Full Orientation Invariance and Improved Feature Selectivity of 3D SIFT with Application to Medical Image Analysis

Stéphane Allaire[1,3], John J. Kim[1,2], Stephen L. Breen[1,2], David A. Jaffray[1,3], Vladimir Pekar[4]
[1]Radiation Medicine Program, Princess Margaret Hospital, Toronto, ON, Canada
[2]Department of Radiation Oncology, University of Toronto, ON, Canada
[3]Department of Medical Biophysics, University of Toronto, ON, Canada
[4]Philips Research North America, Markham, ON, Canada
allaire.stephane@gmail.com

Abstract

This paper presents a comprehensive extension of the Scale Invariant Feature Transform (SIFT), originally introduced in 2D, to volumetric images. While tackling the significant computational efforts required by such multi-scale processing of large data volumes, our implementation addresses two important mathematical issues related to the 2D-to-3D extension. It includes efficient steps to filter out extracted point candidates that have low contrast or are poorly localized along edges or ridges. In addition, it achieves, for the first time, full 3D orientation invariance of the descriptors, which is essential for 3D feature matching. An application of this technique is demonstrated to the feature-based automated registration and segmentation of clinical datasets in the context of radiation therapy.

1. Introduction and related work

The automatic extraction of salient interest points in different images and their matching has many important applications in image processing. It can be used, for example, for feature-based image registration [1, 11], object recognition [9], image segmentation, atlas generation and variability analysis [14], and image retrieval in databases. The Scale Invariant Feature Transform (SIFT) is an approach which has grown very popular in the last decade, since its original introduction by D.G. Lowe [9, 10]. It focuses on extracting salient interest points that are stable, and that can be represented by feature descriptors in the most invariant way with respect to scaling, translation, orientation, affine changes, and illumination within the images. Even though this transform was designed for and tested on 2D images of 3D objects, mathematical theory does not prevent its extension to higher dimensions. To our knowledge, only two multi-dimensional SIFT approaches published to date have been proposed. – W.A. Cheung and G. Hamarneh [2, 3], have generalized the scale space principle, which is the support base for scale invariance. Hyperspherical coordinates are subsequently used for image gradients, and multidimensional histograms contribute to the descriptors. The method has been tested on 3D magnetic resonance (MR) images of the brain, using different contrasts (Proton Density, T1 and T2 weighting), and a 4D computed tomography (CT) image series of a beating heart (i.e. 3D + time); – P. Scovanner *et al*. [12], have extended to 3D and exploited only the descriptor side to SIFT, while dropping its scale invariance and saliency appeals, since descriptors are computed online at randomly chosen interest points, and then clustered in a bag-of-word paradigm to classify actions in video sequences (i.e. 2D + time). The focus of both these recent studies has been on feature descriptor representation and matching / classification. The not reoriented n-SIFT feature is deemed satisfactory in [2]; a 3D reoriented SIFT descriptor is proposed in [12]. Yet the rotational invariance achieved is only partial.

In this work, we address the two issues not tackled so far: a) blob filtering for location stability, and b) canonical 3-angle orientation assignment for full invariance. This leads us to solve the two mathematical problems occurring when fully extending SIFT to 3D. This in turn provides distinctive features, which can thus be reliably extracted and matched. In the context of medical image analysis, salient points are sharply prominent anatomical features that remarkably protrude from their surroundings. Note that this does not exclude edges or ridges, e.g. in the scapula bones in the shoulder. The issue with edges and ridges is that they cannot be reliably distinguished from similar neighboring areas nor attached to a localized point. Another example is the ribs, which all look similar. For the sake of stability, it is advantageous to focus more specifically on blobs, which are either brighter or darker than their surroundings in all directions.

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

Moreover, blobs can be point-like, similar to corners, but they can also provide a complementary description of image structures in terms of regions. This coincides with anatomical landmarks defined by physicians: they often correspond to distinct volumes, not strictly points. The multiscale approach used in SIFT precisely targets fine as well as large blobs. This blob targeting is performed by filtering steps which are not included in [2] and which extend [10] to the 3D case.

The main issue when extending this technique to 3D, is dealing with the orientation, namely achieving the full orientation invariance of the SIFT descriptors with respect to 3 degrees of rotational freedom. Large bin sizes in the 2D SIFT histograms allow for 3D object recognition accuracy above 50% with 50° of out-of-plane rotation [10]. However, this is limited to textured planar objects, or to the outer surface appearance of 3D characters, which has virtually nothing to do with matching the volumetric content of datasets reconstructed by medical imaging modalities. The orientation invariance is especially important if one wants to match medical images with different acquisition setups, or images of different patients. Yet, already with a constrained setup, allowing for only limited error margins, considerable changes to the patient's anatomy may still occur, e.g. due to weight loss in the course of radiation treatment. For instance, flex angles of the spinal cord can sometimes be up to 15°. This is already challenging for the available SIFT implementation [2, 3], where only a partial 2-angle normalization (azimuth and elevation) is performed. The tilt angle of "self"-rotation about the gradient vector is not accounted for.

To fulfill these needs, this paper highlights the two main contributions of the comprehensive 3D extension of the SIFT algorithm which we have implemented: a) improved blob selectivity and b) the complete orientation invariance of the descriptors. Note that this requires the true size of the objects to be considered; the case of isotropic images is rare which allowed the exclusive use of voxel coordinates in [2, 3].

This paper is organized in the following way: the mathematical methods are described in the next section, full orientation invariance is experimentally assessed in Section 3, and an application to clinical images is presented in Section 4.

2. Methods for full 3D SIFT extension

Extending from [7, 8], the scale space of a 3D image can be defined as a 4D function, $L(x, y, z, \sigma)$, produced from the convolution of a variable-scale Gaussian, $G(x, y, z, \sigma)$, with an input image $I(x, y, z)$:

$$\begin{aligned} L(x, y, z, \sigma) &= G(x, y, z, \sigma) * I(x, y, z) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^3} e^{-(x^2+y^2+z^2)/2\sigma^2} * I(x, y, z). \quad (1) \end{aligned}$$

Extending from [9], it is efficient to detect stable feature point locations in the 4D scale space using extrema out of the convolution of the difference-of-Gaussian (DoG) function with the image, $D(x, y, z, k^i\sigma)$. Moreover it is reported that the extrema of such a close approximation to the scale-normalized Laplacian-of-Gaussian $\sigma^2 \bigtriangledown^2 G$ are in practice the most stable image features. The so-called DoG image is simply computed by subtracting two nearby scales separated by a constant multiplicative factor k:

$$D(x, y, z, k^i\sigma) = L(x, y, z, k^{i+1}\sigma) - L(x, y, z, k^i\sigma). \quad (2)$$

We follow the efficient approach that D.G. Lowe proposed to regularly sample the scale space: a pyramid is built with incrementally blurred versions of the original image which are grouped by octaves describing every doubling of σ in a sampling frequency of 3 increments. This allows a gradual downsampling and the associated speed-up; and incremental smoothing is indeed quicker than direct smoothing thanks to smaller kernels. The candidate feature points are then detected as local minima or maxima in the 3 DoG images, across 3D location and scale (which thus requires 2 extra scale levels). This means that a voxel is rejected as soon as 1 of its 80 neighbors proves against it being a bright spot or dark spot respectively. In practice, on medical images too, such a detection is very sensitive, but not specific enough. The additional filtering stage proposed in 2D by [10] is necessary, as well as the sub-voxel candidate location refinement. Subsection 2.3 describes in detail our efficient extension to 3D of the Hessian-based suppression of non-blob and edge-like features. Subsequently, the classical SIFT algorithm involves: –the generation of orientation-invariant features at the salient points and their description via a high-dimensional vector built from weighted histograms of neighboring gradient orientations; –a matching of features from two different images via simple reciprocal Nearest Neighbors in the descriptor space. The following subsections 2.3 and 2.4 give more details on the developed approach. The full 3D orientation invariance is explained in subsection 2.4. The other following subsections describe the application of the SIFT algorithm to medical images.

2.1. Anisotropic image resolution

It is crucial to account for anisotropic resolution of medical data. First the use of an anisotropic separable 3D Gaussian kernel for convolution is essential in this case for proper detection of candidate points. Up to 50 % of blobs

well localized in the inter-slice direction are missed otherwise. This also incidentally speeds up the building of the scale-space pyramid, which is the principal computational bottleneck. Most importantly, the image derivatives must be scaled appropriately before performing the 4D location strength test and the 3D blob selection, building histograms of gradient orientations for both feature orientation assignment and description. In other words, the true gradients and Hessian of the 3D anatomy are expressed in terms of mm, not of voxels. As well the subregions summarized by the descriptors can be defined according to the voxel size instead of being cubes of voxels.

2.2. Sampling frequency in the spatial domain

Among the many settings ruling the SIFT algorithm, be it in 2D or 3D, one important parameter to establish is the sampling frequency in the spatial domain, that is to say from what starting scale σ in Eq. 2 should the scale space be sampled. Figure 1 reports on the experimental determination of the initial blur present in the given image, and complementarily the amount of prior smoothing needed to be applied to each image level before building the scale space representation. Clearly, the lower the assumption of the initial blur – i.e. the higher the prior smoothing σ – the higher the number of feature points and associated features is detected. Meanwhile, the minimum mean time cost is obtained for an initial blur assumed at 1.1, and it would increase for a lower initial blur (and the added extracted features would look very similar to each other, showing poor reliability). Experimentally, we have found that the assumption of 1.1 provides a good trade-off for CT in terms of detection vs. localization; this value has been used to process all CT images. These results are comparable to those using the prior smoothing amount obtained in 2D in [10]. Note that for MR and cone beam CT (CBCT), we have set a higher value for this initial blur assumption (1.6) based on the same type of experiments. Therefore, less prior smoothing was required for MR and CBCT, which agrees well with a usually superior image resolution of CT. Note that for computational time issues the doubling step suggested by D.G. Lowe is skipped. Yet only the "octave 0" is not built; the processing includes "octave 1" with original image sampling rate (unlike in [3]), and up to 3 octaves depending on the volume size.

2.3. Refinement and filtering

After the detection of *scale-space* extrema, D.G. Lowe introduced a detailed interpolation and filtering stage for accurate feature point localization [10]. Faced with the huge amount of DoG extrema detected in medical images, and the obvious poor saliency and thus poor reliability of most of them, we concluded that the extension of this stage to 3D was important and necessary. This includes: i) an iterative sub-voxel 4D location interpolation, combined with

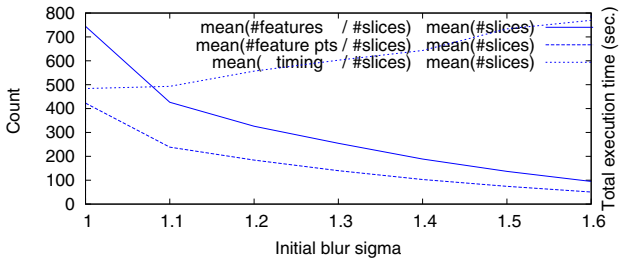

Figure 1. Experimental determination of the initial blur present in the given image. Note: the mean timing for a equivalent 167-slice full CT ranges from 218 to 346 seconds with 45 missing second graduations on the right of this graph.

an upper-thresholding on the local contrast, and a check for the duplicates. Up to 4 adjustment steps are allowed, after which the 4D location is discarded as being weak; ii) the removal of non-blob and edge-like features, by testing the relative ratios of 3D Hessian eigenvalues. Herein, we propose a special algebraic method to avoid direct computation of the eigenvalues and thus to decrease the processing time, which is less straight-forward than the 2D equivalent. An illustration of the results on a CT image is shown in Figure 2.

In addition, we have introduced an optional CT-specific upper threshold on gray values, in order to discard all features in air, background and any plastic mask fixation device. This was incorporated into the implementation as early as at the detection level, to avoid the additional computational burden of tens of thousands of features.

We refine the localization of candidate feature points and reject those that have low contrast (and are therefore sensitive to noise). The method developed by M. Brown and D.G. Lowe [1] for better matching and stability is simply extended by locally fitting the quadratic Taylor expansion of the 4D DoG function $D(x, y, z, \sigma)$, centered at the point considered. In practice, finite differences are used. We have noticed that a lower threshold of 0.03 for the function value at the extremum is suitable for CT images, whereas a less selective 0.01 adapts well to the less contrasted appearance of MR and CBCT images.

Candidates that are poorly localized along edges need to be eliminated since they are unstable in the presence of even small amounts of noise. The DoG function is not specific to blobs and has a strong response along edges [10], and also along ridges in 3D.

Blob-like structures can be characterized by the following properties: i) all principal curvatures are of the same sign, and ii) they all have a magnitude of the same order. One can expect that the first condition is unlikely to be violated due to the properties of DoG extrema. Nevertheless, in practice, a test is necessary to assure this, followed by a second test upon the second condition which is less likely

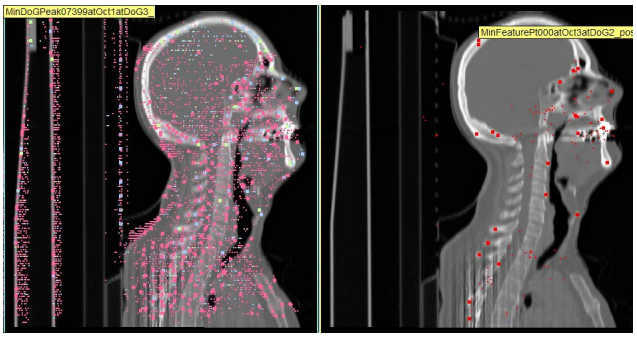

Figure 2. Effect of filtering, e.g. on CT: top – candidate DoG extrema upstream (detected at octave 1, 2, and 3, in pink, purple, and green resp.); bottom – feature points downstream (see typical counts in Table 1). Note: in these pictures, only the points *magnified* are located in the current slice viewed; other dots correspond to feature points in other neighboring slices.

to be satisfied.

Similarly to the 2D case, the principal curvatures are proportional to the eigenvalues of a 3×3 Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}. \quad (3)$$

This matrix is computed from finite differences at the location and scale of the feature point, taking the image anisotropy into account.

Now, in analogy with the approach used in 2D in [5], we have developed an algebraic method to avoid the computationally expensive explicit computation of the eigenvalues of $\mathbf{H}$, since we are only concerned with their ratio. Let $\alpha \geq \beta \geq \gamma$ be the 3 eigenvalues in decreasing order of signed magnitude. They are involved in the trace and the determinant of $\mathbf{H}$ as:

$$\begin{aligned} \mathrm{tr}(\mathbf{H}) &= \alpha + \beta + \gamma = D_{xx} + D_{yy} + D_{zz}; \\ \det(\mathbf{H}) &= \alpha\beta\gamma = D_{xx}D_{yy}D_{zz} + 2D_{xy}D_{yz}D_{xz} \\ &\quad - D_{xx}(D_{yz})^2 - D_{yy}(D_{xz})^2 - D_{zz}(D_{xy})^2. \end{aligned} \quad (4)$$

Let us also introduce the sum of principal second-order minors $\sum \det_2^{\mathrm{P}}(\mathbf{H})$:

$$\begin{aligned} \sum \det\nolimits_2^{\mathrm{P}}(\mathbf{H}) &= \beta\gamma + \gamma\alpha + \alpha\beta = D_{yy}D_{zz} - (D_{yz})^2 \\ &\quad + D_{zz}D_{xx} - (D_{xz})^2 + D_{xx}D_{yy} - (D_{xy})^2. \end{aligned} \quad (5)$$

Condition i) is satisfied when the eigenvalues are either all positive or all negative, corresponding to a dark blob or a bright blob respectively [4]. Deriving a proof by contradiction, we can state that this is equivalent to the condition:

$$\sum \det\nolimits_2^{\mathrm{P}}(\mathbf{H}) > 0, \quad \text{and} \quad \mathrm{tr}(\mathbf{H})\det(\mathbf{H}) > 0. \quad (6)$$

Image Modality	CT	CBCT	MR
Number of voxels	512^2 $\times 152$	512^2 $\times 152$	256^2 $\times 40$
@ Octave 1			
Detected extrema	100,901	17,242	3,893
Meeting thresholds	11,693	5,357	–
Filtered & refined feature pts	93	376	231
Generated features	159	537	317
@ Octave 2			
Detected extrema	14,953	1,690	
Meeting thresholds	1,692	482	
Filtered & refined feature pts	48	91	
Generated features	84	127	
@ Octave 3			
Detected extrema	1,680	189	
Meeting thresholds	267	55	
Filtered & refined feature pts	19	17	
Generated features	35	27	
Total feature points	160	484	231
Total features	278	691	317
Computational time (sec.)	182.3	328.7	25.2

Table 1. Typical statistics obtained on full clinical volumes.

This necessary and sufficient condition allows us to avoid explicitly computing the eigenvalues, in analogy with to how it was avoided in the 2D case, via a simple check that the 2D determinant was indeed negative [10].

At this point, all structures other than blobs, edges and ridges have been rejected. The purpose of condition ii) is also to reject all plate-like or tubular structures. Let r and s be the ratios such that: $\alpha = r\beta$, and $\beta = s\gamma$. Then r and s are greater than $+1$, and rs is the ratio between the largest magnitude eigenvalue and the smaller one. Then, let us consider:

$$\frac{\mathrm{tr}(\mathbf{H})^3}{\det(\mathbf{H})} = \frac{(rs + s + 1)^3}{rs^2}, \quad (7)$$

which depends only on the ratio of eigenvalues rather than their individual values, while comparing their sum with their product, thus highlighting their dispersion. By studying this latter function, which increases both with rs and with s, we have derived that in order to check that the ratio $t = rs$ of principal curvatures is below some threshold, $t_{\max}$, we only need to verify:

$$\frac{\mathrm{tr}(\mathbf{H})^3}{\det(\mathbf{H})} < \frac{(2t_{\max} + 1)^3}{(t_{\max})^2}. \quad (8)$$

Otherwise, features look more like edges or ridges than blobs, and are discarded. Note that this ratio is homogeneous with respect to illumination. Besides, we have experimentally found that an upper threshold $t_{\max} = 5$ is optimal

for CT images, whereas again a less selective $t_{\max} = 20$ is better suited to MR images and CBCT volumes.

2.4. Full orientation invariance

In analogy with the 2D SIFT, the 3D feature descriptor satisfies the properties of scale invariance due to the properties of the DoG function [10] as well as spatial invariance, since no location information is recorded in the descriptor.

The full orientation invariance is achieved in two steps, extending the 2D approach: a) assignment of a canonical orientation to each feature based on local image gradients, and b) representation of the feature descriptor relatively to this orientation, therefore providing invariance to 3D image rotation.

In 2D, there is only one orientation angle to consider. In contrast, in the 3D case, 3 angles need to be handled: azimuth $Az \in [-\pi; \pi]$, elevation $El \in [-\frac{\pi}{2}; \frac{\pi}{2}]$ and tilt $Ti \in [-\pi; \pi]$. Let us denote the spatial finite differences approximating the 3D gradient at the location (x, y, z) of the feature point as: $L'_x = L(x+1, y, z, \sigma) - L(x-1, y, z, \sigma)$, $L'_y = L(x, y+1, z, \sigma) - L(x, y-1, z, \sigma)$, $L'_z = L(x, y, z+1, \sigma) - L(x, y, z-1, \sigma)$ and the voxel size in millimeters as: Δ_x, Δ_y, and Δ_z. Then:

$$Az = \arctan\left(\frac{L'_y/\Delta_y}{L'_x/\Delta_x}\right); \tag{9}$$

$$El = \arctan\left(\frac{L'_z/\Delta_z}{\sqrt{(L'_x/\Delta_x)^2 + (L'_y/\Delta_y)^2}}\right). \tag{10}$$

In 3D, P. Scovanner *et al*. accounted for the azimuth and elevation angles only [12]. For n dimensions, W.A. Cheung and G. Hamarneh accounted for $n-1$ orientation dimensions based on the individual hyperspherical coordinates of the gradient vectors [2, 3]. In particular in 3D, the third self-rotation angle also known as the roll or tilt angle is left undetermined in $[-\pi; \pi]$. For each feature, the missing tilt orientation information cannot stem from the gradient vector only, be it in Cartesian or hyperspherical coordinates, but must encompass its fellow gradients in the region around it. Therefore, the solution we propose in this paper relies on building an additional histogram for tilt determination, once both azimuth and elevation have been assigned to the feature, which we will refer to as its *2-angle orientation*. The tilt histogram is formed by binning the gradient vector components orthogonal to the 2-angle orientation, for all sample points within the same region already considered around the feature point for computing the azimuth and elevation. The idea is then to select peaks in the tilt histogram, which correspond to secondary dominant directions orthogonal to the feature's 2-angle orientation. In practice, this can be done in a computationally efficient way in the coordinate system rotated relatively to the feature 2-angle orientation, since this coordinate system is already needed for the descriptor creation. For each feature point, and for each 2-angle orientation assigned to it, the local 3D coordinate system is used to describe temporarily the local image region. Note that this temporary 3D frame is not repeatable yet, since only its x-axis is aligned with the feature. In addition, for the sake of matching stability, not only the highest peak in the histogram is detected but also any other local peak that is within 80% of the highest peak is also used to create a distinct feature with a distinct tilt orientation at the same location (in the exact same way as it has previously been done in the combined spherical histogram for azimuth and elevation, and as in the 2D case [10]). Subsequently, the same series of refinement sub-steps are performed as for the two first angles.

The final scale-invariant orientation-invariant feature generation algorithm thus consists of these main steps:

For a given feature point:

1) build the gradient-magnitude and Gaussian-weighted 2D histogram of gradient 2-angle orientations in a centered spherical neighborhood. As pointed out in [12], this histogram needs to be bias-corrected through normalizing the accumulated value in each bin by its associated solid angle;

2) find the 2-angle bin with maximum peak and generate other features with a different 2-angle orientations within a 80% ratio if any (with pre-smoothing and post-interpolation);

Then, 3) for each retained 2-angle orientation at this feature point:

3-1) build the Gaussian-weighted histogram of gradient tilt orientations considering the same neighborhood;

3-2) find the tilt bin with maximum peak and generate other features with different tilt orientations within an 80% ratio (with pre-smoothing and post-interpolation).

Only at this moment, a repeatable local 3D coordinate system is obtained in which the local image region can be described.

Then, 3-3) for each retained full 3-angle orientation at this feature point:

3-3-1) assign the orientation to the feature;

3-3-2) compute the descriptor encompassing a centered cubic region of interest of neighboring voxels. Therefore, in order to achieve the orientation invariance: 3-3-2-1) rotate the coordinates in the descriptor neighborhood by the opposite of the feature orientation in 3 Euler angles: first azimuth, then elevation, then tilt; 3.3-2-2) rotate the image gradient orientation in neighborhood by the opposite orientation of the feature point. It should be noticed that this cannot be performed in hyper-spherical coordinates. Handling Cartesian coordinates is necessary in order for the gradient to also be affected by the tilt component.

At this point, one should differentiate between achieving the full orientation invariance and accounting for tilt orien-

tations in the descriptor. For now, we have decided not to incorporate the tilt angle information in the descriptor. This angle already contributes to the consistency and stability of the 2 other angles, azimuth and elevation which are themselves included in the descriptor. In this way, this allows us to avoid increasing the dimensionality of the descriptor and of the matching problem. This is analogous with how the mutual information of gradient intensity may reduce the dimensionality of the rotational registration problem [13].

Therefore, the descriptor dimensionality is not changed and is still 2,048 when using a direct extrapolation from the 2D settings [10]: $4 \times 4 \times 4$ subregions, each summarized by an orientation histogram with 8 azimuth directions (every $\frac{\pi}{4}$) and 4 elevation directions (every $\frac{\pi}{4}$ too), as in [12]. Incorporating the tilt information over 8 directions (again every $\frac{\pi}{4}$) would otherwise make the descriptor dimensionality increase to $4^n \cdot 4 \cdot 8^{n-1} = 2^{5n-1} = 16,384$, instead of $4^n \cdot 4 \cdot 8^{n-2} = 2^{5n-4} = 2,048$, with the dimension $n = 3$ here. The same additional improvements proposed by D.G. Lowe are applied in 3D to reinforce the invariance to non-linear transformations: Gaussian weighting; distance-to-bin-center weighting; normalization, low cap and re-normalization in order to focus on the distribution of gradient orientations and thus reduce the effects of even non-linear illumination changes.

2.5. Computational issues

The main challenge in the practical implementation of the above approach is coping with the computational complexity of multi-scale processing of large volumetric datasets and optimization of data processing to minimize the memory requirements. In particular, we do not perform pre-computation of gradient magnitudes and orientations for the whole pyramid, which was deemed efficient in the 2D case [10]. Instead each salient point is processed separately.

Our current ANSI C implementation is able to process full-resolution CT scans having a 512×512 image matrix in 3-4 minutes on a standard PC with a 3.4 GHz Pentium 4 CPU and 1 GB of RAM. The number of processed DoG extrema, even after excluding the background, is of the order of 11,000 - 16,000. Our implementation extracts 200 salient points and generates their fully reoriented feature descriptor from a full $256 \times 256 \times 50$ MRI volume in less than 40 seconds. The typical statistics for processing full resolution clinical data are summarized in Table 1.

3. Experimental invariance assessment

The 3D saliency of feature points obtained at all scales can be verified visually by inspecting arbitrarily oblique slices. For instance, in CT data, the detected features are mostly bright blobs corresponding to bony structures, yet there are also a few dark blobs from soft tissue, which are also relevant. In MR images, it is more even.

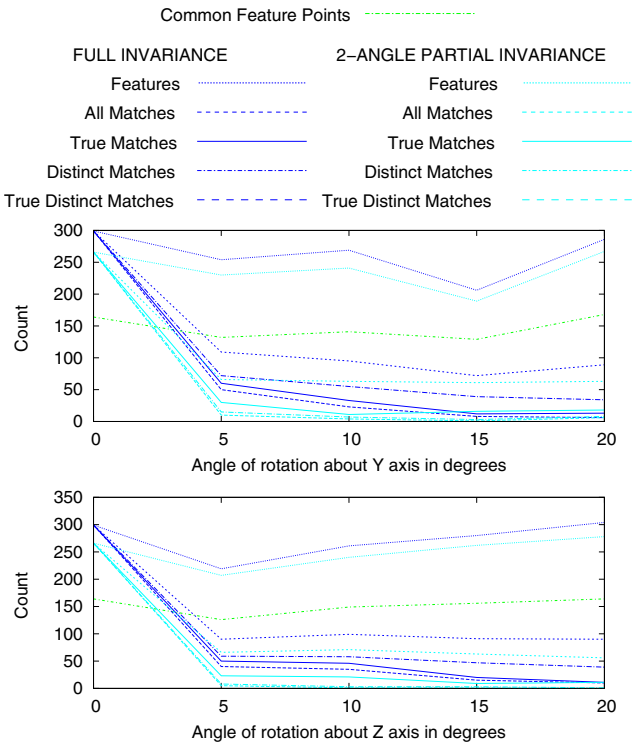

Figure 3. Improvement due to the achieved complete invariance of 3D SIFT extraction and matching over partial rotational invariance (2-angle only), against 3D rigid transformation in Head and Neck CT (effects of rotation about X axis are similar to about Y axis).

In order to verify the invariance with respect to scaling, orientation change and sampling grid, we have applied the following transformation to a clinical $512 \times 512 \times 152$ CT image, and tested the robustness of the feature point extraction, the feature generation and the matching:

– rotation of $5°$, $10°$, $15°$, $20°$ about each of the 3 axes, and an arbitrary oblique axis;

– scaling by factors of 0.6, 0.8, 1.2, 1.4, anisotropically along each of the 3 axes, and isotropically;

– sub-voxel translation by a quarter of voxel, half a voxel, three quarters, along each of the 3 axes independently, and along an arbitrary oblique direction.

Results of testing against rotation are summarized in Figure 3. They show a significant improvement in feature detection and superior matching rate due to the derived full orientation invariance of the descriptors, compared to when the tilt angle normalization is switched off (by simply setting the number of tilt histogram bins to 1). The difference at angle $0°$ corresponds to the added features due to multiple tilt orientations retained. A match is distinct when the ratio to the 2nd best nearest neighboour in the descriptor space is below 90%. A match is deemed true when it lies

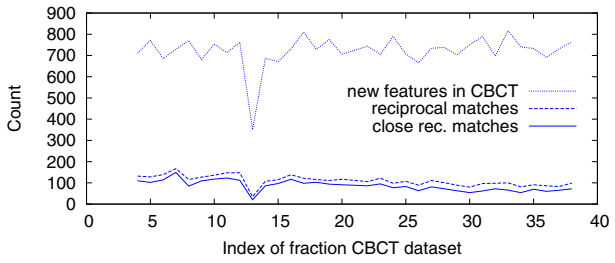

Figure 4. Application to matching images of the same patient over a 7-week treatment course. The max. 1^{st}-to-2^{nd} match ratio was 80%; reciprocal matches within 30 mm were considered close.

within 1 voxel diagonal length (here 3.6 mm) of the ground truth position. This demonstrates how the full-orientation normalization provides more invariance and more stability to the features against rotation, actually in terms of both absolute count of true matches (important for a denser field registration) and of percentage out of all matches. In other words, this increases significantly the sensitivity and the selectivity of the feature extraction and matching.

4. Application to radiation therapy image data

In our first experiment, we applied feature-based registration to propagate contoured organ outlines from a planning CT dataset to cone beam CT data acquired prior to each treatment fraction for the purpose of setup validation. This contour propagation can help to determine a safe dose map to be reflected in planning-organ-at-risk volume margin. When applying direct descriptor matching, the consistent effect of a decrease in the matching rate from about 20% to 12% due to the changes in the patient's anatomy can be observed on Figure 4, including always 65% to 80% of close reciprocal matches (within 30 mm).

In order to address the need for speed and taking into account a limited setup error due to patient immobilization, we have adopted the following scheme: i) extraction of a patient-specific model in terms of 3D SIFT feature points from a single CT volume; ii) retrieval of this model in a CBCT data model using a block-matching technique based on the correlation coefficient, and iii) propagation of delineations using a thin-plate spline transformation derived from the retrieved point correspondences. The results of contour propagation are illustrated in Figures 5, which show an accuracy improvement compared to merely copying the contours, in the presence of changing anatomy due to weight loss.

As for inter-patient descriptor matching, we can show only qualitative results for now, which are promising. Before achieving the full orientation invariance, not a single match was anatomically correct; false positives were omnipresent. With the fully invariant implementation, for each pair of datasets in a group of four patients we have observed 10-12 true matches out of 30 to 40 reciprocally distinct matches. Some examples of true matches are shown in Figure 6. We are currently investigating this application in a more systematic way, with a view to introduce geometric group-wise constraints and/or switch to a scheme where no 1-to-1 correspondence is assumed [14].

On the other hand, interestingly enough, the inter-modality matching results on the same patient, e.g. CT vs. MR, do not show any significant improvement owing to the achieved invariance. However, intra-modality matching with contrast changes, such as CT to cone beam CT shows that the 3D SIFT descriptors are nevertheless robust to changes of illumination.

5. Discussion and perspectives

When developing his SIFT approach, D.G. Lowe was inspired by a model of complex biological vision, where some fuzziness is introduced as for the location of the response to a 2D gradient at a particular and spatial frequency. This turned out to allow for wider matching and recognition of 3D objects from a range of viewpoints [10]. This led to the choice of orientation histograms accounting for as much as 4×4 sample regions, and the addition of a tri-linear interpolation.

The same fuzziness-based robustness principle led to the conclusions in [2] that the n-SIFT feature descriptor, even though not reoriented, was surprisingly outperforming a reoriented global histogram-based feature, even against a rotation alteration. Only 8 bins summarized each of the hyperspherical coordinates, thus widely covering $\frac{\pi}{4}$ of azimuth angle and actually $\frac{\pi}{8}$ of elevation angle. The observed relatively poor performance of the reoriented global histogram feature is probably due to the missing tilt angle.

The perspectives of further extending the approach proposed in this paper may include the following steps. A finer representation of the currently used $4 \times 4 \times 4$ subregions can be introduced to obtain more reliable orientations. Also, the use of quaternions instead of Euler angles can be investigated to improve the orientation handling. Finally, the tilt angle can be directly incorporated into the descriptors, which would increase the dimension 8-fold to 16,384. Due to the memory requirements, this approach is only feasible for cropped volumes of interest. However, dimensionality reduction techniques, such as e.g. PCA-SIFT [6], can be applied in this case to make the memory consumption more manageable.

Acknowledgment

This research is supported in part by Philips Healthcare. We thank the anonymous reviewers for their valuable comments.

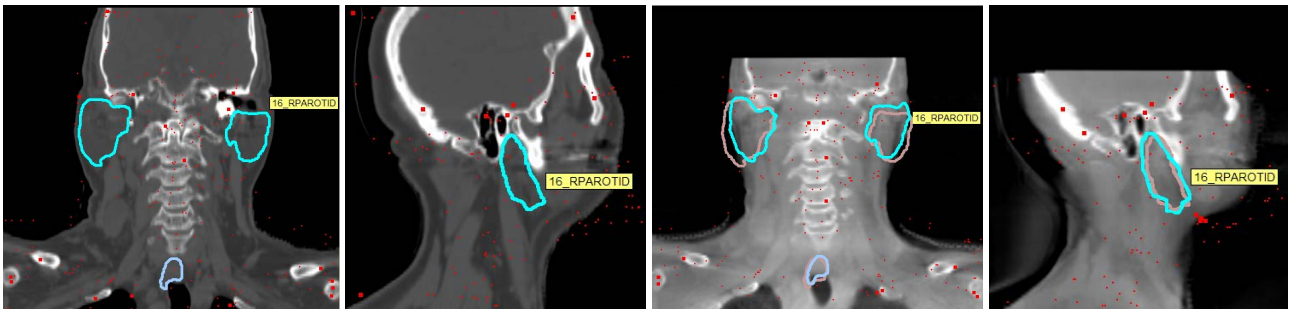

Figure 5. Contour propagation from CT to cone beam CT dataset, based on the matched red points, on the example of parotids. For comparison, the contours simply copied from CT are shown in brown.

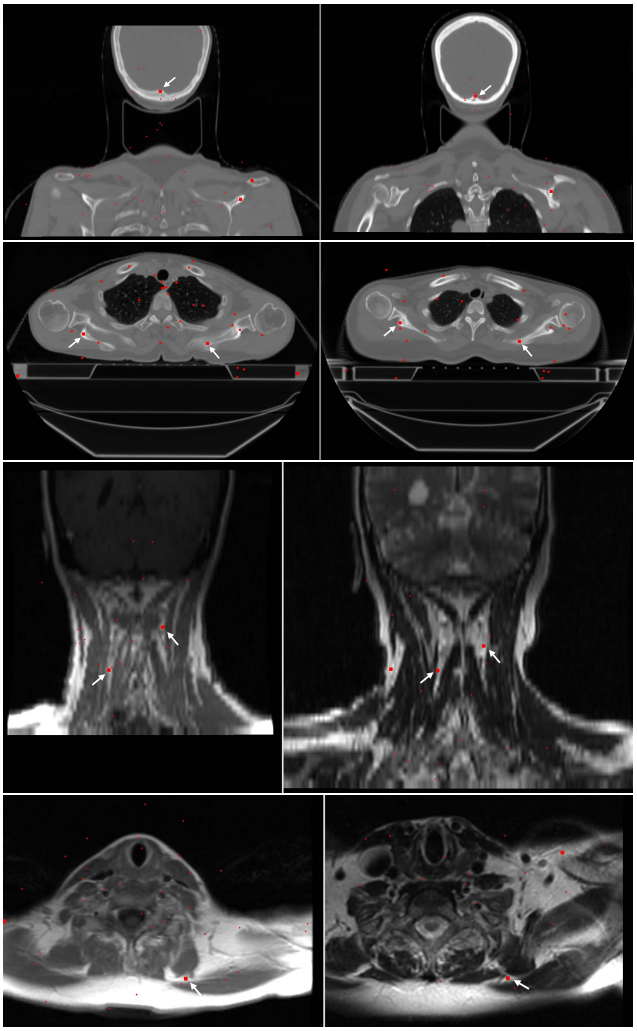

Figure 6. Application to inter-patient feature matching in CT and MR. Note: here again, the dots not magnified are matched points not located in the current slice but in other neighboring slices.

References

[1] M. Brown and D.G. Lowe. Invariant features from interest point groups. *proc. of the 13th Brit. Mach. Vis. Conf. (BMVC)*, pages 656–665, Cardiff, Wales, September 2002.

[2] W.A. Cheung and G. Hamarneh. n-SIFT: n-dimensional scale invariant feature transform for matching medical images. *proc. of the 4th Int. Symp. on Biomed. Imag. (ISBI)*, pages 720–723, Washington, DC, April 2007.

[3] W.A. Cheung and G. Hamarneh. Scale invariant feature transform for n-dimensional images (n-SIFT). *The Insight Journal*, IJ-2007 (207):7 pages, December 2007. http:insight-journal.orgInsightJournalManagerview_reviews.php?pubid=207.

[4] A.F. Frangi, W.J. Niessen, K.L. Vincken, and M.A. Viergever. Multiscale vessel enhancement filtering. *proc. of the 1st int. conf. on Med. Image Comp. & Comp.-Assist. Interv. (MICCAI)*, pages 130–137, Cambridge, MA, Oct. 1998.

[5] C. Harris and M. Stephens. A combined corner and edge detector. *proc. of the 4th Alvey Vis. Conf.*, pages 147–151, Manchester, UK 1988.

[6] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *proc. of the 17th IEEE conf. on Comp. Vis. & Patt. Recogn. (CVPR)*, pages 506–513, Washington, DC, June-July 2004.

[7] J.J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–396, 1984.

[8] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *J Appl Stat*, 21(2):224-270, '94.

[9] D.G. Lowe. Object recognition from local scale-invariant features. *proc. of the 7th IEEE Int. Conf. on Comp. Vis. (ICCV)*, pages 1150–1157, Corfu, Greece, Sept. 1999.

[10] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comp. Vis.*, 60(2):91–110, Nov. 2004.

[11] M. Moradi, P. Abolmaesumi, and P. Mousavi. Deformable Registration Using Scale Space Keypoints. *proc. SPIE Med. Imag. - Imag. Proc.*, Vol. 6144:G1-G8, San Diego, CA, 2006.

[12] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. *proc. of the 15th ACM int. conf. on Multimedia*, pages 357–360, Augsburg, Germany, September 2007.

[13] R. Shams, R. Kennedy, P. Sadeghi, and R. Hartley. Gradient intensity-based registration of multi-modal images of the brain. *proc. of the 11th IEEE Int. Conf. on Comp. Vis. (ICCV)*, 8 pages, Rio de Janeiro, Brazil, October 2007.

[14] M. Toews and T. Arbel. A statistical parts-based model of anatomical variability. *IEEE Transactions on Medical Imaging*, 26(4):497–508, April 2007.