Computer Vision for Food Image Analysis

Dario Allegra, Giovanni Maria Farinella, Francesco Ragusa, Filippo Stanco
Dipartimento di Matematica e Informatica, Università degli studi di Catania
{allegra, gfarinella, fstanco}@dmi.unict.it, francesco.ragusa@unict.it

Abstract

La Computer Vision ha un ruolo fondamentale nell'estrazione automatica delle informazioni sugli alimenti consumati. In questo documento, descriveremo le ricerche effettuate dal gruppo IPLab, dell'Università degli Studi di Catania, per la realizzazione di un sistema intelligente di riconoscimento e analisi del cibo. Discuteremo le soluzioni proposte presentando i risultati scientifici ottenuti in questo contesto.

1 Introduzione

L'alimentazione rappresenta un importante aspetto della vita umana al punto da poter condizionare anche l'economia globale. Le scelte alimentari sono inoltre strettamente legate alla cultura di un popolo, rappresentano un indice della condizione finanziaria e, sopra ogni cosa, hanno un forte impatto sulla salute: si pensi ad esempio alla diffusione di malattie legate all'alimentazione come l'obesità, il diabete oppure le intolleranze e le allergie, che possono rendere estremamente rischiosa l'assunzione di alcuni alimenti. In questi casi risulta necessario un attento monitoraggio del cibo consumato al fine di minimizzare i rischi e di migliorare le condizioni di vita di un paziente. Tali considerazioni, hanno portato allo sviluppo di strategie di monitoraggio assistito e automatizzato attraverso l'uso di dispositivi portatili e indossabili equipaggiati con algoritmi di visione artificiale. Tuttavia, la ricerca di soluzioni per il riconoscimento automatico degli alimenti, la stima del volume e il successivo calcolo dei valori nutrizionali rappresenta un problema tutt'altro che banale, data la natura estremamente variabile con cui si presenta il cibo una volta trattato, cucinato e servito.

2 Problematiche e soluzioni proposte

Molti problemi di visione artificiale possono essere risolti costruendo modelli matematici di apprendimento automatico che elaborano grandi set di dati. Sebbene negli anni siano stati pubblicati un certo numero di dataset per la valutazione degli approcci di Computer Vision in diversi domini, nel 2014 esistevano solo poche collezioni di contenuti visuali (immagini, video, ecc.) relativi al cibo. Per questo motivo, parte del lavoro del nostro laboratorio si è concentrato sulla raccolta e l'organizzazione di dataset di cibo ad oggi pubblici e disponibili per la comunità scientifica ed industriale. Dal punto di vista algoritmico, il primo passo è quello di capire se il contenuto visuale analizzato in un'immagine rappresenta o meno del cibo. Il secondo passo è stato focalizzato, invece, nel riconoscere quale pietanza si ha davanti. Sebbene questo permetta di avere già una stima approssimata degli ingredienti che costituiscono un piatto, non è possibile fare precise valutazioni sui valori nutrizionali, poiché sarebbe richiesto un ulteriore passo, ossia la stima del volume.

2.1 Food VS Non-Food

Gli approcci dai noi proposti per la distinzione tra immagini raffiguranti cibo e quelle raffiguranti altri tipi di contenuti (classificazione binaria) si basano sul paradigma One-Class Classification, ossia la possibilità di costruire un modello matematico possedendo solamente i campioni di una delle due classi (cibo nel nostro caso). Questa idea nasce dal fatto che la quantità di possibili contenuti non appartenenti alla categoria "cibo" risulta pressoché illimitata. Per rappresentare le immagini, sono state utilizzate feature visuali classiche (SIFT, Bag of Words, ecc.) e feature estratte tramite reti neurali convoluzionali [Farinella *et al.*, 2015b; Ragusa *et al.*, 2016]. Per poter testare gli approcci proposti sono stati collezionati da Flickr due dataset da 4805 immagini di cibo e 8005 immagini raffiguranti altro. I risultati migliorano lo stato dell'arte raggiungendo un'accuratezza del **91.99%** con approcci di Deep Learning.

2.2 Riconoscimento

Il problema del riconoscimento e della classificazione del cibo risulta complesso a causa della grande varietà di ingredienti e delle numerose forme, tessiture e colori in cui essi si presentano. Intuendo il potere discriminatorio delle tessiture per questo tipo di immagini, nel 2014 abbiamo proposto un approccio basato su "Texton" che ha mostrato risultati migliori di quelli allo stato dell'arte sul dataset PFID [Farinella *et al.*, 2014; Farinella *et al.*, 2015c]). In ogni caso, dato il limitato numero di campioni in tale dataset e il fatto che essi erano stati acquisiti in ambiente controllato, abbiamo introdotto nel 2015 un nuovo dataset da 3583 immagini acquisite durante pasti reali e suddivise in 889 classi (UNICT-FD889) [Farinella *et al.*, 2015a]. Esperimenti di retrieval sul nuovo dataset, con l'approccio Bag of Texton, hanno portato ad un accura-

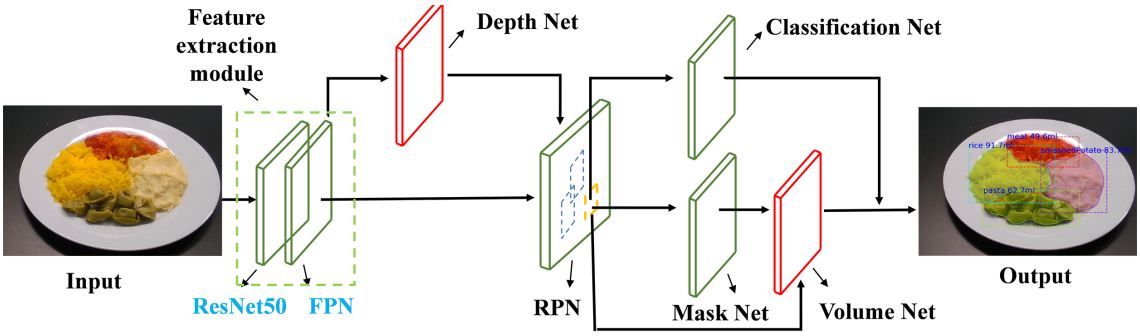

Figura 1: Pipeline proposta per la stima del volume.

tezza del **60.20%**. Nel 2016, il dataset è stato esteso a 1200 classi e 4754 immagini. Inoltre è stata fornita un'ulteriore categorizzazione in 8 classi, relative al tipo di portata (primo piatto, antipasto, dolce, ecc.) e una in 5 classi, relative alle posate da utilizzare per ingerirlo [Ragusa *et al.*, 2017]. Lavorando su uno spazio di colore differente e usando diversi tipi di filtri convolutivi si è riusciti a raggiungere un'accuratezza del **87.44%** sul test di retrieval. Avendo ormai appurato il vantaggio nell'utilizzo di informazioni estratte con operatori convoluzionali per rappresentare questo tipo di immagini, abbiamo proposto un nuovo descrittore chiamato Anti-Texton, che permette di catturare le relazioni spaziali tra i Texton. I risultati sulla classificazione rispetto al tipo di portata e rispetto al tipo di posate suggerite, hanno mostrato rispettivamente un'accuratezza del **92.60%** e del **86.27%**.

2.3 Stima del volume

Per avere una stima effettiva dei valori nutrizionali del pasto analizzato è necessario segmentare il contenuto del piatto e fornire una stima del volume di ogni elemento. Per affrontare il problema e validare la nostra strategia è stato introdotto un nuovo dataset con 80 portate differenti, acquisendo per ognuna di esse un modello 3D e una serie di immagini RGB comprensive di mappa di profondità (RGB-D), per un totale di 21807 immagini [Allegra *et al.*, 2017; Lu *et al.*, 2018]. Per ogni piatto, è stata annotata la massa dei singoli elementi e il volume. Utilizzando un approccio di multi-task learning basato su reti neurali convoluzionali, è stato costruito un modello capace di rilevare il piatto, identificare i singoli ingredienti, stimare la distanza dalla camera e valutare il volume di ognuno dei costituenti (Figura 1). Grazie all'approccio proposto si riesce a stimare il volume con un errore del **19.1%**, migliorando di molto rispetto allo stato dell'arte (errore del **36.1%**). Inoltre, il metodo proposto risulta di gran lunga più efficiente, con un tempo di elaborazione inferiore a $0.2s$ contro i $5.5s$ dello stato dell'arte.

3 Conclusioni

L'impatto sulla salute e la conseguente richiesta di sistemi di monitoraggio, ha spinto alla ricerca di metodi per l'inferenza automatica di informazioni sul cibo consumato. Sebbene una soluzione completa al problema risulti articolata sotto molti aspetti, i progressi del nostro studio, sulla stima del volume, sono molto promettenti. L'esperienza acquisita con i lavori esposti e la diffusione di dispositivi di acquisizione sempre più moderni (es. smartglasses indossabili), permetteranno di continuare a migliorar i risultati raggiunti. Le pubblicazioni scientifiche e i dataset utilizzati in questo ambito sono reperibili al seguente link: http://iplab.dmi.unict.it/UNICT-FD889/

Riferimenti bibliografici

[Allegra *et al.*, 2017] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, e S. Mougiakakou. A multimedia database for automatic meal assessment systems. In *Lecture Notes in Computer Science*, volume 10590, 2017.

[Farinella *et al.*, 2014] G. M. Farinella, M. Moltisanti, e S. Battiato. Classifying food images represented as bag of textons. In *International Conference on Image Processing*, pages 5212–5216, 10 2014.

[Farinella *et al.*, 2015a] G. M. Farinella, D. Allegra, e F. Stanco. A benchmark dataset to study the representation of food images. In *Lecture Notes in Computer Science*, volume 8927, pages 584–599, 3 2015.

[Farinella *et al.*, 2015b] G. M. Farinella, D. Allegra, F. Stanco, e S. Battiato. On the exploitation of one class classification to distinguish food vs non-food images. In *Lecture Notes in Computer Science*, volume 9281, pages 375–383, 2015.

[Farinella *et al.*, 2015c] G. M. Farinella, M. Moltisanti, e S. Battiato. Food recognition using consensus vocabularies. In *Lecture Notes in Computer Science*, volume 9281, pages 384–392, 2015.

[Lu *et al.*, 2018] Y. Lu, D. Allegra, M. Anthimopoulos, F. Stanco, G. M. Farinella, e S. Mougiakakou. A multi-task learning approach for meal assessment. In *Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*, pages 46–52, 2018.

[Ragusa *et al.*, 2016] F. Ragusa, V. Tomaselli, A. Furnari, S. Battiato, e G. M. Farinella. Food vs non-food classification. In *International Workshop on Multimedia Assisted Dietary Management*, pages 77–81, 2016.

[Ragusa *et al.*, 2017] F. Ragusa, A. Furnari, e G. M. Farinella. Understanding food images to recommend utensils during meals. In *Lecture Notes in Computer Science*, volume 10590, pages 419–425, 2017.