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]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne,
CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, LYON, France
[2]Institut Pascal,
UMR 6602 CNRS/UCA/CHU, Clermont-Ferrand, France
lastname@creatis.insa-lyon.fr Adrien.Bartoli@gmail.com

Abstract

A large body of methods exist for keypoint detection and description in 2D images, but only the handcrafted SURF3D and SIFT3D methods handle 3D images. We propose to extend SuperPoint, a learning-based method for 2D images, to form a novel end-to-end method we call SP3D, dedicated to medical 3D images. We create a 3D convolutional neural architecture which implements the full procedure of keypoint detection and description for surrounding 3D patches. We propose to train SP3D in two steps with transfer learning. The first step uses predefined keypoints in a synthetic dataset of simple 3D shapes. The second step uses a semi-synthetic dataset of warped CT volumes; the keypoints are detected via SP3D and by SURF3D, and kept only if they are sufficiently repeated across the warped volumes. Experimental results on synthetic data and registration comparison on real data show the superiority of SP3D.

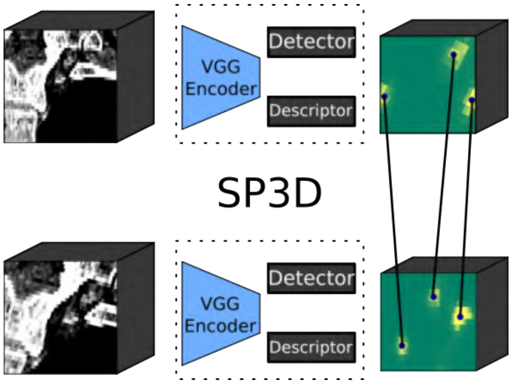

Figure 1. We propose SP3D, a 3D extension of SuperPoint, a fully-convolutional neural network. SP3D computes keypoint locations and descriptors in 3D medical images, in a single forward pass. This figure shows the objective of SP3D training, which is to identify matching keypoints across two 3D images. The input volume contains 96^3 voxels.

1. Introduction

The general goal of computational anatomy is to quantify the variability of anatomical shapes, such as the shape of a tumor in time. An important tool required to carry out this task is medical volume registration. Registration can be estimated densely by considering all voxels of the source and target volumes or sparsely by considering keypoints and their descriptors. We focus on sparse registration, which crucially depends on keypoint extraction and local descriptor computation. Keypoints must be well distributed, repeatable and their descriptors must be discriminant [8, 3]. A keypoint is repeatable if it is detected in both volumes at corresponding anatomical locations. The descriptor is discriminant if it uniquely characterizes each keypoint.

Learning-based methods for keypoint detection and description with Convolutional Neural Network (CNN) have recently outperformed handcrafted methods in 2D images [9, 4, 5], because they can be explicitly trained to generate repeatable and discriminant keypoints. The handcrafted methods SIFT and SURF [3, 8] were extended to volumes [1, 10]; however, a learning-based approach for 3D images has yet not been attempted.

We propose SP3D, illustrated in figure 1, a 3D extension of the 2D learning-based method SuperPoint [4]. We face three main tasks. First, to define an architecture which handles large volume data. Second, to create datasets to train SP3D. Third, to train SP3D appropriately. For the second task, we propose to generate two datasets. The first dataset is used to pre-train the SP3D network with exact positions of keypoints. We will refer to the pre-trained SP3D network as SP3D-pt. This phase is intended as an initial SP3D training phase, designed to transfer the weights learned to the subsequent training phase using our medical dataset.

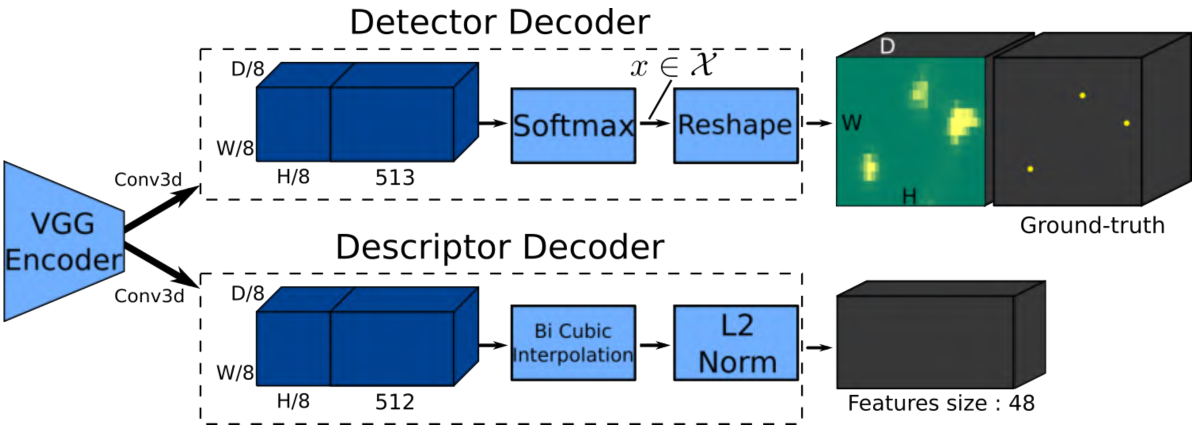

Figure 2. Overview of the SP3D network that jointly extracts repeatable keypoints and their discriminant descriptors. The network begins with a VGG-style encoder whose output is conjointly utilized by the keypoint detector and descriptor. The reshape phase following the Softmax operation ensures an output 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.

The first dataset is composed of generated 3D shapes such as lines, cubes and spheres. The second dataset is used to train SP3D with our data using transfer learning starting from SP3D-pt weights. The second dataset is created using 3D-SURF [1] and the detector from the pre-training stage. The second dataset is semi-synthetic. We create it by transforming real volumes, extracting the most repeatable keypoints and establishing keypoint correspondences between the multiple transformed volumes. Our experimental results show that our extension of the SuperPoint method outperforms hand-crafted methods such as 3D-SURF and 3D-SIFT [1, 10] in terms of repeatability and mean distance between landmarks after registration, with similar run-time.

2. Network Architecture and Losses

SP3D-pt and SP3D use the same architecture, which is a VGG-style [12] encoder illustrated in figure 2. SP3D uses a combination of two losses for the training step, one for the keypoint detection, and the second for keypoint description.

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 [11] (referred to as PixelUnShuffle in the PyTorch library) is applied. We set $H_c = H/8$, $W_c = W/8$ and $D_c = D/8$, where H, W, D are the dimensions of the input volume.

During training, as shown in figure 1, SP3D is applied to two input volumes: V and V^τ. The latter V^τ is the result of applying an affine transform to the former, V. The detector loss is computed independently for both volumes, while the descriptor loss is computed taking into account both network outputs.

2.1. Detector Loss

The loss function for the keypoint detector is a cross-entropy loss applied between $x_{hwd} \in \mathcal{X}$ and the ground-truth labels $y_{hwd} \in \mathcal{Y}$ properly reshaped. The final loss is :

$$\mathcal{L}_p(\mathcal{X}, \mathcal{Y}) = \frac{1}{H_c W_c D_c} \sum_{h=1,w=1,d=1}^{H_c,W_c,D_c} l_p(\mathrm{x}_{hwd}; y_{hwd}), \quad (1)$$

where $\mathrm{x}_{hwd} \in \mathcal{X}$ corresponds to each vector across the channels in the output of the decoder, $\mathcal{Y}$ represents the ground truth provided to the network and l_p is the Negative Log-Likelihood (NLL) of softmax.

2.2. Descriptor Loss

The descriptor loss is composed of two hinge loss terms, the first hinge loss tends to bring together descriptors which correspond to the same location and the second term serves to discriminate descriptors from different locations. The descriptor loss is computed for all descriptor pairs. For a given pair of descriptors d and d' from a pair of volumes V and V^τ, it is defined as:

$$l_d(d, d') = \Big(\lambda_d \; s \; \max(0, m_p - d^T d')\Big) + \Big((1-s) \; \max(0, d^T d' - m_n)\Big), \quad (2)$$

where m_p is the positive margin and m_n is the negative margin, d' corresponds to the descriptor at position $h'w'd'$, d corresponds to the descriptor at position hwd and d^T is its transposed version. Similarly to SuperPoint, we use the term λ_d to mitigate the fact that the network will more often encounter incorrect (negative) correspondences than correct (positive) correspondences. In equation (2), s computes correspondences between each pair of descriptors in input volumes V and V^{τ} and can be written as follows:

$$s_{hwd,h'w'd'} = \begin{cases} 1 & ||p_{hwd}{}^{\tau} - p_{h'w'd'}|| \leq \delta \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In equation (3), p_{hwd} represents the keypoint location, $p_{hwd}{}^{\tau}$ denotes the coordinate transformation of the point by the affine matrix of transformation τ, δ corresponds to the maximal allowed distance between the centers of two keypoints and is expressed in voxels. Empirically we set it to $\delta = 8$.

3. Training

We train SP3D-pt on a synthetic generated dataset composed of simple 3D shapes with exact keypoint locations. Examples of volumes can be seen in figure 3. To generate this dataset, we examined our medical images and constructed shapes based on content analysis. For instance, we aimed to extract keypoints within blobs or spherical structures such as small muscles, or at the extremities of a vessel-like structures. However, we avoided placing keypoints inside vessels since their precise detection locations are uncertain. To simulate a random background, we generated multiple spheres and added Gaussian noise to the images.

For the second training step, we used Silver and Gold subsets from the Visceral dataset [6], to build a semi-synthetic dataset. The first subset, named Gold, contains 20 CT volumes, each annotated with about 40 landmarks. The second subset, Silver, contains 60 CT volumes without anatomical landmarks. Using 3D-SURF and SP3D-pt, we can detect keypoints P and P^{τ} respectively from V and V^{τ}, with V^{τ} the volume V transformed by τ. Using the inverse transform τ^{-1} we warp points P^{τ} to P, and with a threshold empirically found on distances between each set of points, we obtain the best set of corresponding points. The weights of SP3D-pt are fine-tuned using this second dataset. Unlike the original SuperPoint approach, where the detector was trained only on 2D shapes during the initial training and the weights were combined with the addition of the descriptor part for the second training, we conducted fine-tuning on all parameters of both the detector and descriptor networks in both training stages.

For both datasets, the transforms τ are generated with the same method as [7], using landmarks from the Gold group to estimate the distribution of transformations.

We note that in the second training step volumes are too large to be used in a single mini-batch due to memory limitations, with each volume having an approximate size of $512 \times 512 \times 400$ or larger. Training SP3D on such volumes, unlike in the 2D case, is not feasible. To avoid this problem, we train SP3D using sub-volumes extracted from the original volume. We use a patch size of 96^3 for every image or sub-volume image, and simultaneously consider one patch along with its corresponding transformed version. Each sub-volume is paired with its corresponding ground truth.

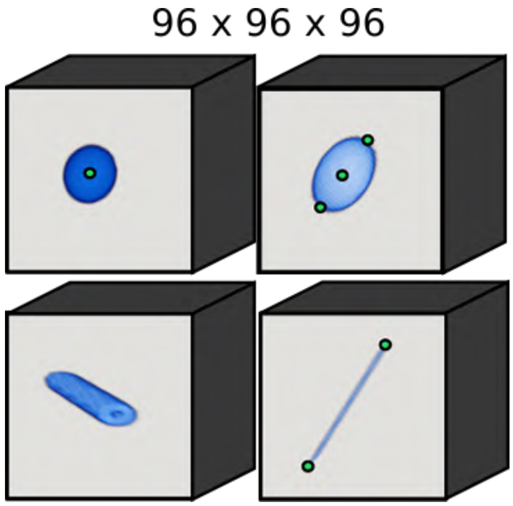

Figure 3. The initial input volumes are 3D shapes that were generated with precise keypoint locations and contain 96^3 voxels. In order to enhance the visualization of shapes, we removed the background.

4. Training Details, Datasets and Metrics

4.1. Datasets

We first learn SP3D-pt with 3D shapes; the data consists of 24000 volumes for training and 2400 volumes for validation. The volumes are randomly generated from 8 different shape families, each shape being automatically annotated with a list of ground-truth keypoints, see figure 3. For the second learning step, we use 100 subjects for the training step, 20 for the validation and 20 from the Gold subset, 20000 keypoints have been extracted for each of these subjects.

4.2. Training Details and Hyper-parameters

We optimize both SP3D-pt and SP3D networks via ADAM with a learning rate of 0.001 and a mini-batch size of 1 patch. To compensate the mini-batch size, we use gradient accumulation during training, we accumulate 8 mini-batches before the learning step. Our GPU-based implementation uses the PyTorch library. The training for a single epoch takes around 90 minutes and approximately 20 Go of memory running on a Linux 64-bit platform, with the utilization of NVIDIA V100 GPU.

4.3. Repeatability, Matching Score and Registration

We evaluate the network using three different metrics. The first metric is the repeatability between keypoints. After keypoint extraction, we obtain two sets of points P and P^{τ} from two volumes V and V^{τ}, V^{τ} being the volume V transformed by τ. Using τ^{-1} we transform points P^{τ} to P. With a simple k-d Tree, we compute distances between P and $P^{\tau^{-1}}$. For a distance lower than a threshold between two points, we consider that these two points are repeated.

The second metric is the matching between descriptors, which measure how much the network can discriminate descriptors. For random keypoints extracted from two images, we extract corresponding descriptors. Similarly to repeatability, we use a k-d Tree to measure distance between each descriptor to find closest points. We consider these two descriptors as a match when the distance between them is less than the first descriptor and all other descriptors.

The third metric is the Mean Distance Between Landmark (MDBL) computed on ground-truth landmarks in Gold volumes after registering them to a common space using the FROG registration algorithm [2]. Low mean distance between landmarks indicates good results, contrary to the repeatability and matching scores. The repeatability and matching score are measured in intra-patient, instead of the MDBL metric which can measure the performances in inter-patient volumes.

The third metric assesses the complete end-to-end pipeline, encompassing both detection and description, while also accounting for the registration method. On the other hand, the first two metrics, repeatability and descriptor matching, solely focus on their respective aspects without considering any additional factors and are only applicable to intra-patient measurement.

5. Results and Observations

We evaluate the performance of the pre-trained network SP3D-pt on shapes and patient volumes from the Gold group. SP3D-pt achieves excellent results on 3D shapes, with a repeatability of 0.68 and a descriptor matching score of 0.61. However, these two metrics experience a significant drop in performance when applied to the medical image dataset. The repeatability score is only 0.2, and the matching score falls to 0.16. These results justify the transfer learning approach with SP3D on the medical dataset, whose results are given in table 1. We can see that SP3D outperforms 3D-SIFT and 3D-SURF on all metrics.

The left part of figure 4 displays the response keypoint map extracted using SP3D, while the right part shows the keypoints extracted from this response map. There is a higher rate of keypoint detection on bone tissues compared to soft tissues such as organs and muscles.

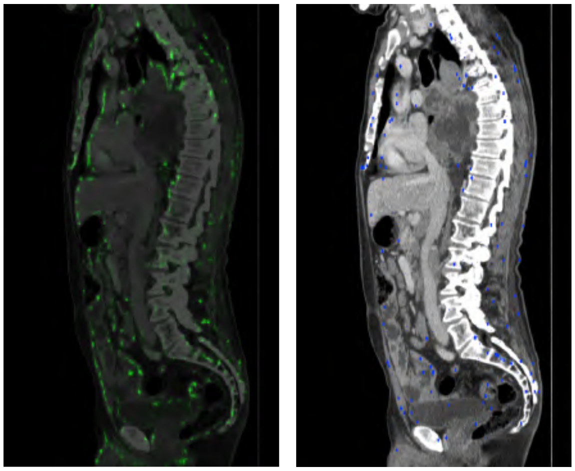

Figure 4. Cross-section of keypoint detection example. Left: detector output, highlighting regions with strong responses in green. Right : same cross-section with keypoints extracted from the response map.

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**

Table 1. Performance comparison between 3DSURF, 3DSIFT and our SP3D approach. We used MDBL for Mean Distance Between Landmarks.

6. Conclusion

Our results show that a learned 3D detector and descriptor can outperform handcrafted methods, namely 3DSIFT and 3DSURF.

Our next objective is to train the network using the repeatability metric directly, as we consider it the most important of keypoint characteristics extracted from medical images. In the realm of 2D distant supervised keypoint extraction, numerous methods have demonstrated superior results. Distant supervised learning refers to a technique where no ground-truth data is available, except for the knowledge of transformations between images. Future research endeavors will focus on training a 3D keypoint detector and descriptor using the distant supervised learning approach.

Acknowledgments

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

References

[1] Rémi Agier, Sébastien Valette, Laurent Fanton, Pierre Croisille, and Rémy Prost. Hubless 3d medical image bundle registration. In *VISAPP 2016 11th Joint Conference*, 2016.

[2] Rémi Agier, Sébastien Valette, Razmig Kéchichian, Laurent Fanton, and Rémy Prost. Hubless keypoint-based 3d deformable groupwise registration. *Medical image analysis*, 59:101564, 2020.

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

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CoRR*, abs/1712.07629, 2017.

[5] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint detection and description of local features. *CoRR*, abs/1905.03561, 2019.

[6] Georg Langs, Allan Hanbury, Bjoern Menze, and Henning Müller. Visceral: Towards large data in medical imaging — challenges and directions. In Hayit Greenspan, Henning Müller, and Tanveer Syeda-Mahmood, editors, *Medical Content-Based Retrieval for Clinical Decision Support*, pages 92–98. Springer Berlin Heidelberg, 2013.

[7] Nicolas Loiseau-witon, Razmig Kéchichian, Sebastien Valette, and Adrien Bartoli. Learning 3D medical image keypoint descriptors with the triplet loss. *International Journal of Computer Assisted Radiology and Surgery*, 2021.

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

[9] Jérôme Revaud, Philippe Weinzaepfel, César Roberto de Souza, Noé Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. *CoRR*, 2019.

[10] Blaine Rister, Mark Horowitz, and D. Rubin. Volumetric image registration from invariant keypoints. *IEEE Transactions on Image Processing*, 26:4900–4910, 2017.

[11] 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.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.