SP3D: Learning Keypoint Detection and Description in 3D Medical Images

N. Loiseau–Witon[1,2], R. Kéchichian[1], S. Valette[1] and A. Bartoli[2]

[1] CREATIS; CNRS (UMR 5220); INSERM (U1206); INSA Lyon; Université de Lyon, France

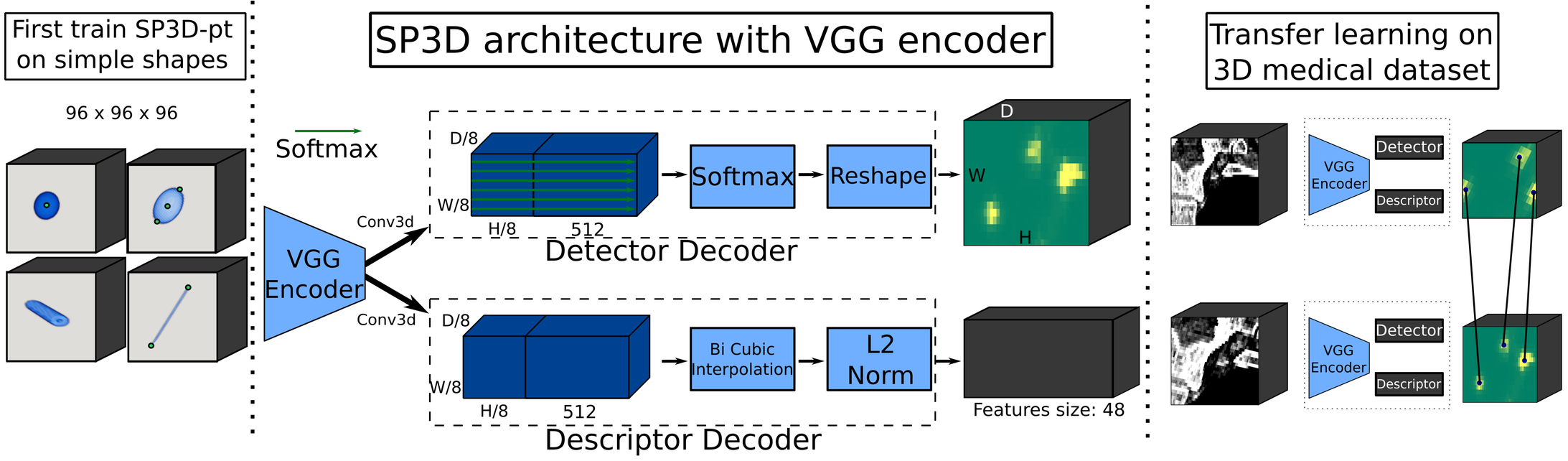
[2] Institut Pascal, UMR 6602 CNRS/UCA/CHU, Clermont-Ferrand, France

Introduction and Objectives

Computational anatomy aims to quantify anatomical shape variations, particularly in scenarios like tumor evolution over time. We primarily focus on sparse registration, which hinges on the extraction of well-distributed, repeatable keypoints with discriminative descriptors [2, 1].

We focus on two main objectives :

1. Establishing a training program for a 3D key point detection and description network.
2. Assess this network using various metrics in order to facilitate comparisons with different established manual methods.

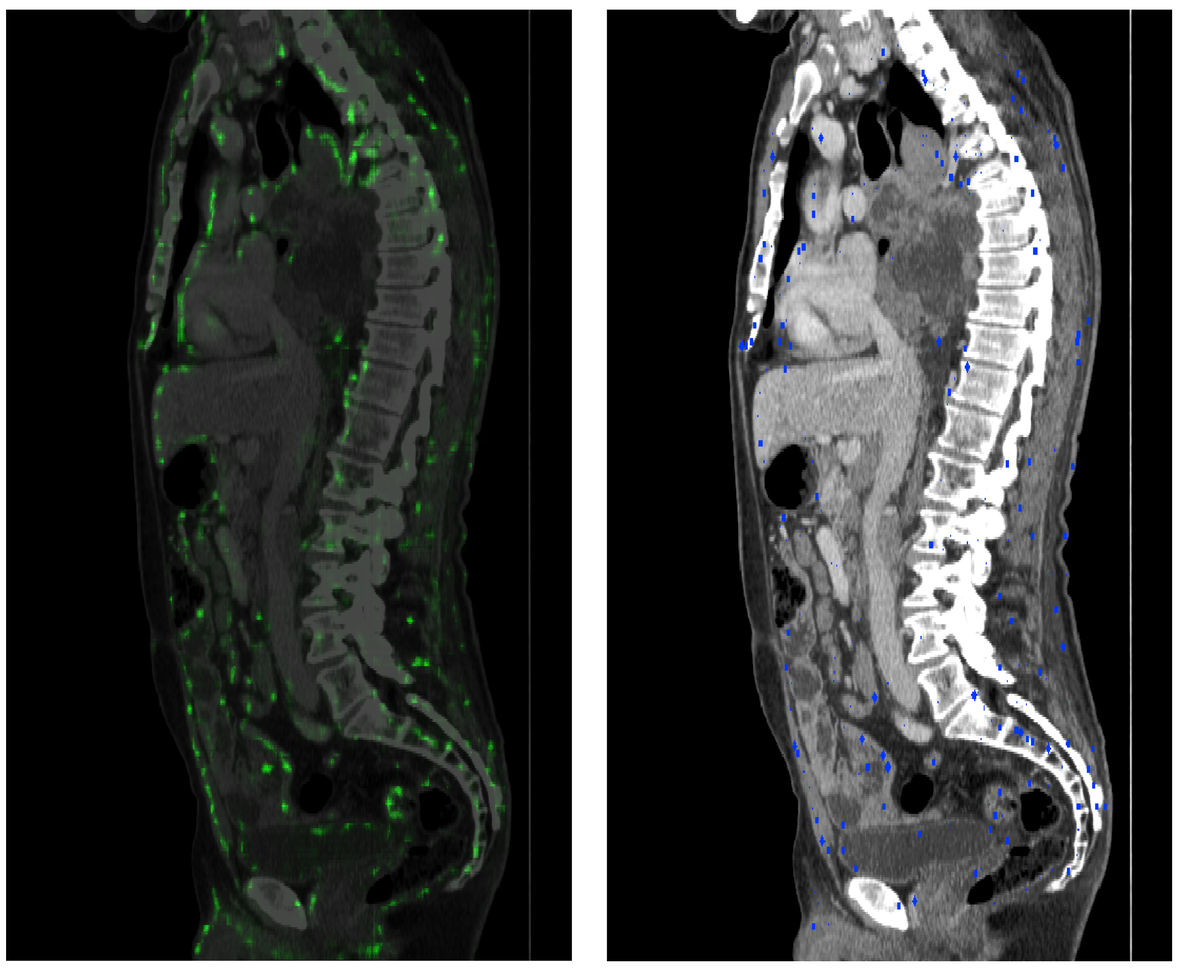


Overview of the SP3D network that jointly extracts repeatable keypoints and their discriminant descriptors. The network begins with a VGG-style encoder whose output is conjointly utilized by the keypoint detector and descriptor. The reshape phase following the Softmax operation ensures an output volume of similar size to the input for the detection part, and the ground truth volume is only utilized during the training phase. The descriptor volume size is half that of the volume used for keypoint description.

Method : CNN Architecture

SP3D use a VGG-style encoder illustrated in figure . The VGG encoder uses eight convolutional layers, and the dimensions of the input data are each reduced by a factor of 8 while obtaining 513 channels. Among these channels, the final one serves as a dustbin channel, discarding non-interest points. Following the application of the Softmax function, the last channel is removed, and an automatic reshaping technique called SubPixelConvolution [3] (referred to as PixelUnShuffle in the PyTorch library) is applied.

Method : Training



Cross-section of keypoint detection example. Left: detector output, highlighting regions with strong responses in green. Right : same cross-section with keypoints extracted from the response map.

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}'_{det} + \lambda \mathcal{L}_{desc}$$

With $\mathcal{L}_{det}$ a cross-entropy loss, which is applied between the generated image and the ground-truth labels. $\mathcal{L}_{desc}$ comprises two hinge loss terms. The first hinge loss term works towards clustering descriptors associated with the same location, while the second term is aimed at distinguishing descriptors from disparate locations.

Results

Method	Repeatability (2 mm)	Matching Score	MDBL
3D-SIFT	0.37	0.16	12.2 mm
3D-SURF	0.46	0.34	8.20 mm
SP3D-pt	0.20	0.16	16.6 mm
SP3D	**0.51**	**0.48**	**7.98 mm**

Performance comparison between 3DSURF, 3DSIFT and our SP3D approach. We used MDBL for Mean Distance Between Landmarks.

Conclusion and Future Work

Our results show that a learned 3D detector and descriptor can outperform handcrafted methods, namely 3DSIFT and 3DSURF. Following these results, we have two new objectives :

- ➤ Train the network using the repeatability metric directly
- ➤ Train a 3D keypoint detector and descriptor using the distant supervised learning approach

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[2] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004.

[3] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CoRR*, abs/1609.05158, 2016.

This work was funded by the project TOPACS ANR-19- CE45-0015 from the National Research Agency (ANR).