Tecniche avanzate per l'analisi di dati visuali tridimensionali attraverso apprendimento automatico

La nascita di nuove applicazioni legate al campo della robotica e della guida autonoma richiede l'elaborazione di dati di tipo tridimensionale. Grazie alla disponibilità dell'informazione di profondità definita come la distanza tra il sensore che ha acquisito i dati e gli oggetti presenti nella scena, è possibile avere una migliore conoscenza dello spazio in cui viviamo e con il quale interagiamo. Inoltre, la recente integrazione all'interno di dispositivi low-cost, come l'Apple Iphone, di sensori *time-of-flight* (ToF) ha alimentato la richiesta non solo a livello industriale ma anche a livello *consumer* di tecniche che siano in grado di elaborare dati 3D. Tali dati possono essere acquisiti con sensori di tipo *Lidar* (*Light Detection and Ranging)* o *RGB-D* (immagine a colori unitamente all'informazione di profondità) e solitamente sono rappresentati come una collezione di coordinate tridimensionali chiamate nuvole di punti o *point clouds*.

Negli ultimi anni, gli algoritmi basati su apprendimento automatico, Deep-Machine Learning, hanno ottenuto prestazioni eccezionali in diversi task di Computer Vision, soprattutto per quanto riguarda l'elaborazione di immagini con *reti neurali convoluzionali* (CNN). Diversamente dalle immagini, le nuvole di punti non contengono una struttura spaziale regolare, rendendo difficile la loro elaborazione per mezzo di standard CNNs. Al fine di superare questa problematica, nella prima parte del percorso previsto da questo assegno di ricerca si vogliono esplorare due recenti tecniche introdotte nel campo dell'intelligenza artificiale, i *Transformer* [1] e le *Implicit Neural Representations* (INR) [2,3,4] per poter progettare algoritmi in grado di eseguire task come la classificazione di superfici [5] o la loro segmentazione semantica [6].

I Transformer sono stati introdotti in [1] come un nuovo meccanismo per l'analisi del linguaggio naturale (NLP). Uno dei principali vantaggi di questo modello è la capacità di elaborare in modo parallelo elementi di una stessa sequenza dati, ad esempio le parole di una frase, aggregando le relazioni tra i vari campioni attraverso un meccanismo denominato *self-attention*. Rappresentando una nuvola di punti come una sequenza di coordinate in un sistema di riferimento cartesiano, tale paradigma può essere sfruttato per individuare le relazioni semantiche tra i punti all'interno di una superficie.

Le Implicit Neural Representations costituiscono una nuova modalità per rappresentare i segnali. I segnali vengono convenzionalmente rappresentati in modo discreto, ad esempio le immagini sono una griglia di pixel, diversamente le INR parametrizzano un segnale come una funzione continua descrivendolo tramite una rete neurale. Codificando una point cloud attraverso una INR è possibile usare tale rappresentazione come segnale in ingresso di una rete neurale senza dover ricorrere ad alcun tipo di meccanismo per la gestione dell'irregolarità del segnale [8]. Successivamente, il concetto di INR può essere esteso per navigare lo spazio dei dati multimodali e progettare un meccanismo basato su *Multimodal Learning* [14] per la traduzione tra segnali acquisiti con modalità diversa ma rappresentanti lo stesso concetto. Per esempio, si potrebbe pensare di tradurre un'immagine raffigurante una sedia nella forma tridimensionale che rappresenta quella sedia o nel segnale audio che codifica la parola sedia.

Nella seconda parte di questo assegno di ricerca, si prevede l'esplorazione di tecniche per l'analisi di dati di profondità acquisiti non da un sensore, ma bensì prodotti da una rete neurale. In passato la diffusione nel mercato di dispositivi a basso costo per l'acquisizione di dati RGB-D, come il Kinect V1 [7] o l'Intel

RealSense, ha alimentato un'intesa attività di ricerca nel campo della 3D Computer Vision, soprattutto in task di basso livello come la stima della trasformazione rigida per allineare due point cloud acquisite da punti di vista diversi in un unico sistema di riferimento (*Surface Registration*). Grazie ai progressi fatti dall'intelligenza artificiale, ad oggi è possibile stimare l'informazione di profondità direttamente da una singola immagine monoculare RGB[9,10] senza ricorrere all'utilizzo di un sensore. Tuttavia, pochi sono i metodi che utilizzano tali dati in applicazioni di Computer Vision. Vista la proliferazione di nuovi algoritmi per dati RGB-D [11,12,13], un'interessante direzione di ricerca riguarda lo sviluppo di algoritmi per Surface Registration operanti su dati di tipo RGB-D Learned, ovvero ottenuti stimando l'informazione di profondità attraverso una rete neurale. Questo nuovo filone consentirebbe l'utilizzo delle innumerevoli collezioni di immagini disponibili, sopperendo così alla mancanza di insieme di dati per 3D Computer Vision. Un primo passo in questa direzione potrebbe essere quello utilizzare alcune tecniche di Surface Registration basate su apprendimento automatico addestrate su dati reali RGB-D e progettare tecniche di *domain adaptation* per operare su dati RGB-D Learned.

References

1. Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

2. Park, Jeong Joon, et al. "DeepSDF: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

3. Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

4. Chen, Zhiqin, and Hao Zhang. "Learning implicit fields for generative shape modeling." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

5. Wu, Zhirong, et al. "3d shapenets: A deep representation for volumetric shapes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

6. Yi, Li, et al. "A scalable active framework for region annotation in 3d shape collections." *ACM Transactions on Graphics (ToG)* 35.6 (2016): 1-12.

7. Newcombe, Richard A., et al. "Kinectfusion: Real-time dense surface mapping and tracking." *2011 10th IEEE international symposium on mixed and augmented reality*. IEEE, 2011.

8. Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

9. Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

10. Lasinger, Katrin, et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." *arXiv preprint arXiv:1907.01341* (2019).

11. Banani, Mohamed El, Luya Gao, and Justin Johnson. "UnsupervisedR&R: Unsupervised Point Cloud Registration via Differentiable Rendering." *arXiv preprint arXiv:2102.11870* (2021).

12. Yang, Heng, et al. "Self-supervised Geometric Perception." *arXiv preprint arXiv:2103.03114* (2021).

13. Pham, Quang-Hieu, et al. "LCD: learned cross-domain descriptors for 2D-3D matching." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 07. 2020.

14. Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018): 423-443.