SIGRAD 2013, pp. 11–16
T. Ropinski and J. Unger (Editors)

Feature Tracking in Time-Varying Volumetric Data through Scale Invariant Feature Transform

Khoa Tan Nguyen[1] and Timo Ropinski[1]

[1]Scientific Visualization Group, Linköping University, Sweden

Abstract

Recent advances in medical imaging technology enable dynamic acquisitions of objects under movement. The acquired dynamic data has shown to be useful in different application scenarios. However, the vast amount of time-varying data put a great demand on robust and efficient algorithms for extracting and interpreting the underlying information. In this paper, we present a gpu-based approach for feature tracking in time-varying volumetric data set based on the Scale Invariant Feature Transform (SIFT) algorithm. Besides, the improved performance, this enables us to robustly and efficiently track features of interest in the volumetric data over the time domain. As a result, the proposed approach can serve as a foundation for more advanced analysis on the features of interest in dynamic data sets. We demonstrate our approach using a time-varying data set for the analysis of internal motion of breathing lungs.

1. Introduction

As the major focus of medical imaging has been the understanding of anatomical structures, vast research efforts has been dedicated to the acquisition and interpretation of anatomical modalities. While these techniques enable interpretation of high-resolution static anatomical images, time-varying data is now becoming more important, as the diagnostic workflow can be significantly improved by better understanding of organ function. This has led to the emergence of many functional modalities, which allow multimodal imaging of physiological processes alongside the anatomical image data serving as a context. In some cases the functional information is extracted from originally anatomical modalities,. The most prominent case of this development is probably fMRI (functional magnetic resonance imaging) that enables imaging of brain activity by detecting oxygen level changes in the blood flow.

The latest exploitation of anatomical modalities in a functional context arose with the recent advances in CT (computed tomography) imaging. Driven by the demands of imaging the beating heart, the scanning times of modern CT scanners nowadays enable a dynamic acquisition under movement. Through this technological advancement, new application cases become possible. For instance, 4D CT and 4D MRI can depict the breathing motion of the internal organs. As the amount of acquired data increases, there has been a great demand on new techniques that enable robust and efficient ways to interpret the vast amount of data under investigation.

In this work, we present a GPU-based implementation of the scale invariant feature transform (SIFT) algorithm [Low99, Low04] applied to the analysis time-varying volumetric data. The advantage of the proposed approach is two fold. First, it supports interactive feature detection. Second, it enables us to robustly and efficiently track features of interest in volumetric data over the time domain.

The remainder of the paper is structured as follows. In the next section, we review works that are related to our approach. In Section 3, we present an overview of the SIFT algorithm. We then present our GPU-based implementation in Section 4. We report the result of the proposed approach applied to the analysis of the internal motion of a time-varying volumetric data set of the lung in Section 5, and conclude the paper in Section 6

2. Related Work

In order to track features throughout a time-varying volumetric data sets as well as across different acquisitions, a robust feature tracking approach is mandatory. The SIFT algorithm proposed by Lowe fulfills this criterion [Low99]. SIFT focuses on extracting points of interest with a high saliency,

and that are stable across different scales. These points of interest are then represented by feature descriptors, which are invariant with respect to scaling, translation, and orientation. Since its introduction, SIFT has been widely used in the field of computer vision for image matching. While initially proposed for 2D images, SIFT has been extended to work with higher dimensional data and has been applied to different applications involving salient feature localization and matching such as motion tracking [SAS07, AKB*08], group-related studies [TWICA10], volumetric ultrasound panoramas [NQY*08], and complex object recognition [FBMB]. As an extension to the standard SIFT algorithm, Lowe proposed a guideline for optimal parameter settings that improve the accuracy as well as the performance [Low04]. To further improve its performance and make it interactively applicable, Heyman et al. proposed a GPU-based implementation of the SIFT algorithm enabling real-time feature detection and matching between images [HMS*07]. Although this algorithm was designed for, and tested on, 2D images of 3D objects, the underlying mathematical theory does not limit its extension to handle higher dimensional data. Scovanner et al. proposed a new approach to the creation of SIFT descriptors for the application of action recognition in video (2D images + time domain) [SAS07]. Cheung et al. generalized the scale space principle and applied SIFT to n-dimensional dataset [CH07, CH09]. Their extension has been applied to 3D MR images of the brain and 4D CT of a beating heart. In order to extend SIFT to handle high dimensional data set, they proposed the use of hyperspherical coordinate representations for volume gradients as well as multi-dimensional histograms to capture the distribution of gradient orientations in the neighborhood of detected feature locations. To improve the quality of the detected feature locations, Allaire et al. made use of the 3×3 Hessian matrix to compute the principle curvature at the detected feature locations [AKB*08]. This enables the filtering of features that are of less interest in medical data, such as non-blob and edge-like locations. In addition, the authors presented a technique that takes into account the tilt angle at the detected feature locations during the construction of SIFT descriptors to achieve full rotation invariance. The proposed extensions have been applied to complex object recognition in 3D volumetric CT data [FBMB]. Paganelli et al. further reported the result of the preliminary feasibility study on the application of SIFT to feature tracking in time-varying data sets [PPP*12]. Recently, Yu et al. compared SIFT to other feature detection algorithms and showed that SIFT achieve a balanced result between stability and performance [YWC12].

As we are interested in interactive feature tracking, besides the robustness, also the performance of the feature tracking is of interest. In comparison to previous works, we exploit the computing performance of the GPU through a GPU-based implementation applied to 4D data sets (3D volume + time domain).

3. 3D SIFT

While the SIFT algorithm has been initially proposed for 2D data, our work is based on recent extensions which have generalized it to 3D [CH07, SAS07, CH09]. The algorithm is performed in three successive stages: feature location detection, feature descriptor construction, and feature identification.

Feature location detection. In the first stage, the volumetric input data, $I(x,y,z)$, is convoluted with variable-scale Gaussian functions, $G(x,y,z,k\sigma)$, to generate a scale space, $L(x,y,z,k\sigma)$, as follows:

$$L(x,y,z,k\sigma) = G(x,y,z,k\sigma) * I(x,y,z) \tag{1}$$

where k is a constant multiplicative factor for separating scales in the scale-space.

The local extrema of the difference-of-Gaussian functions applied to this scale space are considered to be potential local features in the original volumetric data:

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

Lindeberg and colleagues could show that these local extrema are a close approximation to the scale normalized Laplacian-of-Gaussian [Lin94], $\sigma^2 \bigtriangledown^2 G$, which are the most stable features in the input image [MTS*05].

In order to improve the stability of the detected feature location in volumetric data sets, Allaire et al [AKB*08] proposed the principal curvature thresholding technique to filter out the blob-like features, which are usually of no interest. The proposed technique is based on the analysis of the Hessian matrix, which describes the local curvature at a detected feature location:

$$H = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{xy} & D_{yy} & D_{yz} \\ D_{xz} & D_{yz} & D_{zz} \end{bmatrix}$$

The elements of H are computed using finite differences at the detected feature location in the corresponding scale, taking the anisotropy of the image into account. Let t_{max} be the curvature threshold, which is the ratio between the largest magnitude eigenvalue and the smaller one, the following conditions help to filter out blob-like features and improve the stability of the detected feature locations:

$$\begin{aligned} &(1) \quad tr(H)det(H) > 0 \quad \text{and} \quad \textstyle\sum det_2^P(H) > 0 \\ &(2) \quad \frac{tr(H)^3}{det(H)} < \frac{(2t_{max}+1)^3}{t_{max}^2} \end{aligned}$$

where $det_2^P(H)$ is the sum of second-order principal minors of H, $tr(H)$, and $det(H)$ are the trace and the determinant of the Hessian matrix, H, respectively.

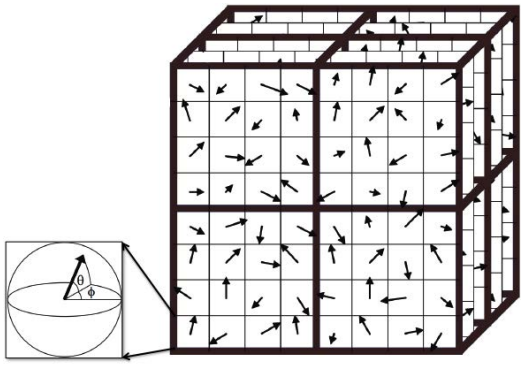


Figure 1: *The formulation of a 3D SIFT descriptor with its corresponding sub-volume.*

Feature descriptor construction. The aim of the second stage is to construct a unique descriptor to represent the detected feature location in such a way that it is most invariant with respect to rotation, scaling, and translation. The construction of such a descriptor is based on the gradient orientations in the neighborhood of the detected feature locations and presented as a histogram of gradient orientations. In 3D volumetric images, there are three angles that need to be handled during the construction of a descriptor: azimuth, elevation, and tilt angles. Consequently, while in 2D a 1D histogram can be used to describe a feature descriptor, in 3D a 2D histogram is required to capture the distribution of the azimuth and elevation angles. It is worth pointing out that while the azimuth and elevation angles can be derived directly from the orientation of the gradient, the tilt angles require more complex analysis [AKB*08]. A Gaussian-weighted kernel is commonly applied to the gradient magnitudes in order to put less emphasis on the gradients that are further away from the detected feature location during the construction of the feature descriptor. Figure 3 illustrates the formulation of a 3D SIFT descriptor with its corresponding sub-volume.

During the construction of a feature descriptor, the maximum peak in the histogram of gradient orientations represents the dominant orientation of the neighborhood around the detected feature location. As a result, the gradient orientations in the neighborhood region are then rotated in relation to the identified dominant orientation to attain a rotation invariance descriptor. While the resulting descriptor is a unique representation of the detected feature, discarding the other gradient orientations of lower magnitude can have negative impact on the feature identification stage. For instance, by creating additional descriptors for smaller peaks, which are of 80% the maximum peak in the histogram of gradient orientations, the result in the feature identification can be improved [Low04].

Feature identification. Once the features descriptors have been constructed for two data sets to be compared, they can be used to identify matching features. Therefore, different techniques such as RANSAC [FB81], Best-Bin-First (BBF) [BL97] have been used. Since a descriptor is basically a multi-dimensional histogram built in a special way, the Euclidean distance between descriptors is usually used, as it is a good indicator for a high probability match:

$$d(p,q) = \sqrt{\sum_{i=1}^{N}(p_i - q_i)}. \tag{3}$$

Here, p and q are two descriptors, p_i and q_i are the i-th elements of these descriptors, and N is the size of the descriptors. To find the matching features in the two data sets, I_1 and I_2, the Euclidean distances from each descriptor in I_1 to all descriptors in I_2 are computed. The minimum distance value is an indicator of a high probability match.

4. GPU-based Implementation

In 2007 Heymann et al. have proposed a GPU-based implementation of SIFT supporting real-time feature detection and feature matching for 2D images [HMS*07]. However, its extension to handle time-varying 3D volumetric data poses some additional challenges. As a result, within this section, we discuss and present our GPU-based implementation to address the issue when applying the SIFT algorithm to time-varying volumetric data sets.

Feature detection. During the construction of the scale-space representation of the input, the convolution operator is the most computational demand factor. Fortunately, the parallelism nature of GPU allows us to overcome this problem. For instance, the performance of the Gaussian convolution operator can be dramatically improved by using separable kernels. Additionally, the gradient calculation as well as the hyperspherical representation calculation can also be parallelized through the GPU-based approach. This enables us to improve the performance on the feature detection stage by a factor of 10 to 20.

Feature descriptor construction. In 2D, the neighborhood of size 8×8 is commonly use. The experimental findings from [Low04] show that the best results were achieved with a 4×4 array of histograms with 8 orientation bins in each which capture 45 degrees orientation differences. Consequently, each descriptor contains 128 elements. This allows a straightforward implementation of histogram calculation on the GPU that exploit the performance of the global and local memory architecture of the hardware.

In 3D, the neighborhood of a larger size, 16×16×16, is commonly used. The neighborhood is divided into 64 subregions of 4×4×4 voxels. Thus, a 1D representation of the descriptor contains 4096 elements. The size of the histogram

makes it difficult to have a GPU-based implementation of the histogram calculation process that exploits the performance of the memory architecture on the GPU. Moreover, it is worth noting that the size of the neighborhood needs to be adapted to the input data and the application in mind. Therefore, standard optimized histogram calculation on the GPU can not be easily adapted as the size of the histogram is large than the size local memory on the GPU. To overcome the hardware limitation, a multi-pass approach for histogram construction is required.

To avoid disruptive changes in the histogram of gradient orientations, the neighborhood around a detected feature location is usually divided into sub-regions. The histograms of gradient orientations of these sub-regions are first calculated and then combined together to form the final feature descriptor [Low04]. This poses a challenge to standard histogram calculation techniques on the GPU. Therefore, in this work, we implemented a generic OpenCL kernel that supports descriptor construction of an arbitrary neighborhood size. Moreover, the algorithm automatically switches to the optimized implementation based on the size of the descriptor.

In the initial SIFT operator, a smoothing operator is applied to the constructed histograms to include interpolation effects. In our implementation, we avoid this smoothing, and instead rotate the neighborhood around the feature locations to the dominant orientation. This enables us to achieve a bi-linear interpolation by default through a GPU-based implementation. With these modifications, we were able to realize an interactive GPU implementation with OpenCL, which runs ten times faster then previous approaches (see Section 5). Thus, we can not only use the SIFT algorithm for matching the different time steps on a global scale, but also can use it for interactive tracking of points of interest.

Feature identification. In the matching process, the Euclidean distances between each descriptor in the first input image and all the descriptors in the second one are calculated. Then the minimum Euclidean distance is identified to determine the highest probability match. The finding of the minimum distance is a reduction problem, which does not exploit the power of the parallelism nature of the GPU. As a result, we implemented a hybrid approach to the problem of feature identification. For instance, the computation of the Euclidean distance between one descriptor in the first input image and all the descriptors in the second image are performed in parallel using the GPU. This result is then passed to the CPU implementation to identify the minimum distance that serves as an indicator of a high probability match.

5. Test Case

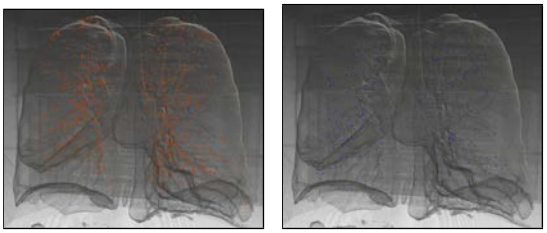
To show the impact of our introduced SIFT algorithm, we compare the results of our GPU-based implementation to the results achieved with the recent approach presented by Paganelli et al. [PPP*12]. We have tested our approach with the same data set, a 4D CT thorax scan[†] [CCG*09, CCZG09]. The data is given by ten equally sampled phases of the respiratory cycle in which the maximum exhale phase and the maximum inhale phase are denoted as $L0$ and $L5$ respectively. The reconstructed volumes have the dimensions of $512\times512\times128$ voxels of $0.97\times0.97\times2.5$ mm, while the reference landmarks in $L0$, and $L5$ were manually setup by an expert [CCZG09, CCG*09]. While the parameters used in our SIFT algorithm were set to match the ones used by Paganelli et al. [PPP*12], our approach allows more than one descriptors per detected feature location by considering orientations that are above 80% of the maximum peak in the histogram. In addition, due to the advantage of the GPU-based implementation, we do not apply smooth operator to the histogram to avoid disruptive changes of gradient orientations but instead rotate the neighborhood region to the dominant orientation, which implicitly takes advantage of the bi-linear interpolation on the GPU.

Feature location detection. We have evaluated the 4D CT SIFT matches, whereby we have computed the error as 3D residual distance between matching SIFT feature locations at the $L0$ and the $L5$ phase (SIFT $L0$–$L5$). Figure 2 illustrates the visualization of the inhale lung overlain with features of interest. While the detected features from the SIFT algorithm are colored as red spheres in Figure 3(a), the manually input reference landmarks from the data set are colored as blue spheres in Figure 3(b). The Mann-Whitney U test [MW47] was applied to this error distribution and the error distribution in the reference landmarks. Table 1 shows our results in comparison to the results reported in [PPP*12].

Besides the slightly higher number of matches between the maximum exhale and maximum inhale phase, the proposed approach has shown to improve the accuracy in the descriptor matching process, as we have achieved a lower median, 11.14 compared to 13.23, as well as a lower variability. In addition, the Mann-Whitney test confirms that the distributions of the residual distances in the reference landmarks and in the result of the enhanced SIFT operator are not significantly different (p-value = 0.736), which means that the proposed approach can be used to identify the feature locations.

Feature identification. To measure the impact on feature identification, we detect the feature locations in the time-series data and apply SIFT to consecutive volumes with two different variants. First, we always use the maximum exhale phase, $L0$, as a reference. Second, we move step-by-step along the breathing cycle such that the previous breathing phase is served as the reference for tracking the feature locations in the next breathing phase. For each approach, we computed the number of feature locations between all phases. Table 2 reports the number of detected feature locations which were preserved along the breathing cycle. While

[†] The data set is available at www.dir-lab.com.



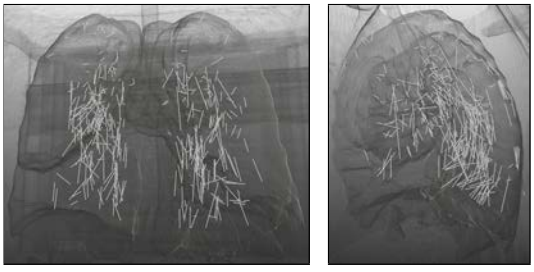
(a) (b)

Figure 2: *Visualization of the inhale lung. (*a*) is the visualization of the lung overlain with the detected features (in red) from the SIFT algorithm, (*b*) is the visualization of the inhale lung with the manually edited landmarks (in blue).*

	#Matches	Median (mm)	Variability (mm)
L0–L5	300	12.98	18.22
SIFT (L0–L5)	509	13.23	17.90
GPU-based SIFT (L0–L5)	525	11.14	12.78

Table 1: *Number of matches, median and variability of error distributions at maximum exhale, L0, and maximum inhale, L5, phases obtained by the proposed enhanced SIFT compared to the manual reference landmarks. Variability is the difference between the 25th and 75th percentiles.*

tracking along the breathing cycle allows us to achieve a higher number of preserved feature locations and excludes the trailing ones, tracking referred to a reference phase exclude the most stable feature locations over time. As seen in Table 2, the proposed approach provides more landmarks as feature locations, which makes it better suitable for motion estimation as well as visualization.



(a) (b)

Figure 3: *Visualization of the inhale lung overlain with the displacement vectors representing the transition of the detected features from L0 to L5. (*a*) is the visualization from the front, and (*b*) is the visualization from the side.*

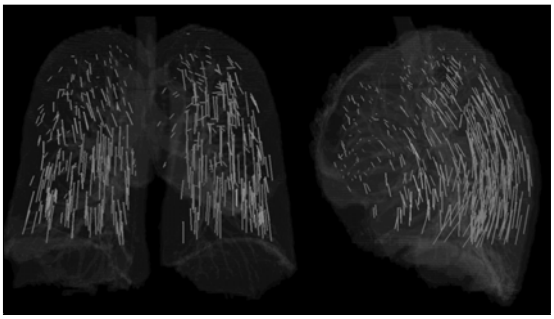


Figure 4: *Visualization of the inhale lung overlain with the displacement vectors representing the transition of the manually edited landmarks from L0 to L5 (courtesy of Castillo et al [CCG*09, CCZG09]).*

	SIFT (L0–L5)	GPU-based SIFT (L0–L5)
Reference (Phase 0)	117	243
Along breathing cycle	9	264

Table 2: *Number of preserved feature locations along the breathing cycle from the maximum exhale, L0, to the maximum inhale, L5, phase.*

As reported in Table 1 and Table 2, the proposed enhanced descriptor construction helps to increase the uniqueness of the descriptors at feature locations. Thus, it helps to improve the accuracy of the feature matching process in time-series data. It is also worth noting that by exploiting the power of the GPU implementation, we also achieved a much better performance in time. For instance, the overall time required for descriptors generation for each volume and descriptors matching between volumes is approximately 5.5 minutes. This is almost ten times faster than the result in [PPP*12], which is 50 minutes. As for the interactive visual analysis, we can precompute the feature descriptors, we can perform this process interactively.

Figure 3 shows the result of the proposed SIFT to track the detected features over the time domain. The displacement vectors (in white) show the transition of the detected features from the maximum inhale and exhale phases. Although the visual result is not very close to the visualization using manually input reference landmarks in Figure 4, the Man-Whitney test shows that there is no significant difference in residual distance distribution between the two results. As a result, this shows that the proposed SIFT is applicable to the application of automatic landmarks identification and tracking in time-varying dataset. This allows the proposed SIFT to serve as a tool for initial landmarks identification in different deformable image registration algorithms. Moreover, it can also be used to evaluate the results of different deformable image registration techniques.

6. Conclusion

In this paper, we presented a GPU-based implementation of the SIFT algorithm. By exploiting the power of the GPU, we do not only achieve better performance but also better results in comparison to the previously published work. The performance improvement of the algorithm enables us to investigate the effect of different parameters settings, such as σ in the scale-space construction, the level of the scale-space pyramid, to the quality of the detected features as well as the quality of the feature identification process in the future work. Furthermore, we would like to apply the proposed technique to the analysis of different dynamic data sets.

References

[AKB*08] ALLAIRE S., KIM J. J., BREEN S. L., JAFFRAY D. A., PEKAR V.: Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (2008), IEEE, pp. 1–8. 2, 3

[BL97] BEIS J., LOWE D.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (1997), pp. 1000–1006. 3

[CCG*09] CASTILLO R., CASTILLO E., GUERRA R., JOHNSON V. E., MCPHAIL T., GARG A. K., GUERRERO T.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine and Biology 54*, 7 (2009), 1849. 4, 5

[CCZG09] CASTILLO E., CASTILLO R., ZHANG Y., GUERRERO T.: Compressible image registration for thoracic computed tomography images. *Journal of Medical and Biological Engineering 29*, 5 (2009), 222–233. 4, 5

[CH07] CHEUNG W., HAMARNEH G.: n-SIFT: N-dimensional scale invariant feature transform for matching medical images. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on* (2007), IEEE, pp. 720–723. 2

[CH09] CHEUNG W., HAMARNEH G.: n-SIFT: n-dimensional scale invariant feature transform. *IEEE Transactions on Image Processing 18*, 9 (2009), 2012–2021. 2

[FB81] FISCHLER M. A., BOLLES R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM 24*, 6 (1981), 381–395. 3

[FBMB] FLITTON G., BRECKON T., MEGHERBI BOUALLAGU N.: Object Recognition using 3D SIFT in Complex CT Volumes. In *British Machine Vision Conference 2010*, British Machine Vision Association, pp. 11.1–11.12. 2

[HMS*07] HEYMANN S., MULLER K., SMOLIC A., FROHLICH B., WIEGAND T.: Sift implementation and optimization for general-purpose gpu. In *Proceedings of the international conference in Central Europe on computer graphics, visualization and computer vision* (2007), p. 144. 2, 3

[Lin94] LINDEBERG T.: Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics 21*, 1-2 (1994), 225–270. 2

[Low99] LOWE D. G.: Object recognition from local scale-invariant features. *Computer Vision, 1999, The Proceedings of the Seventh IEEE International Conference on 2* (1999), 1150–1157. 1

[Low04] LOWE D. G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2 (2004), 91–110. 1, 2, 3, 4

[MTS*05] MIKOLAJCZYK K., TUYTELAARS T., SCHMID C., ZISSERMAN A., MATAS J., SCHAFFALITZKY F., KADIR T., GOOL L. V.: A comparison of affine region detectors. *International journal of computer vision 65*, 1 (2005), 43–72. 2

[MW47] MANN H. B., WHITNEY D. R.: On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics 18*, 1 (1947), 50–60. 4

[NQY*08] NI D., QU Y., YANG X., CHUI Y. P., WONG T.-T., HO S. S., HENG P. A.: Volumetric Ultrasound Panorama Based on 3D SIFT. In *MICCAI '08: Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part II* (Sept. 2008), Springer-Verlag. 2

[PPP*12] PAGANELLI C., PERONI M., PENNATI F., BARONI G., SUMMERS P., BELLOMI M., RIBOLDI M.: Scale invariant feature transform as feature tracking method in 4d imaging: A feasibility study. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (2012), pp. 6543–6546. 2, 4, 5

[SAS07] SCOVANNER P., ALI S., SHAH M.: A 3-dimensional SIFT descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia* (2007), ACM, pp. 357–360. 2

[TWICA10] TOEWS M., WELLS III W., COLLINS D. L., ARBEL T.: Feature-based morphometry: Discovering group-related anatomical patterns. *NeuroImage 49*, 3 (2010), 2318–2327. 2

[YWC12] YU T.-H., WOODFORD O. J., CIPOLLA R.: A Performance Evaluation of Volumetric 3D Interest Point Detectors. *International Journal of Computer Vision 102*, 1-3 (Sept. 2012), 180–197. 2