Open World Face Recognition in Reti di Telecamere non Sovrapposte

Tiberio Uricchio[1], Federico Pernici[1], Matteo Bruni[1], Federico Bartoli[1], Alberto Del Bimbo[1], Francesco Calabrò[2]

[1]Università degli Studi di Firenze
Media Integration and Communication Center
[2]Leonardo S.p.a.
[1]{nome.cognome@unifi.it} [2]{nome.cognome@leonardocompany.com}

Abstract

In questo articolo presentiamo un metodo incrementale per l'apprendimento di identità da flussi video non supervisionati. Il metodo usa descrittori facciali estratti da reti neurali profonde insieme ad un meccanismo di apprendimento basato su una memoria indirizzabile che sfrutta la coerenza temporale del flusso video. I risultati sperimentali mostrano che il metodo proposto raggiunge prestazioni comparabili ad approcci offline che sfruttano le informazioni future del flusso video. Inoltre si ottengono prestazioni superiori nell'identificazione facciale cosiddetta "open world" in cui soggetti non noti vengono automaticamente inclusi (*autoenrollment*) alla galleria dei soggetti noti.

1 Introduzione

Molte applicazioni di sicurezza richiedono il riconoscimento dell'identità delle persone che sono presenti in video. Mentre il riconoscimento di soggetti conosciuti (*"closed set"*) ha raggiunto prestazioni notevoli [Cao *et al.*, 2018], diverso è il caso di identificazione e tracciamento di persone sconosciute, denominato *"open world"*. In tal caso il sistema deve saper rilevare i soggetti non presenti nella galleria e aggiungerli via via che si presentano nel tempo. Questo scenario, estremamente importante per tracciare ed identificare gli spostamenti di soggetti comuni, è tuttora un problema di ricerca aperto.

In questo articolo presentiamo un metodo non supervisionato di tipo online per l'apprendimento di nuove identità da flussi video senza vincoli. Dato un video con persone da identificare, il metodo si basa su descrittori di facce derivati da reti convoluzionali profonde (CNN) che vengono collezionati in un modulo di memoria, distillati sulla base della loro ridondanza rispetto alla rappresentazione corrente. Questo permette la costruzione di una rappresentazione estesa dell'apparenza delle singole identità, via via che il tempo avanza.

2 Lavori associati

Il problema del riconoscimento di identità open world è stato affrontato per la prima volta in [Pernici e Del Bimbo, 2017] e successivamente esteso in [Pernici *et al.*, 2018]. In questi lavori, si introduce l'utilizzo della coerenza temporale e la creazione di una rappresentazione incrementale. In questo articolo espandiamo il metodo con l'utilizzo di più stream contemporanei e una galleria di soggetti già conosciuti all'inizializzazione del sistema. In letteratura il problema corrisponde ad identificare nuove classi in un sistema di apprendimento [Bendale e Boult, 2015] dove i metodi non parametrici risultano favoriti in quanto possono imparare classi mai viste semplicemente memorizzando i nuovi esempi in modo robusto come in [Mensink *et al.*, 2013]. In questo lavoro adottiamo quindi la stessa idea di classificazione basata su prototipi.

Figura 1: Rappresentazione facciale mediante memoria indirizzabile. *Sinistra*: Ogni elemento in memoria è costituito da un descrittore con un'identità associata (box colorato) e da un valore associato (box area grigia) che riflette il grado di ridondanza del descrittore rispetto alla rappresentazione corrente definita dalla memoria. *Destra*: Le regioni colorate mostrano la rappresentazione originale (i.e. VGG-face). I descrittori al di fuori da tali regioni sono stati appresi dal flusso video ed estendono la rappresentazione originale.

3 Approccio proposto

Il metodo incrementale proposto per l'apprendimento di identità si basa su un modulo di memoria $\mathcal{M}$, definito come un insieme di $N(t)$ tuple al tempo t:

$$\mathcal{M}(t) = \{(\mathbf{x}, \mathrm{Id}, e, a)_i\}_{i=1}^{N(t)} \quad (1)$$

dove x_i rappresenta il descrittore, Id_i l'identità associata al descrittore, e_i rappresenta il fattore di *eligibilità* (verrà di-

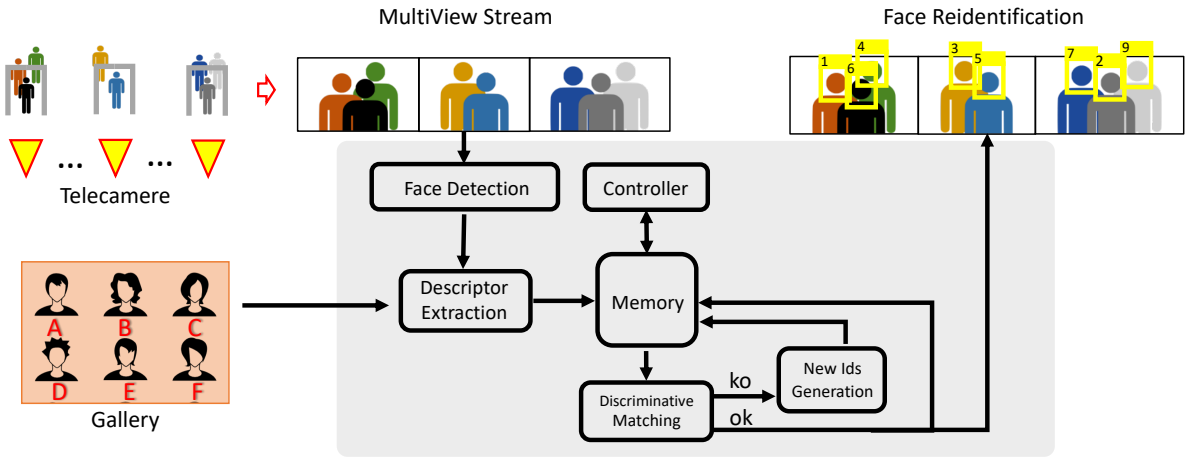

Figura 2: Schema del sistema. La figura mostra un insieme di telecamere non strettamente sincronizzate con campo di vista non sovrapposto. La galleria del sistema viene inizializzata manualmente con un insieme di soggetti noti. Il sistema processa un flusso multivista proveniente da tutte telecamere.

scusso in seguito), mentre a_i il tempo dal quale il descrittore non ha ottenuto più corrispondenze con il frame corrente.

In [Pernici *et al.*, 2018] durante l'inserimento di nuove identità, oltre alla loro persistenza temporale, viene verificato che non vi siano identità duplicate nel frame corrente (unicità dei soggetti). Tale vincolo viene qui esteso nel caso di reti di telecamere non strettamente sincronizzate con campo di vista non sovrapposto come in Figura 2. Dai fotogrammi acquisiti dalla rete di telecamere, vengono rilevate le facce, estratti i relativi descrittori e uniti come se fossero appartenenti ad un unico frame virtuale. Questo poiché un soggetto non può essere visibile allo stesso tempo su più telecamere. Tali descrittori vengono successivamente confrontati con quelli presenti nel modulo di memoria $\mathcal{M}$ considerando una specifica strategia basata su *Reverse Nearest Neighbor* (ReNN, vedi componente *Discriminative Matching* in Figura 2). In particolare, all'istante t vengono selezionati tutti quei descrittori $\mathbf{x}_i$ in memoria che soddisfano il seguente confronto:

$$\mathcal{M}^+ = \left\{ (\mathbf{x}, \mathrm{Id}, e, a)_i \in \mathcal{M}(t) \mid \frac{d_i^1}{d_i^2} < \bar{\rho} \right\} \qquad (2)$$

dove il rapporto $\frac{d_i^1}{d_i^2}$ è tra i due descrittori $\mathbf{x}_i$ più vicini presenti nel frame, con $\bar{\rho}$ una soglia fissata per il rapporto delle distanze. Solo i descrittori nel frame che non soddisfano il *Discriminative Matching* vengono inseriti nella memoria, associando a ciascuno di loro un nuovo identificativo, mentre per quelli rimanenti, viene associata l'identità Id_i che verifica il confronto ReNN. Viceversa, per ciascun descrittore $x_i \in \mathcal{M}^+$ viene resettato a 0 il relativo valore a_i.

Per poter mantenere in memoria solamente i descrittori più rilevanti, riducendo notevolmente lo spazio necessario per memorizzarli, viene considerato il fattore e_i chiamato di *eligibilità*. Al momento dell'inserimento di un nuovo descrittore x_i in $\mathcal{M}$, il relativo fattore e_i viene inizializzato a 1. Per ogni successiva corrispondenza basata su ReNN, tale fattore viene decrementato proporzionalmente in base al rapporto delle distanze:

$$e_i(t+1) = \eta_i \, e_i(t) \;\text{ with }\; \eta_i = \left[\frac{1}{\bar{\rho}} \frac{d_i^1}{d_i^2} \right]^{\alpha}, \qquad (3)$$

dove il parametro η_i permette di ridurre l'effetto del rapporto delle distanze. L'eligibilità permette di considerare la ridondanza spaziale dei descrittori ad un tasso proporzionale al numero di corrispondenze ottenute in frame consecutivi. Infine, un descrittore $\mathbf{x}_i$ viene rimosso dalla memoria se il relativo fattore e_i risulta inferiore a una soglia fissata $\hat{e}$. Nella pratica, questo avviene dopo un certo numero di corrispondenze del descrittore con i successivi estratti nei frame successivi.

Riferimenti bibliografici

[Bendale e Boult, 2015] Abhijit Bendale e Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, e Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 67–74. IEEE, 2018.

[Mensink *et al.*, 2013] Thomas Mensink, Jakob Verbeek, Florent Perronnin, e Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2624–2637, 2013.

[Pernici *et al.*, 2018] Federico Pernici, Federico Bartoli, Matteo Bruni, e Alberto Del Bimbo. Memory based online learning of deep representations from video streams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2324–2334, 2018.

[Pernici e Del Bimbo, 2017] Federico Pernici e Alberto Del Bimbo. Unsupervised incremental learning of deep descriptors from video streams. In *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*, pages 477–482. IEEE, 2017.