An Evaluation of Volumetric Interest Points

Tsz-Ho Yu
University of Cambridge
Cambridge, UK
thy23@cam.ac.uk

Oliver J. Woodford
Toshiba Research Europe Ltd.
Cambridge, UK
oliver.woodford@crl.toshiba.co.uk

Roberto Cipolla
University of Cambridge
Cambridge, UK
cipolla@eng.cam.ac.uk

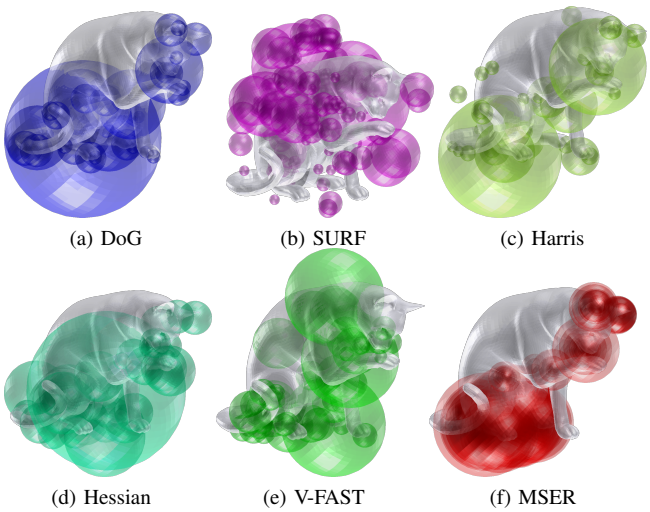
***Abstract*—This paper presents the first performance evaluation of interest points on scalar volumetric data. Such data encodes 3D shape, a fundamental property of objects. The use of another such property, texture (*i.e.* 2D surface colouration), or appearance, for object detection, recognition and registration has been well studied; 3D shape less so. However, the increasing prevalence of depth sensors and the diminishing returns to be had from appearance alone have seen a surge in shape-based methods. In this work we investigate the performance of several detectors of interest points in volumetric data, in terms of repeatability, number and nature of interest points. Such methods form the first step in many shape-based applications. Our detailed comparison, with both quantitative and qualitative measures on synthetic and real 3D data, both point-based and volumetric, aids readers in selecting a method suitable for their application.**

***Keywords*-Volumetric interest points; performance evaluation;**

I. Introduction

The applications of object detection, recognition and registration are of great importance in computer vision. Much work has been done in solving these problems using appearance, helped by the advent of image descriptors such as SIFT and competitions such as the PASCAL challenge, and these methods are now reaching maturity. However, advancing geometry capture techniques, in the form of stereo, structured light, structure-from-motion and also sensor technologies such as laser scanners, time-of-flight cameras, MRIs and CAT scans, pave the way for the use of shape in these tasks, either on its own or complementing appearance—whilst an object's appearance is a function not only of its texture, but also its pose and lighting, an object's 3D shape is invariant to all these factors, providing robustness as well as additional discriminative power.

Shape-based techniques for object recognition range from the very local, with small, indistinct features, *e.g.* single points, combined in a geometrical consistency framework, to the global, with a single descriptor for an entire object. While the former methods offer robustness to occlusion and clutter and provide pose, they do not cope well with shape variability, nor scale well with the number of object classes. The global methods' pros and cons tend to be the inverse. Between these two extremes are those methods that describe local features of limited but sufficiently distinctive scope, thereby gaining the discriminability of global methods *and* the robustness of local methods. The distribution of such features can be used for effective object detection, recognition and registration. The nature of these hybrid methods is reminiscent of the image descriptors of appearance-based methods, not least in the need for shape features to be chosen at points that are repeatably locatable in different data sets, and whose localities are distinctive. A common and crucial stage of such approaches is therefore the detection of *interest points* to be described.



(a) DoG (b) SURF (c) Harris

(d) Hessian (e) V-FAST (f) MSER

Figure 1: Different types of volumetric interest points detected on a test shape.

This paper aims to provide the first performance evaluation of interest point detectors (shown in figure 1) on scalar volumetric data. Such data not only comes directly from volumetric sensors, *e.g.* MRIs, but can also be generated from other 3D data such as point clouds or meshes, making the evaluation as widely applicable as possible. The representation allows interest points to be located off an object's surface, *e.g.* at the centre of a limb. In addition, the nature of the data—voxels, the 3D equivalent of pixels—makes repurposing the many 2D interest point detectors for 3D straightforward. The primary, quantitative evaluation criterion is a novel measure combining both repeatability,

based on the number of corresponding points found across two volumes, and the spatial accuracy of correspondences. Detected interest points have a (subvoxel) location and also a scale, and distances are computed in this space.

The remainder of this section reviews previous work relevant to interest point detectors and our evaluation framework. The following section introduces the interest point detectors we compared, while section III describes our evaluation methodology. Section IV describes our experiments and presents the results, before we conclude in section V.

A. Previous work

We will now discuss relevant prior work on interest point detectors and evaluation methodologies.

Interest point detectors Interest points are typically used to match corresponding points across two or more similar sets of data. The data has predominantly been images, though some detectors have recently been applied to volumes. This review will be brief, focusing on the detectors we evaluate and previous applications to 3D data, as the field is well studied (at least in respect to images); we refer the reader to a recent survey [1] for more details.

Differential approaches to interest point detection find optima in the data response to differential operators such as the second moment matrix (first order derivatives), *e.g.* the Harris corner detector [2], and the Hessian matrix (second order derivatives). Lindeberg [3] studied scale-covariant[1] points using the Laplacian-of-Gaussian kernel (equivalent to the trace of the Hessian), as well as the determinant of the Hessian. Lowe [4] approximated the former with a Difference-of-Gaussians (DoG) operator for efficiency, a method recently applied to CT scan volumes [5]. The SURF detector of Bay *et al.* [6] accelerates computation of the Hessian determinant through the use of integral images, since applied to integral volumes of space-time video [7] and binary volumes generated from 3D meshes [8]. Mikolajczyk & Schmid [9] developed the scale-covariant Harris-Laplace detector by finding Harris corners in the spatial domain which are maxima of the Laplacian in the scale domain, an approach which Laptev [10] applied to classification of video volumes.

In a different approach, the FAST detector of Rosten *et al.* [11] uses the largest number of contiguous pixels on a circle of fixed radius which are significantly darker, or brighter, than the centre pixel. They learn a decision tree classifier that makes the approach extremely efficient, leading it to be applied to space-time video volumes also [12], [13]. The MSER detector of Matas *et al.* [14] finds thresholded regions whose areas are maximally stable as the threshold changes. It is therefore inherently multi-scale, as well as invariant to affine intensity variations and covariant with affine transformations. MSER has already been applied to volumetric data, firstly in the context of segmentation of MRIs [15], then on spatio-temporal data [16].

Image-based detectors have also been made affine-covariant (*e.g.* [9]), in order to approximate the perspective distortion caused by projection of 3D surfaces onto the 2D image plane. Such covariance is not necessary with 3D shape data because most shape acquisition techniques are invariant to view point changes, thus affine transformations are not common among datasets.

Methodologies Empirical performance evaluation is a popular pastime in Computer Vision, and the topic of interest point[2] detection is no exception. Some methods evaluate performance in the context of a particular task, lacking generality to other applications. Most investigate one or more interest point characteristics. One such characteristic, important for registration applications, *e.g.* camera calibration, scene reconstruction and object registration, is the accuracy of interest point localization. Coelho *et al.* [17] measure accuracy by computing projective invariants and comparing these with the actual values, measured from the scene. Brand & Mohr [18] introduce three further measures, including 2D Euclidean distance from detected points to ground truth corner locations (given by line fitting to a grid). This has been extended to using the distance to the nearest point detected in, and transformed from, another image (*e.g.* [19]). Matching scores for interest points found over location and *scale* [9], and also *affine* transformations [9], have since been proposed.

When used for object detection and recognition, two other important characteristics of interest points are their repeatability (*i.e.* is the same feature detected in multiple datasets) and their distinctiveness (*i.e.* can a point, or region of interest be easily recognized among many). Repeatability was proposed by Schmid *et al.* [19] as the ratio of repeated points to detected points. Rosten *et al.* [11] used the area under the repeatability curve as a function of number of interest points, varied using a threshold on the detector response, in their evaluation. When ground truth locations of interest points are given, *e.g.* hand labelled, an alternative measure is the ROC curve [20], which takes into account false matches. Schmid *et al.* [19] also introduce a quantitative measure of distinctiveness—entropy, or "information content". Alternatively, qualitative visual comparison is used (*e.g.* [3], [10]), on test datasets containing a variety of different interest points.

In the context of texture-based interest points, performance of detectors are often measured over variations in image rotation, scale, viewpoint angle, illumination and noise level (*e.g.* [19] covers all these factors), as well as corner properties. Efficiency may be a further consideration [11].

[1] Covariant characteristics, often (inaccurately) referred to as *invariant* characteristics, undergo the same transformation as the data. We prefer "covariant" in order to distinguish truly invariant characteristics.

[2] When referring to interest points in the context of methodology, we include image features such as corners, lines, edges and blobs.

II. Detectors

This section briefly describes the volumetric interest point detectors that we will evaluate: difference of Gaussians (DoG) [5], Harris-based [10] and Hessian-based interest points, speeded up robust features (SURF) [7], [8], V-FAST [13] and maximally stable extremal regions (MSER) [15], [16].

DoG Difference of Gaussians is a blob detection technique for feature localization popularized by the SIFT algorithm [4]. Its volumetric extension has been applied to shape recognition [5]. DoG approximates the Laplacian of Gaussian filter, which enhances features of a particular size. The DoG saliency response S_{DoG} is computed by subtracting two Gaussian smoothed volumes, usually adjacent scale-space representations, of the same signal and taking the absolute values of this. Interest point are detected at the *4D local maxima* (both 3D space and scale) in S_{DoG} within each octave of $\mathbf{V}(\mathbf{x}, \sigma_s)$.

Harris Extending the Harris corner detector [2], Laptev introduces the space-time interest point for video categorization [10]. The second-moment matrix is computed by smoothing the first derivatives of the volume in scale-space $\mathbf{V}(\mathbf{x}; \sigma_s)$ by a spherical (in our case) Gaussian weight function $g(\cdot; \sigma_{\text{Harris}})$:

$$\mathcal{M} = g(\cdot; \sigma^2_{\text{Harris}}) * \begin{bmatrix} \mathbf{V}_x^2 & \mathbf{V}_x\mathbf{V}_y & \mathbf{V}_x\mathbf{V}_z \\ \mathbf{V}_x\mathbf{V}_y & \mathbf{V}_y^2 & \mathbf{V}_y\mathbf{V}_z \\ \mathbf{V}_x\mathbf{V}_z & \mathbf{V}_y\mathbf{V}_z & \mathbf{V}_z^2 \end{bmatrix}, \quad (1)$$

where $\mathbf{V}_x$,$\mathbf{V}_y$,$\mathbf{V}_z$ denote the partial derivatives of the volume in scale-space $\mathbf{V}(\mathbf{x}; \sigma_s)$ along x, y and z axes respectively. The saliency S_{Harris} is computed from the determinant and trace of $\mathcal{M}$:

$$S_{\text{Harris}} = \sigma_s^3 \det(\mathcal{M}) - k\,\text{trace}(\mathcal{M})^3, \quad (2)$$

where k controls the rejection of edge points. Each saliency response S_{Harris} is normalized by its scale σ_s. The window size σ_{Harris} is proportional to expected feature scales σ_s by a factor of 0.7 as suggested in [9]. Interest point are the 4D local maxima in the scale-space of S_{Harris}.

Hessian The Hessian interest point is similar to the Harris detector; instead of a second-moment matrix $\mathcal{M}$, it is based on Hessian matrix $\mathcal{H}$.

$$\mathcal{H} = \begin{bmatrix} \mathbf{V}_{xx} & \mathbf{V}_{xy} & \mathbf{V}_{xz} \\ \mathbf{V}_{yx} & \mathbf{V}_{yy} & \mathbf{V}_{yz} \\ \mathbf{V}_{zx} & \mathbf{V}_{zy} & \mathbf{V}_{zz} \end{bmatrix}, \quad (3)$$

where $\mathbf{V}_{xy}$ denotes the second derivative of the volume at scale σ_s, along x and y axes: $\frac{\partial \mathbf{V}(\mathbf{x};\sigma_s)}{\partial x \partial y}$. The Hessian saliency is the determinant of $\mathcal{H}$ normalized by scale σ_s:

$$S_{\text{Hessian}} = \sigma_s^3 \det(\mathcal{H})\,. \quad (4)$$

Similarly, the interest points are located at the 4D scale-space maxima of S_{Hessian}.

SURF Speeded up robust features (SURF) [6] is a feature extraction algorithm optimized for efficiency. SURF is similar to Hessian in terms of the feature model it is based on; Haar wavelets are used to approximate derivatives of Gaussian functions, accelerating computation. The first 3D version of SURF was introduced in Willems *et al.* [7] for video classification. Recently, it was used in a 3D shape object recognition task [8].

V-FAST Building on the success of FAST corner detector [11], Yu *et al.* [13] proposed V-FAST for video recognition. Koelstra & Patras [12] introduced an alternative FAST-3D interest point. The V-FAST algorithm performs accelerated segment tests (AST) on three orthogonal Bresenham circles. Along each plane, the saliency score is computed by maximizing the threshold t that makes at least n contiguous voxels brighter or darker than the nucleus voxel by t in equation 5.

$$AST_{xy}(n,t) = \begin{cases} t & \text{if } ||v_{\text{nucleus}} > \mathbf{c}_{xy} + t|| \geq n \\ t & \text{if } ||v_{\text{nucleus}} < \mathbf{c}_{xy} - t|| \geq n \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$S^{xy}_{\text{vfast}} = \max(AST_{xy}(n,t))\,, \quad (6)$$

$$S_{\text{vfast}} = \sqrt{(S^{xy}_{\text{vfast}})^2 + (S^{xz}_{vfast})^2 + (S^{yz}_{\text{vfast}})^2}\,, \quad (7)$$

where $\mathbf{c}_{xy}$ denotes the voxels on an xy-circle centered at c_{nucleus}. The overall saliency S_{vfast} is the Euclidean norm of saliency scores on the three planes in equation 7. Interest points are local maxima in S_{vfast} along both space and scale, with at least two non-zero values in $AST_{xy}(n,t)$, $AST_{xz}(n,t)$ or $AST_{yz}(n,t)$.

MSER Maximally stable extremal regions (MSER) is a blob detection technique proposed by Matas *et al.* [14]. Being inherently advantageous for volumetric interest point detection (*e.g.* robust to rotation and scale changes), MSER has already been applied for detecting volumetric salient regions [15], [16].

Subvoxel refinement MSER locates interest points by fitting an ellipsoid to the detected salient region [14]. We apply the subpixel refinement process of Lowe [4] to all the other detectors; interest points are localized at the subvoxel level by fitting a 4D quadratic function around the local scale-space maxima, and selecting the maxima of those functions instead.

III. Methodology

Whilst previous evaluations have focused on either localization accuracy or repeatability, we combine the two performance metrics into a single score.

Localization accuracy A key component of the score is the distance metric used to measure the closeness of two interest points.

$$D(\mathbf{P}, \mathbf{Q}') = ||\mathbf{P} - \mathbf{Q}'||_2\,, \quad (8)$$

$$\mathbf{P} = [P_x,\ P_y,\ P_z,\ f\log(P_s)]^\mathsf{T}\,, \quad (9)$$

where P_x, P_y and P_z are the spatial components and P_s is the scale component of an interest point's location. The log of scale is used to remove multiplicative bias across detectors, and since spatial location and scale are not fully commensurable, a parameter f is introduced to balance the importance of scale to the distance function.

In our evaluation the interest point $\mathbf{P}$ is in the coordinate frame of the volume, V_P, it was found in, whilst $\mathbf{Q}'$ is the point $\mathbf{Q}$ found in volume V_Q of a given pair of volumes, transformed into the coordinate frame of V_P. Our score is based on the distance of an interest point to the nearest transformed interest point, thus:

$$D(\mathbf{P}, \mathcal{Q}') = \min_{\mathbf{Q}'_j \in \mathcal{Q}'} D(\mathbf{P}, \mathbf{Q}'_j) \,, \tag{10}$$

where $\mathcal{Q}' = \{\mathbf{Q}'_j\}_{j=1}^{q}$, the set of q transformed interest points found in V_Q.

Repeatability Schmid *et al*. [19] defined repeatability as the ratio of correspondences to points:

$$R_{\text{ratio}}(\mathcal{P}, \mathcal{Q}', d) = \frac{\sum_{i=1}^{p} H(d - D(\mathbf{P}_i, \mathcal{Q}'))}{\min(p, q)} \,, \tag{11}$$

where $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^{p}$, the set of p interest points found in V_P, d is a user provided distance threshold, and $H(\cdot)$ is the Heaviside step function.

Combined score Rosten *et al*. [11] computed the area under R_{ratio} as a function of the number of interest points (varied using a contrast threshold on the detector). We use the same idea, but computing R_{ratio} as a function of the distance threshold, d. The score therefore increases both if more points are matched, and also if matches are more accurate. We also compute a symmetric score, by measuring the average score across two directions of transformation , using V_P and V_Q as the reference frames respectively. This score is given as

$$R_{\text{area}} = \frac{1}{2D} \int_0^D R_{\text{ratio}}(\mathcal{P}, \mathcal{Q}', d) + R_{\text{ratio}}(\mathcal{Q}, \mathcal{P}', d) \, \mathrm{d}d, \tag{12}$$

where D is a maximum distance threshold.

IV. Evaluation

A comprehensive evaluation is performed on the volumetric interest point detectors in this section. The main objective is to investigate how the interest points perform under different variations. Our R_{area} score is advantageous over traditional repeatability, as it reflects both repeatability and accuracy of interest points in one measurement.

A. Test data

Three sources of data are used in our evaluation. Two of these are synthetic, as large sets of real, registered, 3D data are not commonly available. Synthetic data are used because we can create new test data with arbitrary noise levels, transformations and sampling density with accurate ground-truths for evaluation. The first set, **Mesh**, contains 25 shapes (surface meshes) chosen from the *Princeton shape database* [21] and *TOSCA* [22] sets. This selection contains a wide range of geometric features, from coarse structures to fine details. Point clouds are created by randomly sampling 3D points, with a uniform distribution, over the area of the meshes. Homogeneous Gaussian noise is added to the points. The point clouds are then *voxelized* using kernel density estimation with a Gaussian kernel $g(\cdot, \sigma_{KDE})$. The conversion process, illustrated in figure 3, allows sampling density and noise to be varied when generating the volumes. The **MRI** dataset consists of two synthetic MRI scans of a human brain, generated from BrainWeb [23], again with known transformation between the two. The **Stereo** dataset contains 16 point clouds of 8 objects captured from a variety of viewpoints using a multi-view stereo system [24]. Relative transformations are computed by aligning each point cloud with a reference model using the *iterative closest point* algorithm [25]. The same voxelization technique is used to convert stereo point clouds to volumetric data.

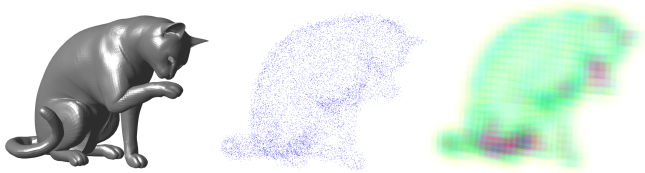


Figure 3: **Mesh to volume conversion**: *Left to right* – mesh, point cloud and voxel array.

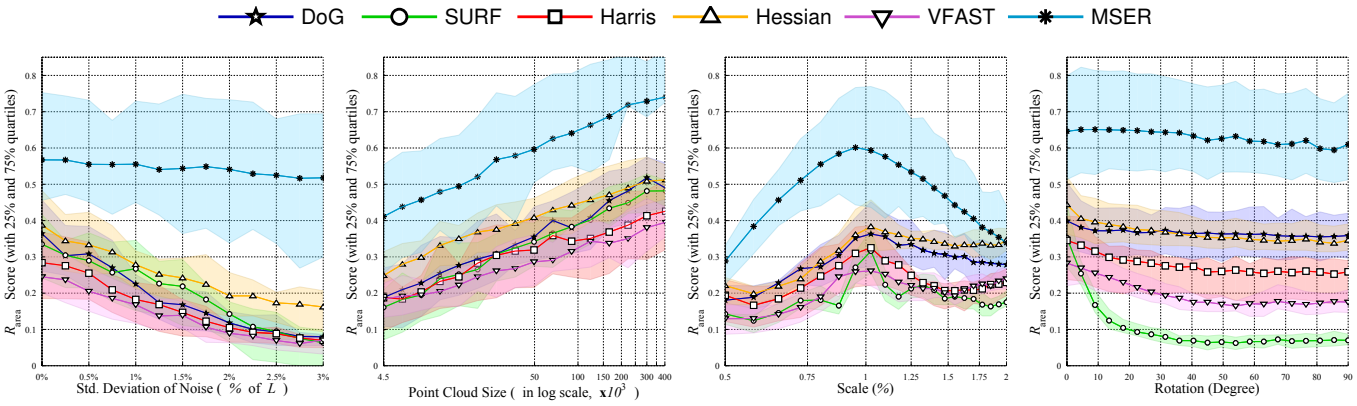
While the synthetic shape instances of the same object completely overlap one another, avoiding bias to the repeatability score [26], the real stereo data contains occlusions (the underside of each object, which varied across instances, was not captured), as well as uneven sampling density and generally more sampling noise. The applicability to real applications of our performance evaluation using synthetic data will therefore be tested by comparing the results with those on the **Stereo** dataset.

B. Experimental setup

We evaluate the robustness of the detectors under variations of several test parameters: rotation, translation, scale, sampling density and noise. The three datasets in section IV-A are tested under different variations:

- **Mesh**: Sampling noise and density, scale, rotation.
- **MRI**: Rotation + translation.
- **Stereo**: Scale + rotation + translation.

Performances of the candidate detectors are measured as each test parameter is varied individually, keeping all the other parameters at their default values (given in table I). Sampling parameters (*i.e*. noise level and sampling density) are applied to all shape instances, whilst pose parameters (*i.e*. rotation, translation and scale) are applied to only



(a) Sampling Noise (b) Sampling Density (c) Scale (d) Rotation

Figure 2: R_{area} scores of testing dataset under changing (a) *sampling density (point cloud size)*, (b) *scaling* and (c) *rotation*. Solid lines indicate the average, and the light colour bands show the 25% and 75% quartiles.

one instance in each matching pair. Some parameters are defined in terms of L, the largest dimension of the voxelized reference shapes. We set the maximum value of L to 200 voxels.

C. Experiments on synthetic meshes

Sampling noise Sampling noise and density are crucial factors in shape-based interest point detection. Existing shape acquisition techniques (*e.g.* multi-view stereo) often produce data with sampling noise. In this test, different levels of Gaussian white noise, with standard deviations 0–0.03L, are applied to the **Mesh** dataset.

The result is shown in figure 2a. While the R_{area} scores of other detectors decline rapidly, MSER maintains a high R_{area} score, demonstrating high robustness to noise, since it does not use derivatives, which are vulnerable to it. The Hessian-based detector shows a relatively stronger tolerance than the remaining detectors at higher noise levels.

Sampling density The R_{area} score of shapes with various sampling densities (point clouds range from 4K points to 405K points) are measured. Figure 2b presents the change of R_{area} scores versus point cloud size. The scores vary linearly in log scale, therefore a diminishing return is observed with increasing sampling density. MSER achieves the best average performance but it also has the largest variance across different shapes. Hessian and DoG produce satisfactory results, with high scores but smaller intra-dataset variance than that of MSER.

Rotation This experiment evaluates susceptibility of the detectors to rotational aliasing effects. Eight rotation axes are generated randomly for each shape, and rotations of increasing magnitude, up to 90°, applied about them.

The effect of rotation is shown in figure 2d. Most detectors show excellent tolerance to rotation, inheriting this from their texture-based counterparts. DoG and MSER perform slightly better than others, with almost unchanged average scores. SURF performs worse than other detectors because the use of box filters leads to increased quantization errors as the shapes are rotated.

Parameter	Value
Default point cloud size	50000 points
Default noise	$0.0025L$
Default rotation	0°
Maximum L	200 voxels
Default σ_{KDE} in $g(\cdot, \sigma_{\text{KDE}})$	1.5 voxels
Distance threshold D	$0.03L$
Parameter f in equation (8)	$\sqrt{8}$
Number of octaves in scale-space	4

Table I: The reference parameters for the testing shapes.

Scale Data are scaled from 50% to 200% of their original sizes. The values of R_{area} measured against scale changes are illustrated in figure 2c. DoG and Hessian detectors are comparatively more robust to scale. SURF only works well at 100% and drops significantly outside the original scale, because of the approximated scale-space used. MSER achieves the best result at its original size, its performance decreasing steadily as the shape is scaled. The loss of information in smaller volumes leads to a greater performance drop at smaller scales than larger scales for all detectors.

Number of correspondences Table II presents quantitative statistics for the number of interest points and correspondences at three noise levels (0.0025, 0.01 and 0.02 of L). The MSER detector has the highest percentage of correspondences, yet it gives a smaller set of interest points. By contrast, Hessian, SURF and Harris produce larger sets of interest points with good correspondence ratios. The displacement threshold used here ($D = 0.015L$) is about half the typical value, hence only accurate correspondences are counted towards the ratios.

Saliency Figure 5 shows the repeatability, R_{ratio}, with varying percentages of interest points. For each detector, the

Avg.# Pts. Avg.# Corr.Pts. (*Corr. %*)	**Sampling Noise Level** **Low** (0.0025L)	**Medium** (0.01L)	**High** (0.02L)
DoG	122.0 48.8 (*39.8%*)	118.2 35.1 (*29.7%*)	73.1 9.3 (*12.7%*)
SURF	154.7 54.7 (*35.3%*)	70.4 18.4 (*26.2%*)	28.7 3.84 (*13.4%*)
Harris	303.3 78.6 (*25.9%*)	142.2 33.2 (*23.3%*)	123.8 13.4 (*10.83%*)
Hessian	**330.8** **117.1** (*35.4%*)	**272.0** **72.2** (*26.6%*)	**201.2** **30.2** (*18.2%*)
V-FAST	115.9 33.5 (*28.8%*)	85.5 15.5 (*18.10%*)	74.6 7.4 (*9.85%*)
MSER	99.0 59.9 (**60.5**%)	74.4 44.7 (**60.2**%)	52.5 28.8 (**54.9**%)

Table II: *For each entry, top to bottom:* The average number of interest points detected, the average number of correspondences ($d \leq 0.015L$), percentage of correspondence points.

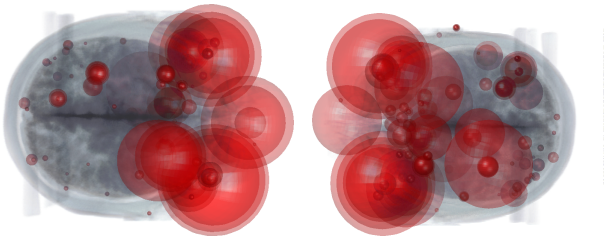
detected interest points are sorted by their corresponding saliency responses in descending order, then the first $p\%$ of interest points for computing R_{ratio}. However, since no saliency measure is defined in the MSER detector, the number of detected interest points cannot be controlled directly; MSER is therefore not included in figure 5. For analyzing the accuracies of the candidate detectors, R_{ratio} is computed using a smaller displacement threshold ($D = 0.015L$) in this experiment.

The performance of DoG and Harris detectors tend to be stable (though Harris performs notably worse than DoG) with increasing numbers of interest points (*i.e.* decreasing saliency threshold), indicating that a saliency threshold is not necessary for these detectors. Hessian's performance, which is initially the best, decreases very slowly and converges with DoG and SURF, indicating that the lower saliency points are less reliable. A saliency threshold for Hessian might benefit applications requiring more accurate point localization. By contrast, the reapeatability scores of SURF and VFAST increases steadily before leveling off, suggesting that some of the high saliency points are unreliable; this poses more of a problem in terms of interest point selection.

D. Experiments on MRI and stereo data

Interest points detectors are also evaluated on the **MRI** and **Stereo** datasets. As a reference for comparison, average scores obtained from the **Mesh** dataset are plotted against displacement threshold D in figure 6a.

MRI dataset R_{area} scores measured on this dataset with varying D are shown in figure 6b. The evaluation results obtained are comparable to that of synthetic mesh data—MSER, DoG and Hessian work slightly better in synthetic meshes, while the Harris detector is good at detecting complicated internal structures in the MRI scans.



(a) MRI scan 1 (b) MRI scan 2

Figure 4: Two volumetric MRI scans of a human brain, with detected MSER features.

Figure 4 shows the MSER interest points detected in the data. It is worth noting that some points detected on the MRIs can be matched easily, such as those on the nose-tips, eye sockets and foreheads, but that there are fewer detections within the brain area.

Stereo dataset The R_{area} scores obtained from the **Stereo** dataset, shown in figure 6c, are lower compared with **Mesh** and **MRI** datasets, especially at small D. Nonetheless, in terms of overall rankings and relative scores of the detectors, our synthetic and real data demonstrate similar behaviour. The decrease in performance for our stereo data could be due to its: *(a)* low sampling frequency and high noise, *(b)* uneven object surfaces, which are infeasible for blob detection algorithms (*e.g.* MSER, SURF and Hessian) and *(c)* small errors in the estimated ground truth poses.

Figure 7 shows two example point clouds and their corresponding volumes, with detected DoG, Harris and SURF features (figures 7b, 7c and 7d respectively). One object exhibits a much sparser reconstruction (due to a lack of texture on the object's surface), but it is interesting to note that the distribution of detected features is no less dense across any of the detectors, suggesting that our synthetic results are representative of sparse point clouds as well as dense.

E. Qualitative analysis

Volumetric interest points can be roughly classified into three categories: region-based blob detection (MSER), derivative-based blob detection (DoG, Hessian and SURF) and corner detection (Harris, V-FAST). The quantitative evaluation results imply that region-based blob detectors work better than derivative-based blob detectors, and blob detectors are better than corner detectors, but this is not the whole story. The candidate detectors demonstrate different behaviours in terms of locations and scales of the detected interest points. Therefore, besides repeatability, it is also important to analyze the characteristics of detectors qualitatively. Figure 1 visualizes interest points detected by the six candidate detectors on the "cat" object from the **Mesh** dataset.

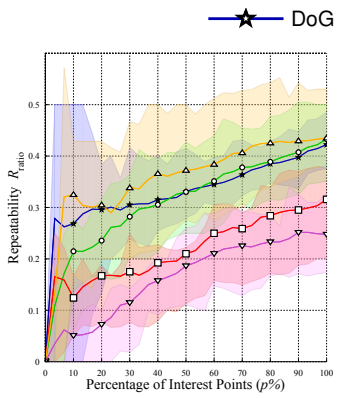


Figure 5: Repeatability R_{ratio} *vs.* percentage of interest points p.

(a) Repeatability Score: *Mesh* (b) Repeatability Score: *MRI* (c) Repeatability Score: *Stereo*

Figure 6: R_{area} scores versus displacement threshold d: (a) **Mesh** dataset, (b) **MRI** dataset and (c) **Stereo** dataset.

Region-based blob detection MSER detects contiguous regions of any shape (*i.e.* not limited to spherical blobs) allowing it to select more robust regions from a greater selection, and hence perform better. It can be seen in figure 1f that MSER finds features at fewer locations, but over multiple scales. The locations tend to center on regions of high surface curvature.

Derivative-based blob detection DoG, Hessian and SURF detectors theoretically find, in order of preference, spherical blobs, corners and planes. DoG and Hessian have qualitatively similar output, as shown in figures 1a and 1d, finding features at limb extremities such as paws, ears and tail, as well as within limbs. By contrast, SURF (figure 1b), despite being an approximation of the Hessian detector, produces features off the surface (both inside and outside), often over regions of low surface curvature.

Corner detection Harris and V-FAST both aim to find areas of high curvature. However, their outputs (figures 1c and 1e) vary qualitatively, with the former tending to find fewer large features and more sharp corners than the latter, which finds an even distribution of features over both scale and translation.

V. Conclusion

In this paper, we evaluated the state of the art for volumetric interest points. Combining both repeatability and accuracy, we proposed a novel metric to evaluate a feature detector. Three types of input data (meshes, MRI and stereo point clouds) are leveraged in the evaluation framework to give a complete picture of different detectors. Summarizing the results with respect to the proposed R_{area} score, MSER achieves the best overall performance, being robust to both noise and rotation. Taking the number of corresponding points into account, Hessian and, to a lesser extent, DoG maintain a balanced performance between this and repeatability. From the experiments, blob detectors (*e.g.* Hessian) appear to perform better than corner detectors (*e.g.* Harris) in 3D shapes, a result that agrees with an evaluation of texture-based detectors [27]. Nevertheless, the choice of volumetric interest points can also be application dependent. This work seeks to provide guidance on selection of interest point algorithm for specific computer vision or machine learning tasks. In a qualitative analysis we discussed the nature of features found by each detector. In addition we have demonstrated that our extensive evaluation on synthetic data can aid detector selection for real applications.

References

[1] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, pp. 177–280, 2008. 2

[2] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc., 4th Alvey Vision Conference*, 1988. 2, 3

[3] T. Lindeberg, "Feature detection with automatic scale selection," *Intl. Jour. of Computer Vision*, 1998. 2

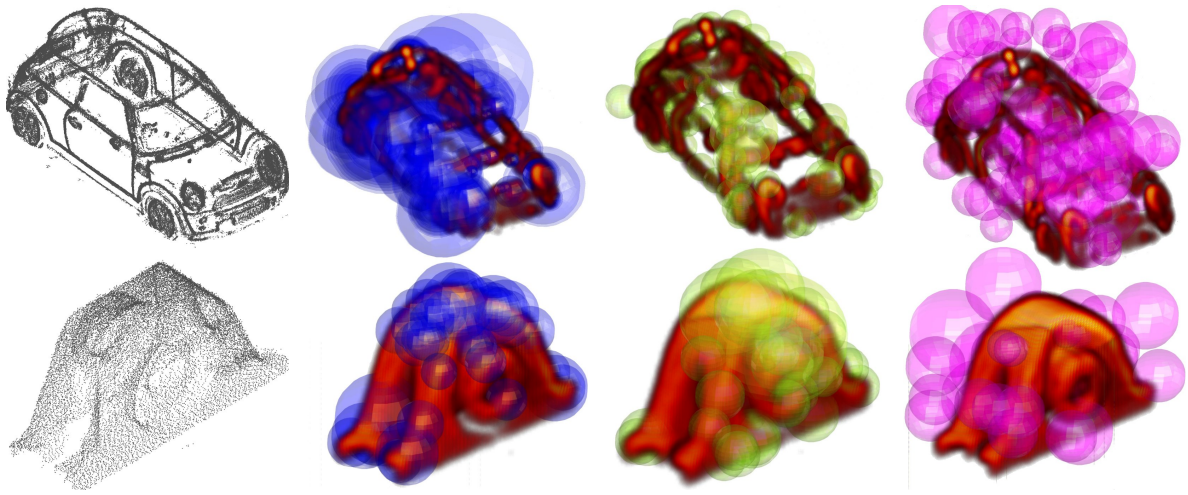
[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Jour. of Computer Vision*, 2004. 2, 3

[5] G. Flitton, T. Breckon, and N. Megherbi Bouallagu, "Object recognition using 3D sift in complex CT volumes," in *British Machine Vision Conf.*, 2010. 2, 3

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, 2008. 2, 3

[7] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conf. on Computer Vision*, 2008. 2, 3

Acknowledgements. We thank the Computer Vision Group of Toshiba Research Europe Ltd. for providing the multi-view stereo data. Tsz-Ho Yu is funded by the Croucher Foundation.



(a) Input point clouds (b) DoG (c) Harris (d) SURF

Figure 7: (a) Two sample point clouds obtained from the **Stereo** dataset. (b) DoG, (c) Harris and (d) SURF interest points visualized on the voxelized data.

[8] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool, "Hough transform and 3D SURF for robust three dimensional classification," in *European Conf. on Computer Vision*, 2010. 2, 3

[9] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Intl. Jour. of Computer Vision*, 2004. 2, 3

[10] I. Laptev, "On space-time interest points," *Intl. Jour. of Computer Vision*, 2005. 2, 3

[11] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *Trans. on Pattern Analysis and Machine Intelligence*, 2010. 2, 3, 4

[12] S. Koelstra and I. Patras, "The FAST-3D spatio-temporal interest region detector," in *Proc., WIAMIS '09*, 2009. 2, 3

[13] T. H. Yu, T. K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forest," in *British Machine Vision Conf.*, 2010. 2, 3

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, 2004. 2, 3

[15] M. Donoser and H. Bischof, "3D segmentation by maximally stable volumes (MSVs)," in *Intl. Conf. on Pattern Recognition*, 2006. 2, 3

[16] H. Riemenschneider, M. Donoser, and H. Bischof, "Bag of optical flow volumes for image sequence recognition," in *British Machine Vision Conf.*, 2009. 2, 3

[17] C. Coelho, A. Heller, J. L. Mundy, D. A. Forsyth, and A. Zisserman, "An experimental evaluation of projective invariants," in *Geometric invariance in computer vision*, J. L. Mundy and A. Zisserman, Eds. Cambridge, MA, USA: MIT Press, 1992, pp. 87–104. 2

[18] P. Brand and R. Mohr, "Accuracy in image measure," in *Videometrics III*, S. F. El-Hakim, Ed., vol. 2350. SPIE, 1994, pp. 218–228. 2

[19] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *Intl. Jour. of Computer Vision*, 2000. 2, 4

[20] K. Bowyer, C. Kranenburg, and S. Dougherty, "Edge detector evaluation using empirical ROC curves," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1999. 2

[21] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The Princeton shape benchmark," in *Shape Modeling International*, 2004. 4

[22] A. Bronstein, M. Bronstein, and R. Kimmel, *Numerical Geometry of Non-Rigid Shapes*, 1st ed. Springer, 2008. 4

[23] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans, "BrainWeb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, 1997. 4

[24] G. Vogiatzis and C. Hernández, "Video-based, real-time multi view stereo," *Image and Vision Computing*, In Press, Accepted Manuscript, 2011. 4

[25] P. Besl and N. McKay, "A method for registration of 3-D shapes," *Trans. on Pattern Analysis and Machine Intelligence*, 1992. 4

[26] A. Willis and Y. Sui, "An algebraic model for fast corner detection," in *IEEE Intl. Conf. on Computer Vision*, 2009. 4

[27] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Intl. Jour. of Computer Vision*, 2005. 7