Learning 3D Medical Image Patch Descriptors with the Triplet Loss

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
lastname@creatis.insa-lyon.fr

[2] Institut Pascal, UMR 6602 CNRS/UCA/CHU, Clermont-Ferrand, France
Adrien.Bartoli@gmail.com

Keywords: Medical imaging · Triplet loss · Convolution Neural Network

1 Purpose

Computational anatomy focuses on the analysis of the human anatomical variability. Typical applications are the discovery of differences across healthy and sick subjects and the classification of anomalies. A fundamental tool in computational anatomy, which forms the central focus of this paper, is the computation of point correspondences across volumes (3D images) such as Computed Tomography (CT) volumes, for multiple subjects. More specifically, we consider automatically detected keypoints and their local descriptors, computed from the image or volume patch surrounding each keypoint. Theses descriptors are essential because they must be discriminant and repeatable [5,10]. Learned descriptors based on Convolutional Neural Networks (CNN) have recently shown great success for 2D images [4]. However, while classical 2D image descriptors were extended to volumes [8], recent CNN based approaches were not. We propose a methodology to construct these learned volume keypoint descriptors. The main difficulty is to define a sound training approach, combining a training dataset and a loss function. In short, we propose to generate semi-synthetic data by transforming real volumes and to use a triplet loss inspired by 2D descriptor learning. Our experimental results show that our learned descriptor outperforms the hand-crafted descriptor 3D-SURF [1], a 3D extension of SURF, with similar runtime.

2 Methods

Our first goal is to create a reference dataset defining keypoint correspondences between multiple volumes. In 2D, these correspondences can be established using Structure-from-Motion [3]. In 3D medical images, keypoints could be defined as anatomical landmarks placed by medical experts. However, no such large annotated dataset is publicly available. We thus propose to create a semi-synthetic reference dataset by transforming real volumes.

2.1 Constructing a semi-synthetic dataset

We use two subsets from the Visceral dataset [9]. The first subset, named *Gold*, contains 20 CT volumes, each annotated with about 40 landmarks. The second subset, *Silver*, contains 60 CT volumes without landmarks.

In order to generate new volumes, we estimate the probability density of possible inter-volume transformations, and sample from this density to warp CT volumes and define keypoint correspondences. We first compute inter-volume local affine transformations in the least-squares sense. For each landmark in each volume of the *Gold* subset, we use the landmark, its three closest landmarks and the four corresponding landmarks in another volume to estimate a local transformation. When the landmark and its three neighbours are almost collinear (e.g. vertebral landmarks), the least-squares problem is ill-conditioned and we therefore discard the transformation. Thus we obtain at most $Lk(k-1)/2$ affine transformation matrices of size 4×4, where $L = 40$ is the number of landmarks and $k = 20$ is the number of volumes. We apply Kernel Density Estimation (KDE) to these matrices to estimate the density of inter-volume transformations. This KDE uses a Gaussian Kernel and Scott's rules to estimate the bandwith. To generate our semi-synthetic dataset, we sample transformations from this density and apply them to volumes in the *Silver* subset. More specifically, for a sampled transformation t and a silver volume V_i, we obtain the volume V_i^t. We detect the keypoints in V_i and V_i^t using 3D-SURF and obtain two keypoint sets P_i and P_i^t. Note that 3D-SURF is both a detector and a descriptor. We then apply the inverse transform t^{-1} to the keypoints from V_i^t. Finally we use a k-d tree to construct the set of corresponding keypoints between the volumes as the set of pairs: $(p \in P_i, q \in t^{-1}(P_i^t))$, whose p to q distance is lower than 8 mm. This threshold was chosen to obtain a large number of correct correspondences.

2.2 Training the descriptor with the triplet loss

We learn a descriptor CNN mapping a 3D patch of 10^3 voxels surrounding a keypoint to a descriptor vector. Recent work in 2D has shown that learning descriptors using triplets yields better results than using pairs [7]. Triplet learning requires forming triplets of patches $\{a, p, n\}$ where a is an anchor, p a positive representing a different patch of the same class as a, and n a negative representing a patch of a different class. In our case a and p are two patches around corresponding keypoints from different volumes and n is a patch around a different keypoint. The aim is to optimize the CNN parameters in order to bring a and p close together in descriptor space and to push n away from a. Thus the triplet loss is defined by $\mathcal{L}(a,p,n) = \max(\|f(a)-f(p)\|^2 - \|f(a)-f(n)\|^2 + \alpha, 0)$, where $f(\cdot)$ is the CNN and α the margin parameter.

Our CNN is defined by two 3D convolution layers, one maxpooling layer between the convolutions, and a fully-connected layer which gives the final descriptor. The network architecture is illustrated by Figure 1. Passing a patch of size 10^3 through this CNN gives us a descriptor vector of the desired size. We use a size of 48 for direct comparability with 3D-SURF.

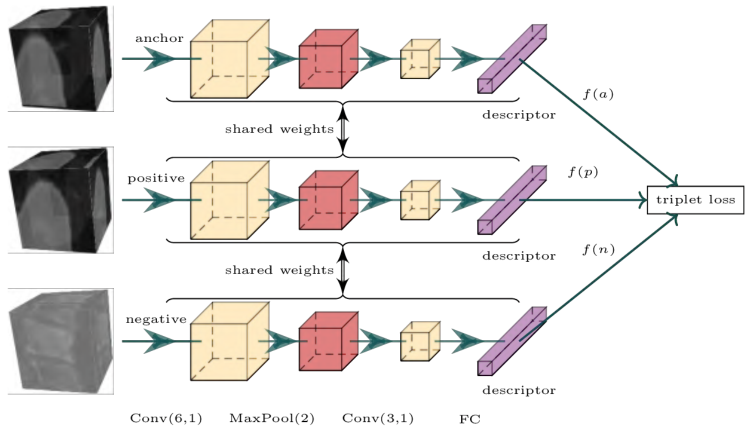

Fig. 1. Architecture of our CNN and the triplet-loss. The input patches are 10^3 voxels. The CNN has two convolutions with *tanh* activation, one maxpooling and a fully-connected layer. The triplet $\{a, p, n\}$ is passed through the network, and descriptor vectors are fed to the triplet loss.

3 Results

Data split. We divide the *Silver* dataset into training and validation subsets. The training data consist of 55 subjects with 10 transformed volumes each, following our procedure of semi-synthetic data generation. The validation data consist of 5 subjects with 10 transformed volumes each. For testing, we use the *Gold* subset with 20 subjects and the associated anatomical landmarks.

Training. Optimization is performed via Stochastic Gradient Descent, with a batch size of 1000 patches, a learning rate of 0.1, a momentum of 0.9, a weight decay of 10^{-6} and a loss margin of 0.2. We also use online triplet mining to find the best triplets for learning. Our CPU-based implementation uses the PyTorch library. The training of a single epoch with 10^6 triplets takes about 30 minutes and approximately 10 GB of memory on a Linux 64-bit platform running on an Intel Xeon 2.6 GHz CPU. Our model is light enough to be trained on the CPU; using the GPU did not significantly reduce training time.

Evaluation. We evaluate the descriptor using two different metrics. The first metric is the false positive rate at point 0.95 of true positive recall (FPR95) [6]. We compute FPR95 based on descriptor distances of randomly selected 10^5 keypoint pairs with 50% corresponding and 50% non corresponding pairs. A low FPR95 indicates good results. The second metric is the mean landmark distance calculated on ground-truth landmarks in *Gold* volumes after registering them to a common space using the keypoint-based FROG registration algorithm [2]. For comparison, we replace the 3D-SURF descriptor used in this algorithm with our learned descriptor. Low mean distance between landmarks indicates good results. Table 1 shows that the proposed descriptor yields better results in terms of both FPR95 and mean landmark distance compared to the 3D-SURF descriptor.

Type of descriptor	FPR95	Mean landmark distance
3D-SURF	0.077	8.74
Learned	0.022	8.54

Table 1. Performance comparison of the 3D-SURF and our learned descriptors.

4 Conclusions and future work

Our results, although preliminary, show that a learned 3D descriptor, trained on semi-synthetic data, can outperform a carefully hand-crafted one. We intend to further explore these promising results by extending our training dataset and conducting more experiments. Future research will address training a 3D keypoint detector.

5 Acknowledgements

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

References

1. Agier, R., Valette, S., Fanton, L., Croisille, P., Prost, R.: Hubless 3d medical image bundle registration. In: VISAPP 2016 11th Joint Conference (2016)
2. Agier, R., Valette, S., Kéchichian, R., Fanton, L., Prost, R.: Hubless keypoint-based 3d deformable groupwise registration. Medical image analysis **59**, 101564 (2020)
3. Altwaijry, H., Veit, A., Belongie, S.J., Tech, C.: Learning to detect and match keypoints with deep architectures. In: BMVC (2016)
4. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Bmvc. vol. 1, p. 3 (2016)
5. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417. Springer (2006)
6. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 43–57 (2010)
7. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition. pp. 84–92. Springer (2015)
8. Kéchichian, R., Valette, S., Desvignes, M.: Automatic multiorgan segmentation via multiscale registration and graph cut. IEEE Transactions on Medical Imaging **37**(12), 2739–2749 (2018)
9. Langs, G., Hanbury, A., Menze, B., Müller, H.: Visceral: towards large data in medical imaging—challenges and directions. In: MICCAI international workshop on medical content-based retrieval for clinical decision support. pp. 92–98. Springer (2012)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)