21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Feature-aligned 4D Spatiotemporal Image Registration

Huanhuan Xu[1], Peizhi Chen[2], Wuyi Yu[1], Amit Sawant[3], S.S.Iyengar[4], and Xin Li*[1]

[1]Louisiana State University, [2] Xiamen University,
[3]UT Southwestern Medical Center, [4]Florida International University.
*Email: xinli@lsu.edu

Abstract

In this paper, we develop a feature-aware 4D spatiotemporal image registration method. Our model is based on a 4D (3D+time) free-form B-spline deformation model which has both spatial and temporal smoothness. We first introduce an automatic 3D feature extraction and matching method based on an improved 3D SIFT descriptor, which is scale- and rotation- invariant. Then we use the results of feature correspondence to guide an intensity-based deformable image registration. Experimental results show that our method can lead to smooth temporal registration with good matching accuracy; therefore this registration model is potentially suitable for dynamic tumor tracking.

1. Introduction

Advancing modern 4-D (3D spatial + 1D temporal) CT techniques provide abundant spatial and temporal data of the patient for clinical monitoring and diagnosis. Temporally parameterizing these scan data can facilitate many clinical analysis and planning tasks. For example, in lung cancer radiation radiotherapy, 4D-CT images can be used to model the motions and deformations of the tumor and surrounding organs, to guide treatment planning [3]. Image registration plays an important role in the current motion estimation methods by establishing temporal correspondences [5, 1].

Compared with the conventional image registration techniques, 4D spatiotemporal registration can avoid the bias caused by a predetermined reference frame, and can enforce both *spatial and temporal smoothness* of the transformations, which indicates physically natural deformations [6].

However, most of the current spatiotemporal dynamic images are fully guided by the image's intensity [5, 1, 9]. The aligning computation therefore reduces to minimizing a non-linear problem having many local minima, which usually has high computational cost and, more importantly, requires a good initial guess to reach a desirable matching. Feature constraints can effectively guide the optimization from getting trapped on locally. For example, in many video tracking tasks, the SIFT descriptor has demonstrated great efficacy and been widely used due to its discriminative feature [4]. Directly generalized SIFT descriptor in 3D [2], however, could be sensitive to scalings and rotations of the deforming objects in the volume images. In this paper, we first introduce a modified 3D-SIFT descriptor that can handle these more reliably, then we develop a feature-constrained 4D dynamic registration algorithm to spatially and temporally match deforming volume images. This paper has two main contributions.

1. We propose an improved 3D feature extraction and matching algorithm based on N-SIFT method. The new method can detect more corresponding features and have less matching error.
2. We formulate a 4D spatiotemporal feature alignment metric that minimizes the position invariance over time to guide the image registration which leads to more accurate results.

2. Method

2.1. Feature Point Extraction and Matching

To handle the registration of volumetric images, Scovanner et al. [7] proposed a 3D SIFT descriptor and applied it in action recognition. Cheung and Hamarneh extended SIFT to N-Dimension SIFT [2] (N-SIFT) and

978-4-9906441-1-6

showed its effectiveness on volumetric images. However, neither descriptor is scale or rotation invariant. To adequately describe images of deforming organs, we shall improve the existing 3D SIFT descriptor.

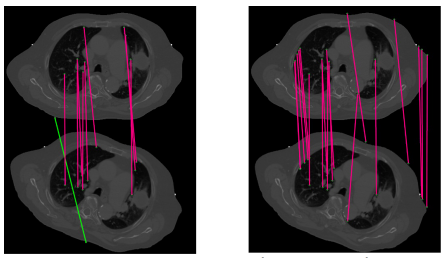
The procedure of N-SIFT includes scale space extrema detection, orientation assignment, descriptor construction and matching [2]. For an input volume image, we first extend method [4] to locate its keypoints with sub-pixel accuracy.

One limitation of N-SIFT is its sensitivity against local rotation. To more robustly handle this, we can assign multiple directions (rather than just one dominant direction used in [7]) to a keypoint region. We calculate an orientation histogram of a region around the keypoint with width $6 * \sigma$ where σ is the scale of the keypoint. This orientation histogram has 36×36 bins covering 360° of the orientations. The highest peak of the histogram corresponds to the dominant direction. Here, we consider local peaks within 80% of the highest peak also to be the directions of the keypoint region. Region that is chosen in the construction of the descriptors can be reoriented according to its directions by multiplying its rotation matrixes [7]. Descriptors are constructed on the reoriented regions. Multiple directions make our 3D SIFT more robust to the image rotation.

N-SIFT is also not scale-invariant, since it computes the descriptor on the original image and the size of the region around the keypoint is fixed. We use a scale selection method to deal with scale change. We construct the descriptors on the corresponding Gaussian smooth image. The region around the keypoint is defined and divided into $4\times4\times4$ patches. We set its patch size to be $3 * \sigma$ which is related to its scale. In this way, our descriptor is robust against scaling.

For the matching process, since N-SIFT matches descriptors directly, a point may be matched to more than one point. Some of the matchings are wrong. Hence, we further conduct a RANSAC algorithm to deal with this one-to-many correspondence issue and remove the outliers. In our work, before doing 4D registration we first perform feature extraction and matching between every two consecutive volume images, then choose those consistent correspondences that appear in all time frames.

A simple example is given in Fig. 1 to demonstrate the rotation invariance of the new descriptor. A lung CT volume image (dimension $465\times300\times20$) is used as the reference; its subsequent image has rotated by 20° along Z axis (this happens when the patient rotates). We compare the correspondences found using N-SIFT and our improved 3DSIFT. N-SIFT method extracts fewer matching pairs and has some error matchings while our algorithm works correctly and find more matched features. Note that this matching is done on volume images although we only illustrate a 2D cross section.



(a)N-SIFT (b)Improved 3DSIFT

Figure 1. Feature Extraction and Matching.

2.2. 4D Free-form B-spline Deformation

We present a 4D deformation model, based on a 4D free-form B-spline incorporating both the spatial and time dimensions [5]. Denote the 4D input image as $I(\mathbf{y})$, where $\mathbf{y} = (\mathbf{x}^T, t)^T \in R^3 \times R$ is a coordinate in I which consists of a spatial location $\mathbf{x} \in R^3$ and temporal location $t \in R$. The B-spline based coordinate transformation $\mathbf{T}_\mu$ is defined as follows:

$$\mathbf{T}_\mu(\mathbf{y}) = \mathbf{y} + \sum_{\mathbf{y}_k \in N_y} \mathbf{p}_k \beta^r(\mathbf{y} - \mathbf{y}_k), \qquad (1)$$

where $\mathbf{y}_k$ is a knot on the parametric domain; $\beta^r(\cdot)$ is the r-th order multidimensional B-spline polynomial; $\mathbf{p}_k$ is the B-spline control points to be solved, and N_y denotes the neighboring region providing local support to the B-spline at $\mathbf{y}$. The knots $\mathbf{y}_k$ are defined on a 4D regular grid, uniformly overlaid on the image. The parameter vector μ consists of the collection of the first 3 elements of each $\mathbf{p}_k$. The last element of each $\mathbf{p}_k$ is fixed to zero, which ensures that only deformation in the spatial domain are allowed. In the following $T_\mu(\mathbf{y})$ is interchanged with $T_\mu(\mathbf{x}, t)$ for convenience.

In order to align all the images, we assume that after correct registration the intensity values at corresponding spatial locations over time are equal. Hence we should minimize the image intensity changes over time. An implicit reference frame is used to eliminate the need to choose a reference time point image. The dissimilarity metric, or cost function, is therefore defined as:

$$C(\mu) = \frac{1}{|S||\Gamma|} \sum_{\mathbf{x}\in S} \sum_{t\in\Gamma} (I(\mathbf{T}_\mu(\mathbf{x}, t)) - \bar{I}_\mu(\mathbf{x}))^2, \quad (2)$$

where $\bar{I}_\mu(\mathbf{x})$ is the average intensity value over time after applying transformation $\mathbf{T}_\mu$,

$$\bar{I}_\mu(\mathbf{x}) = \frac{1}{|\Gamma|} \sum_{t\in\Gamma} I(\mathbf{T}_\mu(\mathbf{x}, t)), \qquad (3)$$

and S and Γ are the set of spatial and temporal voxel coordinates respectively.

As none of the images are chosen as an anatomical reference, it is necessary to add a geometric constraint to define the reference coordinate frame. Similar to [1], we define the reference frame by constraining the average deformation to be the identity transformation

$$\frac{1}{|\Gamma|}\sum_{t\in\Gamma}[\mathbf{T}_\mu(\mathbf{x},t)]_\mathbf{x} = \mathbf{x}, \tag{4}$$

where $[\cdot]_\mathbf{x}$ means get the position component $\mathbf{x}$ from current 4D point $(\mathbf{x}, t)$. Then the optimal deformation field can be computed by the adaptive stochastic gradient descent optimizer (ASGD).

$$\hat{\mu} = \arg\min_\mu C(\mu), \text{ subject to } (4) \tag{5}$$

After this registration all time point images are aligned in the implicit reference frame.

2.3. Feature-aligned Registration

In order to compute the transformation $\mathbf{T}_\mu^{ij}$ which maps coordinates from time point i to time point j, we need to compute the inverse mapping $\mathbf{T}_\mu^{-1}$ which maps coordinates from the input image coordinate frame to the reference frame. Since the mapping $\mathbf{T}_\mu$ may not be bijective, its inverse mapping $\mathbf{T}_\mu^{-1}$ may not actually exist. Here we define an approximate inverse mapping using a B-spline $\mathbf{T}_\mathbf{v}$ by minimizing

$$F_{Pos}(\mathbf{v}) = \frac{1}{|Y|}\sum_{\mathbf{y}\in Y}||\mathbf{T}_\mathbf{v}(\mathbf{T}_{\hat{\mu}}(\mathbf{y})) - \mathbf{y}||^2 \tag{6}$$

where Y is the set of knots. In order to prevent foldings in the transformations we choose smaller grid spacing to yield more accurate results.

After our feature extraction and matching, we get the coherent corresponding features of all spatial images along temporal dimension. We enforce the feature matching constraints in the inverse registration.

Suppose we have N coherent features. We denote the ith feature point on time j (on jth image) as p_{ij}, where $i = 1,\ldots,N, j = 1,\ldots,\Gamma$. Intuitively, after correct registration the corresponding feature should be at the same point in the reference frame. That is, for each i, we shall also minimize the variance of $\mathbf{T}_\mathbf{v}(p_{ij}, j)$ in the reference image, where $(p_{ij}, j)^T$ denotes a 4D vector in the spatial-temporal space.

The cost function for feature alignment is

$$F_{Fea}(\mathbf{v}) = \frac{1}{N|\Gamma|}\sum_{i=1}^{N}\sum_{t\in\Gamma}||[T_\mathbf{v}(p_{it},t)]_\mathbf{x} - [\bar{T}_\mathbf{v}(p_i.)]_\mathbf{x}|| \tag{7}$$

where

$$\bar{T}_\mathbf{v}(p_i.) = \frac{1}{|\Gamma|}\sum_{t\in\Gamma}[T_\mathbf{v}(p_{i,t},t)]_\mathbf{x} \tag{8}$$

The final objective function for estimating the optimal deformation field is formulated as:

$$F_\mathbf{v} = F_{Pos} + \lambda F_{Fea} \tag{9}$$

Table 1. The registration error in mm, on 40 landmarks among 0^{th},5^{th},9^{th} time frames of the POPI-data. $E_{i,j}$ is the matching error from i^{th} to j^{th} frame, $\bar{E}$ is the mean error for the whole sequence.

	$E_{0,5}$	$E_{0,9}$	$E_{5,0}$	$E_{5,9}$	$E_{9,0}$	$E_{9,5}$	$\bar{E}$
[5]	2.94	0.98	2.88	2.86	0.99	2.91	2.26
Our's	2.87	0.85	2.77	2.74	0.87	2.84	2.16

where λ is the weighting factor controlling the strength of the feature constraint term. We determine those transform parameters that minimize the total metric as

$$\hat{\mathbf{v}} = \arg\min_\mathbf{v} F(\mathbf{v}). \tag{10}$$

We also solve Eq-(10) using ASGD, then we can get the transformation from time point i to time point j:

$$T^{ij}_{\hat{\mu},\hat{v}}(\mathbf{x}) = [T_\mu([T_\mathbf{v}(\mathbf{x},t_i)]_\mathbf{x}, t_j)]_\mathbf{x}. \tag{11}$$

3. Implementations and Experiments

We implement our model via a multi-resolution strategy and use linear interpolation in the spatial domain for the derivation of intensity values for any point not on a grid. Our algorithm was implemented in C++ using an Intel Core E7300 @2.66 GHz, 4GB RAM. The registration on the POPI-model and our tumor data take approximately 45 mins, of which 22 mins were spent to compute the 4D forward registration and 23 mins were spent to compute the 4D inverse registration.

Experiments on POPI Dataset. Our first experiment is conducted on the POPI dataset [8]. This dataset contains one 4D CT series including ten 3D volumes representing ten different phases of one breathing cycle. In the 3D volume at time frame t, the coherent landmarks (a set of 3D points, denote as $P_t = \{p_{t,1}, p_{t,2}, ..., p_{t,|P_t|}\}$) are available and can be used to evaluate the registration. We use the time frames 0, 5, and 9 with 571 feature correspondences to do group registration. The registration results were evaluated by the *mean target registration error* (MTRE) between the set of landmark points $\{P_0, P_5, P_9\}$. Denote MTRE as $E_{r,t} = \frac{1}{|P_t|}\sum_{p_{t,i}\in P_t}||T^{r,t}(p_{r,i}) - p_{t,i}||$, where $p_{t,i}$ is a landmark i in time t. In our experiments, we set the control weight in Eq. (9) as $\lambda = 0.1$. Table 1 shows the comparison between our method and the algorithm of [5]: our method outperforms [5] by introducing smaller MTRE errors.

Lung Tumor Registration. Our second experiment is to apply our registration model in dynamic tumor tracking (Fig. 2). We detect 202 feature correspondences among the image sequence. Before registration, we segment the tumor in the first frame by using 3D

graph-cut segmentation [3]. Then with our registration results, we track this tumor in the following second/third time sequence (shows in the second/third column of fig. 2). The bottom of this figure depicts the registration of this tumor among different time sequences.

Furthermore, we compute an unbiased difference image between the deformed image and the target image to evaluate the registration accuracy. Assume the 3D source image is $I^i(\mathbf{X})$ in i-th frame, the 3D target image is $I^j(\mathbf{X})$ in j-th frame. The deformed image is $I(Tx)$ where $Tx = T^{ij}(\mathbf{x})$ and $\mathbf{x}$ from the source image. In order to avoid the influence of the gray value of original pixel, we normalize the difference frame: if $I(Tx) + I^j(Tx) \neq 0$ then $I^d(Tx) = \frac{|I(Tx) - I^j(Tx)|}{I(Tx) + I^j(Tx)}$; otherwise, $I^d(Tx) = 0$. It is easy to check that this metric is symmetric between the deformed image and target image. Smaller I^d indicates more accurate registration.

Fig 3 (a) shows the projection of the difference image between the second and the third frame. (b) shows the histogram of the computed difference value. We construct this histogram based on the normalized difference frame between the deformed second frame and the third frame. We count the occurrence of each difference value and divide it by the total number of the pixels to get its probability. We can see in larger than 90% pixels, the difference value is less than 0.1, and the mean difference value is 0.016. These indicates that our registration introduces very small error between deformed image and the target image. Thus our registration can be used for tumor motion tracking (see Fig.2).

Also, this visualization (Fig 3 (a)) can also help us to identify the region with large registration errors for subsequent matching refinement. We can see around the boundary part and the central of left lung part have larger difference value. In the future, we will develop hierarchically spline scheme to support adaptive refinement, so that we can insert more knots in these regions to reduce the registration error.

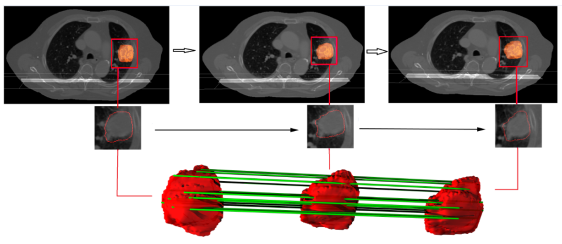


Figure 2. Tumor tracking with our registration.

4. Conclusion

We propose an automatic feature-guided 4D image registration framework. We develop an improved

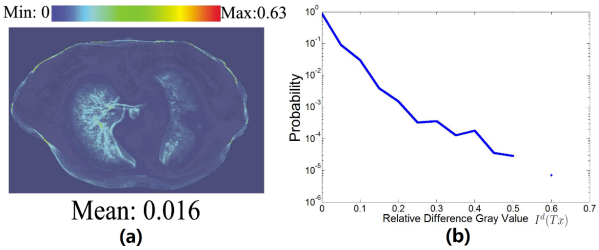


Figure 3. The color-encoded difference image and its histogram distribution. The volumetric difference image is projected onto 2D, (a) shows the accumulated intensity. This intensity distribution is illustrated in (b); 90% pixels has difference value < 0.1. Note that y-axis is logarithmic-scaled.

3D-SIFT descriptor for reliable feature extraction and matching. Compared with existing 4D registration model we achieve better landmark predication accuracy. Our model also has good ability to do tumor motion estimation which can greatly facilitate lung tumor radiotherapy planning and management.

Acknowledgements. This work is supported by Louisiana BOR-(RCS)-LEQSF(2009-12)-RD-A-06 and National Natural Science Foundation of China No. 61170323.

References

[1] K. Bhatia, J. Hajnal, B. Puri, A. Edwards, and D. Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *Intl. Symp. on Biomedical Imaging.*, volume 1, pages 908 – 911, 2004.

[2] W. Cheung and G. Hamarneh. N-sift: N-dimensional scale invariant feature transform for matching medical images. In *ISBI 2007*, pages 720 –723, 2007.

[3] S. S. Iyengar, X. Li, H. Xu, S. Mukhopadhyay, N. Balakrishnan, A. Sawant, and P. Iyengar. Toward more precise radiotherapy treatment of lung tumors. *IEEE Computer*, 45:59–65, 2012.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[5] C. Metz, S. Klein, M. Schaap, T. van Walsum, and W. Niessen. Nonrigid registration of dynamic medical imaging data using nd + t b-splines and a groupwise optimization approach. *Med.Img.Analy.*, 15:238 – 249, 2011.

[6] J.-M. Peyrat, H. Delingette, M. Sermesant, C. Xu, and N. Ayache. Registration of 4d cardiac ct sequences under trajectory constraints with multichannel diffeomorphic demons. *IEEE Trans.Med.Img.*, 29:1351–1368, 2010.

[7] P. Scov., S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. MM'07.

[8] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse. Point-validated pixel-based breathing thorax model. In *ICCR07*.

[9] G. Wu, Q. Wang, J. Lian, and D. Shen. Estimating the 4d respiratory lung motion by spatiotemporal registration and building super-resolution image. MICCAI'11.