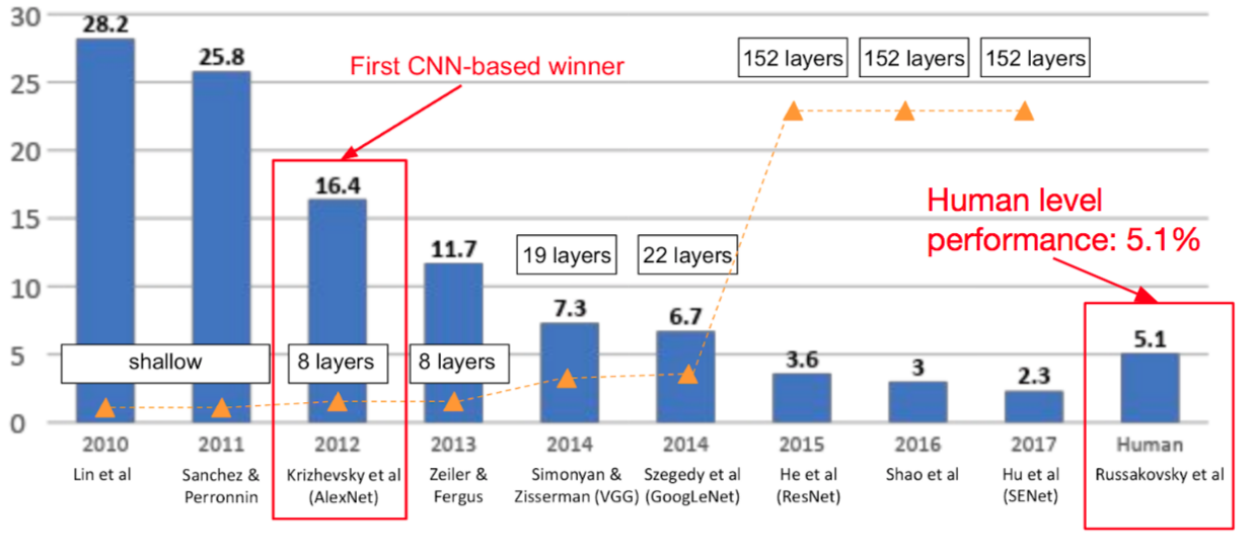
In particolare, una categoria di algoritmi appartenenti alla famiglia del ML, nota come Deep Learning (DL), ha attratto in maniera crescente l'attenzione del mondo accademico ed industriale assumendo un ruolo predominante in applicazioni legate ai campi della Computer Vision, Natural Language Processing, Cyber Security, Fraud Detection, Medical Imaging, ecc. Alla base del successo delle tecniche di DL vi è la capacità di questi algoritmi, noti anche come reti neurali artificiali profonde o Deep Neural Networks (DNNs), di imparare in maniera automatica a riconoscere delle strutture caratteristiche dei dati di input che ne permette l'analisi senza dover progettare ed implementare dei software ad-hoc per ogni caso specifico. Questi algoritmi riescono a raggiungere un tale obiettivo grazie alla loro capacità di imaparare a generare le rappresentazioni numeriche, dette "deep features", di cui hanno bisogno per raggiungere il proprio obiettivo (LeCun, Y., Bengio, Y. & Hinton, G 2015). Tuttavia, il successo delle DNNs è tipicamente legato alla disponibilità di grandi moli di dati che questi algoritmi richiedono per poter essere addestrati per un determinato scopo. Sebbene l'attuale era viene a volte definita come l'era dei "big data", ci possono essere situazioni in cui non sia possibile raccogliere dati a sufficienza per allenare una rete neurale. Per ovviare a tale problema, sono state proposte diverse tecniche tra le quali troviamo:

- "data augmentation", che consiste nell'applicare delle specifiche trasformazioni ai dati in modo da aumentare il numero di istanze presenti nel dataset;
- "one-shot learning" (Li, Fei-Fei, Rob Fergus, and Pietro Perona 2009), nel quale un problema di classificazione viene in genere riformulato in termini di confronto tra coppie di dati;
- "One-Class Classification" (Ruff, Lukas, et al. 2018), la cui prima applicazione ha riguardato la ricerca di anomalie, in cui una rete neurale viene allenata su una tipologia di istanza di dato.

Le tecniche citate sono solo un esempio di possibili approcci appartenenti ad una gamma vastissima di metodi sviluppati per far fronte a delle problematiche in cui ci si può imbattere quando si prova ad applicare algoritmi di DL per risolvere uno specifico problema.

1.1 Computer Vision e Deep Learning

Nel settore della computer vision, le tecniche di ML ed in particolare i metodi basati su rappresentazioni multiple e gerarchiche, ovvero il DL, hanno permesso di raggiungere risultati inimmaginabili fino ad una decina di anni fa. In particolare, grazie alla realizzazione delle cosiddette Deep Convolutional Neural Networks (DCNNs), le reti neurali artificiali hanno raggiunto prestazioni superiori a quelle umane in alcuni task specifici quali ad esempio la classificazione di immagini.



Progressione delle prestazioni di algoritmi di classificazione di immagini basati su tecniche di Deep Learning

Tale progresso scientifico e tecnologico ha permesso la realizzazione di sistemi in grado di processare in maniera automatica dati di input, quali ad esempio le immagini, con una percentuale di errore inferiore rispetto a quella umana. Tale risultati hanno suscitato un forte del mondo dell'industria conferendo alle reti neurali artificiali un ruolo chiave in settori quali, self-driving car, sistemi di sicurezza basati su face detection e recognition, ecc.

Uno dei principali vantaggi legati all'utilizzo di tecniche di DL è che l'implementazione di questi algoritmi non è connessa in maniera intima con l'obiettivo in analisi. Di fatti una stessa rete neurale, allenata su dati differenti è in grado di assolvere a compiti diversi. Ad esempio una DCNN può essere allenata a riconoscere dei caratteri scritti a mano o a segmentare un'immagine andando a classificare ogni singolo pixel contenuto in essa.

Da quanto detto finora si evince quindi che, oltre all'implementazione dell'algoritmo in sè, l'addestramento delle DNN rappresenta una delle fasi più delicate nella progettazione ed implementazione di sistemi basati su reti neurali artificiali.

Nel contesto della computer vision, negli ultimi anni sono stati resi pubblici una gran quantità di banche dati adatte all'addestramento di questi algoritmi. Tra i più comunemente utilizzati per scopi di classificazione di immagini troviamo MNIST, contenente immagini in scala di grigi rappresentanti numeri da 0 a 9 scritti a mano, CIFAR10 e CIFAR100, contenenti rispettivamente 10 e 100 classi di vari oggetti, ed ImageNet, contenente immagini appartenenti ad oltre ventimila classi. Questi dataset sono principalmente utilizzati per testare nuove architetture e/o tecniche di addestramento. Quando invece si passa dal settore accademico a quello industriale, in genere si preferisce utilizzare dataset più specifici per l'obiettivo in questione.

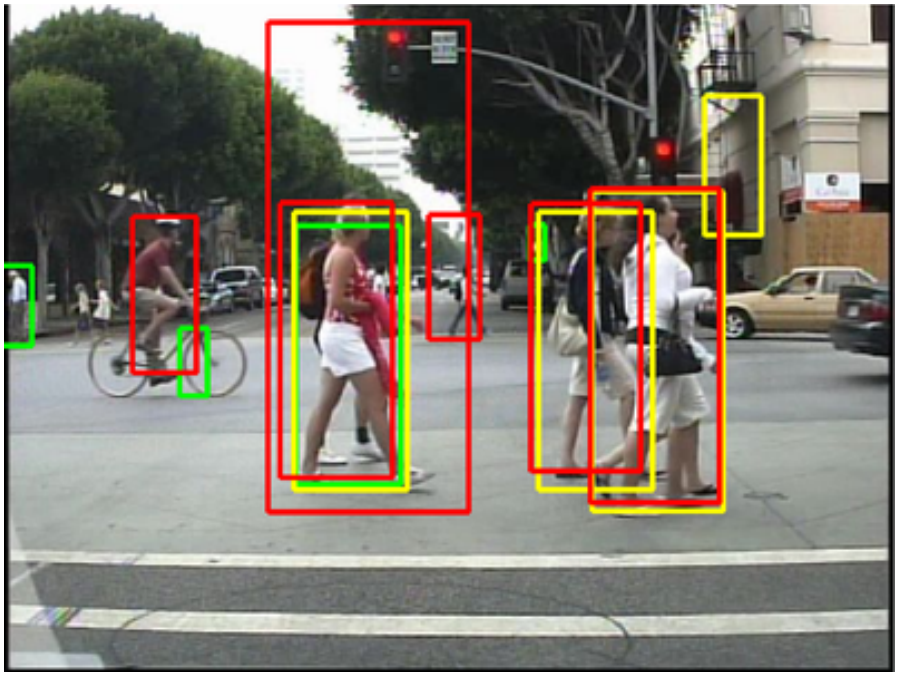
1.2 Tecniche di conteggio delle persone

Il People Counting (o Crowd Counting) consiste nel contare il numero di istanze di persone fisiche presenti in una data immagine. E' utile, ad esempio, per stimare il numero di individui che transitano in una certa area e quindi per fare statistiche sul numero di presenze in un dato istante o identificare zone di sovraffollamento. Applicazioni di maggiore interesse, tipicamente richiedono l'utilizzo di tale tecnica su stream video, ad esempio da telecamere di sorveglianza. Tra gli approcci proposti nella letteratura recente, è possibile raggruppare quelli che hanno raggiunto le prestazioni migliori in due categorie principali:

- metodi basati su detection;

- metodi basati su stime di densità.

I primi sfruttano i cosiddetti "object detectors", ovvero algoritmi in grado di identificare una specifica tipologia di struttura, ad esempio sagome di individui, all'interno di immagini. Una volta individuate le persone, è sufficiente contare il numero di istanze trovate. Tali algoritmi sono in genere allenati su dataset costruiti considerando il task della rivelazioni di pedoni (Pedestrian Detection).



Esempio di un algoritmo per la "Pedestrian Detection"

La metodologia appena descritta è caratterizzata da un elevato indice di flessibilità, in quanto permette di adattare algoritmi precedentemente progettati per scopi differenti. Per esempio, in (Amato et al. 2019) e (Ciampi et al. 2018), gli autori hanno presentato una soluzione per il conteggio di veicoli usando smart cameras, basata su "object detection". Tale approccio può essere facilmente adattato al conteggio di persone ad esempio in situazioni non di grande affollamento.

Nel caso di luoghi affollati invece, tecniche basate sulla stima della densità di istanze (Y. Zhang et al. 2016, Boominathan, Kruthiventi, and Babu 2016), raggiunge prestazioni maggiori rispetto all'approccio descritto in precedenza. Per situazioni di grande affollamento, tipicamente ci si riferisce a luoghi in cui non è sempre possibile distinguere ogni singolo individuo indipendentemente dagli altri (ad esempio: concerti, code d'attesa, strade molto frequentate). In questo caso, l'obiettivo della rete neurale artificiale è di stimare la densità di persone per unità di area da cui poi calcolare il numero totale di individui presenti.

Confronto tra tecniche per il conteggio di persone. Sinistra: Basato su Face Detection. Destra: Mappa di densità

I due approcci finora descritti richiedono una metodologia di addestramento detta supervisionata. Ovvero, necessita di dati di input di cui si conosce l'output desiderato. Tuttavia, più di recente, metodi basati su strategie di tipo self-supervised e unsupervised sono state proposte in letteratura. In particolare, in (X. Liu, Van De Weijer, and Bagdanov 2018) gli autori hanno utilizzato un approccio self-supervised sfruttando il paradigma del "learning-to-rank", mentre in (Sam, Sajjan, et al. 2019) è stata proposta una tecnica basata sul principio del "winner-takes-all" applicato alle rappresentazioni prodotte da una DCNN dando vita ad un framework non supervisionato. Differentemente, gli autori in (H. Xiong et al. 2019) hanno proposto un approccio self-supervised basato sull'idea che il conteggio si può decomporre lungo le componenti spaziali dell'immagine agendo indipendentemente su partizioni di esse. In (Shen et al. 2018) viene utilizzato un approccio di adversarial learning per ridurre gli effetti di blurring tipici dei metodi di stima tramite mappe di densità, utilizzando anche in questo caso una loss che garantisce consistenza spaziale del conteggio.

Recentemente, alcuni lavori come (Lian et al. 2019), (Sam, Peri, et al. 2019) e (Y. Liu et al. 2020) hanno tentato con successo di unire approcci di regressione con approcci di detection, nel tentativo di sfruttare al meglio i vantaggi di entrambe le metodologie.

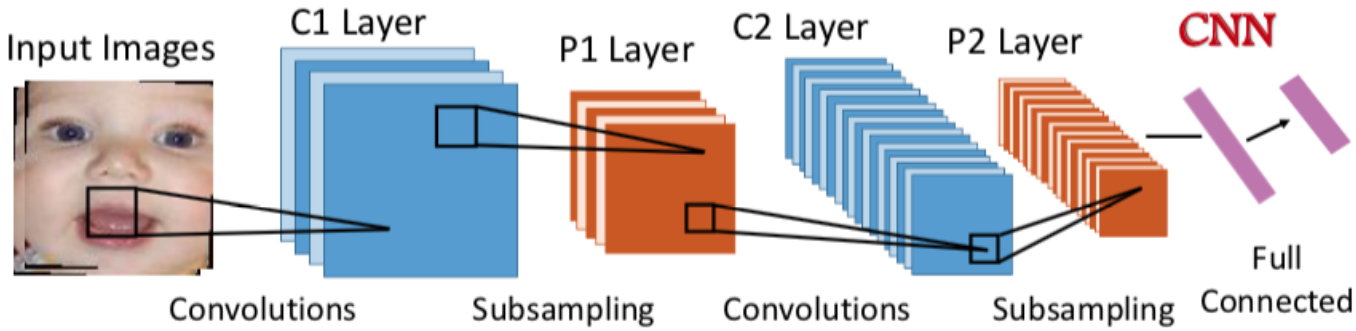
1.3 Riconoscimento delle emozioni

Il riconoscimento automatico delle espressioni facciali (FER) rappresenta una sfida aperta per sistemi di apprendimento automatico. Oltre all'interesse scientifico, le tecniche di FER riscuotono un grande interesse da parte del settore industriale tramite applicazioni quali: sociable robots, medical treatment, driver fatigue surveillance, e svariati altri campi in cui l'interazione uomo-macchina assume un ruolo di primo piano. Nel contesto del DL, una delle tecniche più comunemente utilizzata è basata sulla discretizzazione e categorizzazione delle emozioni in termini di sei classi di base. La scelta di tali espressioni come indipendenti dalla cultura dell'osservatore, e quindi in grado di essere identificate in maniera universale, è stata oggetto di dibattiti scientifici nel secolo scorso (Ekman, Paul, and Wallace V. Friesen 1971, Ekman, Paul 1994, Matsumoto, David 1992). Nel caso in esame, le reti neurali vengono quindi addestrate a riconoscere le seguenti sei espressioni: rabbia, disgusto, paura, felicità, tristezza, e sorpresa. A tali categorie viene poi implicitamente aggiunta l'espressione cosiddetta neutra che rappresenta l'assenza di una delle sei classi precedentemente menzionate.

Esempi di immagini presenti in dataset pubblici utilizzati per l'addestramento di reti neurali sul task del FER

I sistemi utilizzati per FER si possono distinguere in due categorie a seconda della tipologia di input:

- "static-based methods", nel caso in cui la rete neurale lavori su singole immagini (Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. 2016, Dehghan, Afshin, et al. 2017, Kuo, Chieh-Ming, Shang-Hong Lai, and Michel Sarkis 2018);
- "dynamic-based methods", nel caso in cui un sistema di DL viene impiegato per analizzare strewam video (Jung, Heechul, et al. 2015, Zhao, Xiangyun, et al. 2016, Baddar, Wissam J., and Yong Man Ro. 2018, Meng, Debin, et al. 2019).



Esempio schematico di una rete convoluzionale utilizzata nella classificazione delle emozioni

Sebbene viviamo nell'era dei "big data" l'addestramento di modelli di DL per il riconoscimento di espressioni facciali rimane un compito arduo a causa della mancanza di dataset di grandi dimensioni progettati per questo specifico obiettivo. Tra i dataset di maggiore rilevanza vi sono "Affect-in-the-wild 2" (Kollias, Dimitrios, and Stefanos Zafeiriou. 2018) e "Google facial expression comparison dataset" (Vemulapalli, Raviteja, and Aseem Agarwala 2019).

Per mitigare tale problema, in letteratura sono state proposte tecniche di apprendimento che prevedono l'utilizzo di stadi di pre-addestramento su dataset progettati per scopi diversi quali, ad esempio, il riconoscimento facciale (Face Recognition, vedi sezione successiva). Ad esempio, in (Kaya, Heysem, Furkan Gürpınar, and Albert Ali Salah 2017) gli autori suggeriscono che una fase di pre-allenamento sul dataset VGGFace (Parkhi, Omkar M., Andrea

Vedaldi, and Andrew Zisserman, 2015) seguita poi dall'addestramento sul task FER permette ad una rete neurale di raggiungere prestazioni più elevate rispetto a quando l'intero processo di addestramento si basa esclusivamente sull'utilizzo di dataset per il FER. Considerando modelli che accettano come input singole immagini, in (Yao, Anbang, et al. 2016) gli autori propongono un modulo di tipo "inception" in grado di apprendere features discriminative per il task di FER. Con l'obiettivo di ridurre il più possibile la similarità tra espressioni differenti, aumentando nel contempo quella tra emozioni simili, in (Guo, Yanan, et al. 2016) gli autori formalizzano una funzione obiettivo denominata "exponential triplet-based" mentre in (Liu, Xiaofeng, et al. 2017) la "(N+M)-tuples cluster loss" viene proposta per alleviare il costo della formazione di tripletti. Tecniche di "ensemble" (Hamester, Dennis, Pablo Barros, and Stefan Wermter 2015, Bargal, Sarah Adel, et al. 2016, Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu 2019) e "multi-task" (Jang, Youngkyoon, Hatice Gunes, and Ioannis Patras 2017, Pons, Gerard, and David Masip 2018, Zhang, Zhanpeng, et al. 2018) hanno mostrato prestazioni di rilievo permettendo nello specifico di generare delle rappresentazioni altamente discriminative al fine del FER. Negli ultimi anche approcci basati su architetture di tipo "GAN" hanno mostrato enormi potenzialità. Ad esempio in (Yang, Huiyuan, Zheng Zhang, and Lijun Yin 2018), gli autori propongono un modello costituito da due componenti in cui una si occupa di generare immagini di uno stesso soggetto con diverse espressioni mentre l'altra soddisfa il task di FER vero e proprio. In (Chen, Jiawei, Janusz Konrad, and Prakash Ishwar 2018), viene proposta un'architettura che combina insieme due modelli generativi in grado di disaccoppiare informazioni riguardanti l'identità di un soggetto dalle espressioni facciali. Per quel che concerne invece i "dynamic-based methods", in (Kim, Dae Hoe, et al. 2017) gli autori sfruttano cinque diverse funzioni obiettivo allo scopo di regolarizzare la rete neurale nel tentativo di minimizzare l'errore di classificazione. In (Yu, Zhenbo, Qinshan Liu, and Guangcan Liu 2018) viene introdotta un'architettura denominata "Cascaded Peak-Piloted Network" allenata tramite la tecnica del "cascade fine- tuning" in grado di ridurre fortemente problematiche legati al fenomeno dell'overfitting, mentre in (Kim, Youngsung, et al. 2017) viene sfruttato un approccio di tipo generativo per eliminare informazioni non critiche per il task del FER.

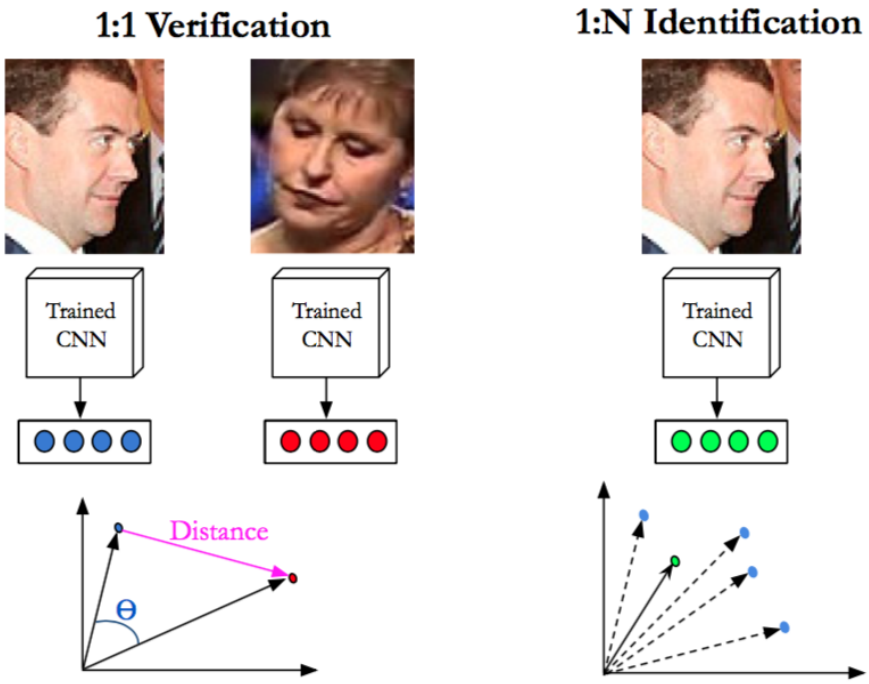
1.4 Riconoscimento facciale

Il Face Recognition (FR) o riconoscimento facciale rappresenta uno dei temi di maggiore interesse nella comunità della computer vision. Quando si parla di FR ci si riferisce a sistemi in grado di identificare una persona a partire da una foto del suo volto. Per misurare le prestazione di tali algoritmi, tipicamente si utilizzano due protocolli (Whitelam, Cameron, et al. 2017):

- 1:1 Face Verification, date due immagini di due volte la rete neurale stabilisce se esse appartengono alla stessa persone oppure no;
- 1:N Face Identification, data un'immagine di una faccia, l'algoritmo ne genera una rappresentazione numerica, ovvero un vettore di features, che viene poi utilizzata per cercare all'interno di un database se il soggetto ripreso è noto oppure no. Tale protocollo si distingue inoltre in "close" qualora si è sicuri che il soggetto è presente nel database di identità note e "open" in caso contrario.

Questa tipologia di tecniche trova impiego in applicazioni quali sistemi di sorveglianza, forensi, ecc. Ad esempio, il protocollo di "face verification" potrebbe essere utilizzato per l'autorizzazione all'ingresso in zone ad accesso ristretto mentre il protocollo di "face identification" trova impiego qualora si voglia sapere se una data persona appartiene alla lista di latanti. Algoritmi di DL applicati nel settore del FR vengono in genere allenati tramite addestramento supervisionato e per tale motivo richiedono dataset di grandi dimensioni per tener conto della enorme variabilità in termini fisici ed ambientali con cui un volto umano si presenta ad una telecamera. Dunque, dataset di grandi dimensioni sono necessari. Tra le banche di dati più utilizzate troviamo "Labeled Face in the Wild (LFW)" (Huang, Gary B., et al. 2008) e "Visual Geometry Group Face 2 (VGGFace2)" (Cao, Qiong, et al. 2018),

entrambi caratterizzati da una molteplicità di immagini dell'ordine del milione. In (Wu, Yudong, et al. 2020) gli autori focalizzano il proprio studio sul problema del FR considerando features quantizzate (low-bit quantization). Nello specifico, l'obiettivo dello studio è la riduzione dell'errore derivante da operazioni di quantizzazione tramite la ridefinizione del problema in un cosiddetto "angular space". In (Liu, Hao, et al 2019) viene proposta una soluzione al problema dello sbilanciamento delle classi all'interno dei dataset. A tale scopo, gli autori formulano la "Adaptive Margin Softmax", ovvero una funzione obiettivo usata durante l'addestramento delle reti neurali in grado di modificare il margine a seconda della classe in esame. In (Yin, Xi, et al. 2019) lo stesso problema viene affrontato invece nello spazio delle features dove si cerca di cerca di riprodurre la stessa variabilità tra le classi meno rappresentate e quelle con maggiore molteplicità.



Protocolli di misura delle prestazioni di sistemi per il Face Recognition. Sinistra: distanza euclidea tra rappresentazioni numeriche generate a partire da due immagini per determinare se appartengono alla stessa persona oppure no. Destra: la rappresentazione numerica generata a partire da un'immagine di input viene confrontata con un database di identità note per stabilire se il soggetto in questione è noto oppure no.

Nonostante il grande successo raggiunto nel FR, le prestazioni degli algoritmi di deep learning sono soggette a forte degrado quando vengono utilizzate immagini a risoluzioni differenti (Massoli, F.V., et al. 2020). Tale problema è legato principalmente alla composizione dei dataset che vengono utilizzati in fase di addestramento. Di fatti, questi contengono in genere immagini ad alta risoluzione. Ad esempio, la risoluzione media delle immagini presenti nel dataset VGGFace2 (Cao, Qiong, et al. 2018) è maggiore di 64 pixels (considerando il lato minore dell'immagine) mentre in sistemi di sorveglianza si possono avere immagini con risoluzioni basse fino ad 8x8 pixels.

Immagine di un volto acquisita tramite una telecamera di sorveglianza

Tra le diverse tecniche proposte per fronteggiare tale problema troviamo:

- SuperResolution (SR),
- proiezioni in uno spazio comune,
- Cross-Resolution training

Per quanto riguarda il primo approccio, a volte anche chiamato anche Face Hallucination, esso consiste nel generare un'immagine ad alta risoluzione a partire da una di minore qualità. In (Singh, Maneet, et al. 2018) gli autori propongono un approccio basato sull'apprendimento di rappresentazioni sparse per le immagini ad alta e bassa risoluzione e di una trasformazione tra le due rappresentazioni in grado di fornire un vettore di features per il volto di bassa qualità quanto più simile possibile a quello ottenuto da facce ad alta risoluzione. In (Kolouri, Soheil, and Gustavo K. Rohde 2015) viene proposto un algoritmo in grado di apprendere un modello Lagrangiano non-lineare per le immagini ad alta risoluzione utilizzato successivamente per generare un'immagine ad alta risoluzione a partire dalla query di minor qualità.

Per quanto concerne le tecniche di proiezioni in spazi comuni per immagini di diversa risoluzioni, troviamo il lavoro di (Li, Pei, et al., 2019) nel quale gli autori addestrano una rete neurale a mappare immagini proveniente da una stessa identità, ma di diversa risoluzione, in uno spazio delle rappresentazioni comune cercando di ridurre quanto più possibile la variabilità delle features relative ad un singolo individuo a dispetto della risoluzione delle immagini.

Nonostante le tecniche di SR abbiano in generale offerto una soluzione molto interessante al problema del FR in contesti di immagini con risoluzione diverse, questi soffrono di un problema non trascurabile: poiché tali algoritmi sono progettati per generare immagini di alta qualità visuale, nel processo di sintesi non tengono conto dell'obiettivo finale del FR per cui le immagini prodotte, sebbene di alta qualità, risultano "difficili da comprendere" da parte di una rete neurale che si occupa di FR (Zhang, Kaipeng, et al. 2018) nel senso che le features generate non riescono ad avere un elevato contenuto discriminante al fine del riconoscimento facciale.

Approcci basati su un addestramento di tipo Cross-Resolution hanno dato prova di poter raggiungere risultati a livello dello stato dell'arte pre-esistente se non migliori. In (Massoli, F.V., et al. 2020) gli autori propongono un approccio basato su di un paradigma di addestramento denominato "Teacher-Student" in cui un modello allenato su immagini ad alta risoluzione funge da guida per un altro modello addestrato con immagini di varia risoluzione. Tale approccio ha dimostrato delle prestazioni elevate, al di sopra dello stato dell'arte in diversi setting per quanto

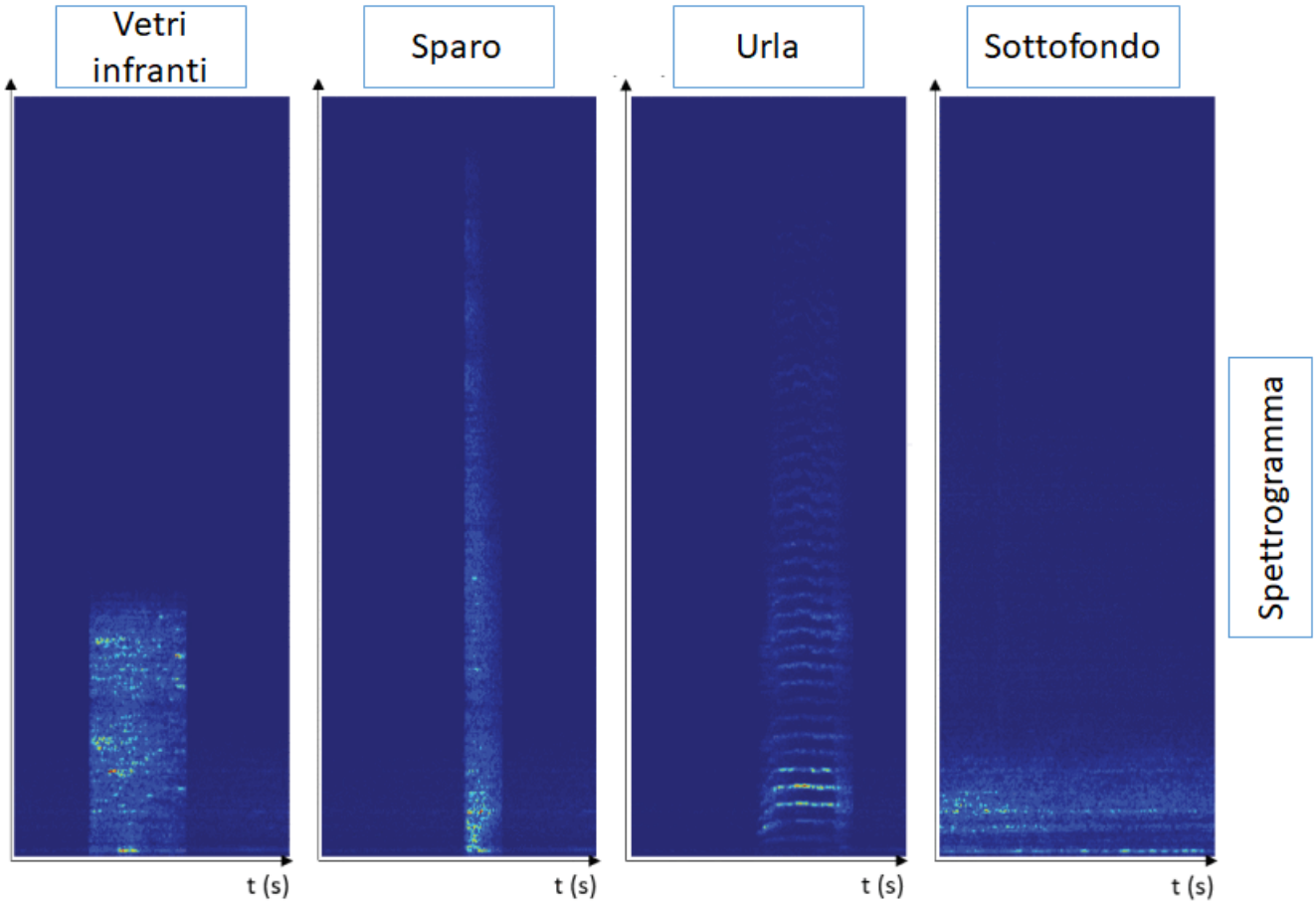
riguarda il problema del FR in contesti cross-resolution. considerando risoluzioni da 8x fino a 64x pixel. Per risoluzioni maggiori, questa tecnica fornisce invece prestazioni comparabili con lo stato dell'arte.

1.5 Analisi audio

Per quanto riguarda il riconoscimento di segnali sonori catturati utilizzando microfoni, le modalità di analisi ed elaborazione saranno, a livello di tecnologie utilizzate, simili a quelle descritte per l'analisi delle immagini. Le attività di interesse sono principalmente due:

- L'analisi di un segnale audio finalizzata ad identificare caratteristiche che contribuiscono a definirne la tipologia (a partire dal livello di rumore fino al riconoscimento di suoni che possano ricondurre a urla, segnali di pericolo, spari, etc.)
- L'analisi multi segnale, tramite l'utilizzo di più microfoni, che permetta di riconoscere la provenienza e in qualche modo aiuti a localizzare la fonte che genera il segnale audio.

Lo stato dell'arte nel riconoscimento di segnali audio in vari tipi di ambienti (e.g. strade trafficate, locali, foresta, piazze, bus, auto, etc.) si basa essenzialmente su cinque tipi di modelli: Gaussian Mixture Model (GMM), Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Deep Neural Network (CNN) e i-vector. Considerando che nel nostro caso ci troviamo in un ambiente urbano è molto importante poter effettuare il riconoscimento in un ambiente rumoroso, (Cha Zhang, Dinei Florencio, et al. 2008) considera il riconoscimento di suoni in presenza di rumore costante, mentre (L. Gerosa, G. Valenzise, et al. 2007) si concentra su problemi di sicurezza andando ad analizzare urla e spari in ambiente rumoroso.



Spettrogrammi ricavati da diversi suoni di interesse ai fini delle telesorveglianza tratti dal dataset MIVIA

In figura sono riportati alcuni spettrogrammi ricavati da suoni di interesse ai fini della sorveglianza urbana, oltre al rumore di sottofondo. Una tendenza comune che si è affermata nell'ultimo periodo è quella di ottenere dei sistemi di classificazione per l'audio mutuandoli da metodi già realizzati per la classificazione di immagini. In particolare, un suono viene rappresentato attraverso il suo spettrogramma che convertito in una immagine simile a quelle riportate in figura può essere quindi classificato da reti neurali originariamente pensate per l'analisi di immagini (Greco, A., Saggese, A., Vento, M., & Vigilant 2019).

Anche il riconoscimento vocale (Lawrence R. Rabiner, et al. 1993) sarà preso in considerazione e testato durante la fase di sperimentazione. Le features sulle quali vengono testati i modelli sono tipicamente Mel-frequency cepstral coefficients (MFCC), Binaural MFCC, log Mel-spectrum e Largescale temporal pooling features estratte utilizzando OpenSMILE. Prima dell'avvento del Deep Learning, i classificatori convenzionali come GMM erano i più utilizzati, ma ad oggi soluzioni tipo DNN, RNN, CNN sono quelle che danno i risultati migliori con una maggiore percentuale di accuratezza. Per quanto concerne il riconoscimento spaziale di suoni (ovvero capire da dove viene generato un suono) si utilizza come minimo una coppia di microfoni; all'aumentare del numero di microfoni aumenta la precisione nel riconoscimento. Le tecniche utilizzate si basano principalmente sul BeamForming; a tal proposito (A. Mahajan and M. Walworth, 2001) descrivono diverse approssimazioni di come calcolare la posizione della sorgente audio.

Per il training, il test e la validazione degli algoritmi oltre ai dati cha sarà possibile raccogliere nell'ambiente urbano della città di Lucca ci si riferirà ad alcuni datasets disponibili che risulteranno particolarmente utili per il benchmarking e il confronto con gli algoritmi e i metodi proposti da altri gruppi di ricerca. Si cita tra i dataset

disponibili il già menzionato "Mivia Audio Events Dataset" (Foggia, P. et al 2015). Tale dataset è composto da un totale di 6000 eventi per applicazioni di sorveglianza, ovvero rottura di vetri, spari e urla. I 6000 eventi sono già suddivisi in un training set (composto da 4200 eventi) e un test set (composto da 1800 eventi) per agevolare il confronto di algoritmi basati su machine learning. Il dataset è stato pensato per essere rappresentativo di diverse situazioni in cui ad esempio gli eventi di interesse (un urlo o uno sparo) possono verificarsi a diverse distanze dal microfono, con diversi livelli del rapporto segnale-rumore. Il dataset inoltre include eventi registrati con rumori di background complessi che sono rappresentativi sia degli ambienti indoor (ad esempio elettrodomestici in funzione) ma soprattutto degli ambienti outdoor (applausi, persone che parlano, traffico intenso, passaggio di veicoli rumorosi come motociclette). Ogni evento è rappresentato da sei tracce audio con diversi valori del rapporto segnale-rumore e/o con sovrapposizioni di rumori ambientali. I suoni sono stati registrati con un modulo audio Axis P8221 e un microfono omnidirezionale Axis T83. Sono stati campionati a 32000 Hz e quantizzati a 16 bit con modulazione a impulsi codificati.

Un ulteriore dataset di interesse per la localizzazione delle sorgenti audio mediante array di microfoni è rappresentato dal "MIVIA Audio Localization" (Saggese et al, 2017), che ha il duplice scopo di valutare le prestazioni dei sistemi di localizzazione con dispositivi low-cost a bassa fedeltà e di valutare le prestazioni dell'approccio proposto variando la distanza tra i microfoni. I suoni sono stati registrati utilizzando l'array di microfoni di un sensore Kinect. Il set hardware è stato ampliato in modo da realizzare un array di quattro microfoni, la cui distanza massima tra loro è di 138 cm.

Referenze

Amato, G., L. Ciampi, F. Falchi, and C. Gennaro. 2019. "Counting Vehicles with Deep Learning in Onboard UAV Imagery." In *2019 IEEE Symposium on Computers and Communications (ISCC)*, 1–6.

Baddar, Wissam J., and Yong Man Ro. "Learning spatio-temporal features with partial expression sequences for on-the-fly prediction." *arXiv preprint arXiv:1711.10914* (2018).

Bargal, Sarah Adel, et al. "Emotion recognition in the wild from videos using images." *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016.

Boominathan, Lokesh, Srinivas S. S. Kruthiventi, and R. Venkatesh Babu. 2016. "CrowdNet: A Deep Convolutional Network for Dense Crowd Counting." In *Proceedings of the 24th ACM International Conference on Multimedia*, 640–44. MM '16. New York, NY, USA: Association for Computing Machinery.

Cao, Qiong, et al. "Vggface2: A dataset for recognising faces across pose and age." *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018.

Chen, Jiawei, Janusz Konrad, and Prakash Ishwar. "Vgan-based image representation learning for privacy-preserving facial expression recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

Ciampi, Luca, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti. 2018. "Counting Vehicles with Cameras." In *SEBD*. http://ceur-ws.org/Vol-2161/paper12.pdf.

Dehghan, Afshin, et al. "Dager: Deep age, gender and emotion recognition using convolutional neural network." *arXiv preprint arXiv:1702.04280* (2017).

Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." *Journal of personality and social psychology* 17.2 (1971): 124.

Ekman, Paul. "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique."

(1994): 268.

Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., & Vento, M. (2015). Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, *65*, 22-28.

Georgescu, Mariana-Iuliana, Radu Tudor Ionescu, and Marius Popescu. "Local learning with deep and handcrafted features for facial expression recognition." *IEEE Access* 7 (2019): 64827-64836.

Greco, A., Saggese, A., Vento, M., & Vigilante, V. (2019, October). SoReNet: a novel deep network for audio surveillance applications. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (pp. 546-551). IEEE.

Guo, Yanan, et al. "Deep neural networks with relativity learning for facial expression recognition." *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016.

Huang, Gary B., et al. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." 2008.

Jang, Youngkyoon, Hatice Gunes, and Ioannis Patras. "SmileNet: registration-free smiling face detection in the wild." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017.

Jung, Heechul, et al. "Joint fine-tuning in deep neural networks for facial expression recognition." *Proceedings of the IEEE international conference on computer vision*. 2015.

Hamester, Dennis, Pablo Barros, and Stefan Wermter. "Face expression recognition with a 2-channel convolutional neural network." *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015.

Kaya, Heysem, Furkan Gürpınar, and Albert Ali Salah. "Video-based emotion recognition in the wild using deep transfer learning and score fusion." *Image and Vision Computing* 65 (2017): 66-75.

Kim, Dae Hoe, et al. "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition." *IEEE Transactions on Affective Computing* 10.2 (2017): 223-236.

Kim, Youngsung, et al. "Deep generative-contrastive networks for facial expression recognition." *arXiv preprint arXiv:1703.07140* (2017).

Kollias, Dimitrios, and Stefanos Zafeiriou. "Aff-wild2: Extending the aff-wild database for affect recognition." *arXiv preprint arXiv:1811.07770* (2018).

Kolouri, Soheil, and Gustavo K. Rohde. "Transport-based single frame super resolution of very low resolution face images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

Kuo, Chieh-Ming, Shang-Hong Lai, and Michel Sarkis. "A compact deep learning model for robust facial expression recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436–444 (2015)

Lian, Dongze, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. 2019. "Density Map Regression Guided Detection Network for Rgb-D Crowd Counting and Localization." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1821–30.

Li, Fei-Fei, Rob Fergus, and Pietro Perona. "One-shot learning of object categories." *IEEE transactions on pattern analysis and machine intelligence* 28.4 (2006): 594-611.

Li, Pei, et al. "On low-resolution face recognition in the wild: Comparisons and new techniques." *IEEE Transactions on Information Forensics and Security* 14.8 (2019): 2000-2012.

Liu, Hao, et al. "Adaptiveface: Adaptive margin and sampling for face recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Liu, Xialei, Joost Van De Weijer, and Andrew D. Bagdanov. 2018. "Leveraging Unlabeled Data for Crowd Counting by Learning to Rank." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7661–69.

Liu, Xiaofeng, et al. "Adaptive deep metric learning for identity-aware facial expression recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

Liu, Yuting, Zheng Wang, Miaojing Shi, Shin 'ichi Satoh, Qijun Zhao, and Hongyu Yang. 2020. "Towards Unsupervised Crowd Counting via Regression-Detection Bi-Knowledge Transfer." In *Proceedings of the 28th ACM International Conference on Multimedia*, 129–37. MM '20. New York, NY, USA: Association for Computing Machinery.

Massoli, Fabio Valerio, Giuseppe Amato, and Fabrizio Falchi. "Cross-resolution learning for Face Recognition." *Image and Vision Computing* (2020): 103927.

Matsumoto, David. "More evidence for the universality of a contempt expression." *Motivation and Emotion* 16.4 (1992): 363-368.

Meng, Debin, et al. "Frame attention networks for facial expression recognition in videos." *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

Mollahosseini, Ali, David Chan, and Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016.

Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015).

Pons, Gerard, and David Masip. "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition." *arXiv preprint arXiv:1802.06664* (2018).

Ruff, Lukas, et al. "Deep one-class classification." *International conference on machine learning*. 2018.

Saggese, A., Strisciuglio, N., Vento, M., & Petkov, N. (2017, August). 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)* (pp. 1-7). IEEE.

Sam, Deepak Babu, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R. Venkatesh Babu. 2019. "Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1906.07538.

Sam, Deepak Babu, Neeraj N. Sajjan, Himanshu Maurya, and R. Venkatesh Babu. 2019. "Almost Unsupervised Learning for Dense Crowd Counting." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8868–75.

Shen, Zan, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. 2018. "Crowd Counting via Adversarial Cross-Scale Consistency Pursuit." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5245–54.

Singh, Maneet, et al. "Identity aware synthesis for cross resolution face recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

Vemulapalli, Raviteja, and Aseem Agarwala. "A compact embedding for facial expression similarity." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019.

Whitelam, Cameron, et al. "Iarpa janus benchmark-b face dataset." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

Wu, Yudong, et al. "Rotation consistent margin loss for efficient low-bit face recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Xiong, Haipeng, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. 2019. "From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2019.00845.

Yao, Anbang, et al. "HoloNet: towards robust emotion recognition in the wild." *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016.

Yang, Huiyuan, Zheng Zhang, and Lijun Yin. "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks." *2018 13th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.

Yin, Xi, et al. "Feature transfer learning for face recognition with under-represented data."*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Yu, Zhenbo, Qinshan Liu, and Guangcan Liu. "Deeper cascaded peak-piloted network for weak expression recognition." *The Visual Computer* 34.12 (2018): 1691-1699.

Zhang, Kaipeng, et al. "Super-identity convolutional neural network for face hallucination." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Zhang, Yingying, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 589–97.

Zhao, Xiangyun, et al. "Peak-piloted deep network for facial expression recognition." *European conference on computer vision*. Springer, Cham, 2016.

Zhang, Zhanpeng, et al. "From facial expression recognition to interpersonal relation prediction." *International Journal of Computer Vision* 126.5 (2018): 550-569.

A. Mahajan and M. Walworth. 3-d position sensing using the difference in the time-of-flights from a wave source to various receivers, IEEE Transactions on Robotics and Automation, 17(1), pages 91-94, 2001.

J.M. Varin, F. Michaud, J.Rouat, D.Letourneau, Robust sound source localization using a microphone array on a mobile robot, IROS 2003, Vol.2, pages 1228-1233, Oct. 2003.

Cha Zhang, Dinei Florencio, Demba E. Ba, Zhengyou Zhang Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings, IEEE Transactions on Multimedia, April 2008, Vol. 10, num. 3, pages 538-548.

L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, A.Sarti Scream and gunshot detection in noisy environments, Signal Processing Conference, pages 1216 - 1220, 3-7 Sept. 2007.

Lawrence R. Rabiner, B. H. Juang, Fundamentals of speech recognition, Prentice Hall, 1◦ Edition, April 1993.

Juncheng Li*, Wei Dai*, Florian Metze*, Shuhui Qu, and Samarjit Das "A comparison of deep learning methods for environmental sound detection"