Computer Vision and Image Understanding 157 (2017) 117–137

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A generalised framework for saliency-based point feature detection

Mark Brown[a,*], David Windridge[a,b], Jean-Yves Guillemaut[a]

[a] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey GU2 7XH, UK
[b] School of Science and Technology, Middlesex University, London, NW4 4BT, UK

ARTICLE INFO

Article history:
Received 30 November 2015
Revised 26 August 2016
Accepted 13 September 2016
Available online 14 September 2016

Keywords:
Point detection
Feature detection
Feature matching
2D-3D registration
Saliency

ABSTRACT

Here we present a novel, histogram-based salient point feature detector that may naturally be applied to both images and 3D data. Existing point feature detectors are often modality specific, with 2D and 3D feature detectors typically constructed in separate ways. As such, their applicability in a 2D-3D context is very limited, particularly where the 3D data is obtained by a LiDAR scanner. By contrast, our histogram-based approach is highly generalisable and as such, may be meaningfully applied between 2D and 3D data. Using the generalised approach, we propose salient point detectors for images, and both untextured and textured 3D data. The approach naturally allows for the detection of salient 3D points based jointly on both the geometry and texture of the scene, allowing for broader applicability. The repeatability of the feature detectors is evaluated using a range of datasets including image and LiDAR input from indoor and outdoor scenes. Experimental results demonstrate a significant improvement in terms of 2D-2D and 2D-3D repeatability compared to existing multi-modal feature detectors.

1. Introduction

Light Detection And Ranging (LiDAR) scanners have been used to obtain 3D data for decades, but it is only in recent years that they have seen more widespread applicability due to the high computational capacity required to cope with such large datasets. However, the integration of LiDAR scans with data from other modalities (e.g. images) remains a difficult problem, with many approaches relying on line features for their registration (Liu and Stamos, 2012; Wang and Neumann, 2009), which may not always be available. This causes significant bottlenecks in practical applications such as digital film production, where LiDAR scans and images are captured on-set to obtain data about the scene, but subsequently need to be manually registered during post-production. The problem is further exacerbated by the high resolution and large scale of the data, requiring scalable methods for registration that are robust to the diverse, multi-modal aspect of the data.

To address this, here we propose a point feature detector that may be naturally and meaningfully applied between both 2D and 3D data. Feature detection is a typical first stage in many registration pipelines (Li et al., 2010; Liu and Stamos, 2012; Wu et al., 2008b), whereby considering only a small subset of discriminative features in each dataset the registration parameters may be obtained in a relatively straightforward manner. However, obtaining suitably repeatable features between both 2D and 3D data is a particularly challenging problem due to the large heterogeneity between the two modalities.

Instead, existing point feature detection methods are typically centred around images. Recent advances in 3D data acquisition (e.g. Microsoft Kinect) has resulted in a significant interest in 3D feature detection (Guo et al., 2014; Tombari et al., 2013b). However, it is clear that the majority of 2D and 3D feature detectors are constructed in very separate ways. The more popular 2D feature detectors are based on the derivative of the image, and provide a principled approach to scale selection using scale-space theory (Lowe, 2004; Mikolajczyk and Schmid, 2004). Yet, very few may be extended to operate on 3D data, with many 3D feature detectors based on surface curvature (Tombari et al., 2013b), and since the traditional scale-space approach typically cannot be applied to 3D data without altering the geometry. The differences between 2D and 3D feature detectors are further exacerbated by the range of existing 3D data types (point cloud, volumetric, mesh, textured / untextured), leading to different 3D feature detectors for each case (Guo et al., 2014; Tombari et al., 2013b; Yu et al., 2013).

As such, it is very difficult to use existing point feature detectors jointly across 2D and 3D due to the incomparable nature of their constructions, and the limited scope to which 3D detectors may be applied. Applications such as registration, that would

* Corresponding author.
E-mail addresses: m.r.brown@surrey.ac.uk (M. Brown), d.windridge@mdx.ac.uk (D. Windridge), j.guillemaut@surrey.ac.uk (J.-Y. Guillemaut).

http://dx.doi.org/10.1016/j.cviu.2016.09.008

typically rely on point feature detectors, instead use other techniques in the 2D-3D case (e.g. learning a bag of features across multiple viewpoints (Tombari et al., 2013a), or Mutual Information alignment (Mastin et al., 2009)). These approaches are not as general as their feature-based counterparts; often making restrictive assumptions about the scene, or requiring a good initial alignment.

To address this issue, here we propose a more general approach to point feature detection, based on the Kadir-Brady (KB) saliency detector (Kadir and Brady, 2001). Its histogram-based approach does not exclusively depend upon data-type specific quantities such as derivatives or curvatures. Instead, it defines a salient point as having a high information content (as measured by the entropy of its histogram) at a particular scale. This histogram-based approach allows it to be formulated across different modalities in a more meaningful manner than other feature detectors due to the vast array of ways in which histograms may be constructed.

Based upon the KB saliency detector, and inspired by the success of the 2D Harris corner detector (Aanæs et al., 2012; Harris and Stephens, 1988) we propose a novel extension to the 2D KB saliency detector. Whereas the original KB saliency detector constructs a histogram of pixel intensities in a circular region, we propose a derivative-based approach whereby the histogram is constructed based on the distribution of eigenvalues of the second moment matrix. This allows our approach to detect salient points with respect to the derivative of the image, where it may operate in a more general manner than a typical corner detector and avoid repetitive parts of the scene.

By using the generalisable histogram-based approach of the KB saliency detector, the above approach may be naturally extended to 3D data by constructing a histogram based on the 3D second moment matrix (Sipiran and Bustos, 2010). Furthermore, the histogram-based approach allows for the detection of salient points based on both the geometry and texture of the scene by constructing a 2D histogram based on the texture of the 3D surface, and combining the 2D and 3D histograms. This allows it to operate in a meaningful manner regardless of whether or not the 3D data is textured, and is able to combine the best of both sets of features for textured data.

The contributions of this paper are three-fold. Firstly, a generalisation to the KB saliency detector is formulated, demonstrating its broad applicability to operate wherever histograms may be meaningfully constructed within a metric space. Secondly, in light of this generalisation, we propose a 2D derivative-based KB saliency detector based on the second moment matrix. Thirdly, the derivative-based KB saliency detector is naturally extended to 3D, where it may operate on both textured and untextured 3D data. It is, to the best of our knowledge, the first 3D feature detector to operate based on both the geometry and texture of the scene simultaneously. The proposed detectors are evaluated in a 2D-2D and 2D-3D manner where it is shown to be more repeatable than existing detectors (Harris 2D and 3D (Harris and Stephens, 1988; Sipiran and Bustos, 2010), and SIFT 2D and 3D (Lowe, 2004; Zaharescu et al., 2012)).

This paper is structured as follows. In Section 2 we describe related work in point feature detection between 2D and 3D. In Section 3 a description of the KB saliency detector is given, along with proposed extensions and modifications (Kadir et al., 2004; Shao and Brady, 2006; Shao et al., 2007). In Section 4 we propose a generalisation of KB saliency. The generalisation is subsequently implemented for a 2D derivative-based KB saliency detector (Section 5), and a 3D KB saliency detector (Section 6) that may operate on textured or untextured 3D data. In Section 7 results will be given, involving qualitative and quantitative results in both 2D and 3D; finally, conclusions and future work are presented in Section 8.

2. Related work

There has been a significant amount of research in point feature detection; both in 2D (Li et al., 2015; Tuytelaars and Mikolajczyk, 2008) and in 3D (Guo et al., 2014; Tombari et al., 2013b). Here we aim to give a brief overview of point feature detection in each modality, describing and comparing the mechanisms involved.

2.1. 2D point feature detection

A significant number of 2D point feature detectors may be categorised as *derivative-based*. The early Harris corner detector (Harris and Stephens, 1988) is a prime example, based on the second moment matrix M (made up of the partial derivatives of the image in a neighbourhood of the point). When both eigenvalues of M are large, it implies a corner is present; a 'corner measure' is constructed accordingly. Alternatively, the Hessian matrix may be used (Beaudet, 1978) as the basis for a feature detector. It detects 'blob' structures, where a point is of relatively high or low intensity compared to its immediate surroundings. The eigenvectors and eigenvalues describe the size and shape of the blob, with the determinant of the Hessian typically used as a response value.

In the case of both the Harris and Hessian detectors, they may be made affine-invariant by constructing the matrices from image derivatives over an elliptical regions (Mikolajczyk and Schmid, 2004). Furthermore, they may be made scale-invariant by constructing the matrices over ellipses of varying size while convolving with a Gaussian kernel (Mikolajczyk and Schmid, 2004). It is observed that detecting keypoints based on the magnitude of the scale-normalised Laplacian of Gaussians (LoG) produces the highest percentage of correct scales. This has led to the popular SIFT detector (Lowe, 2004) that detects keypoints by the magnitude of the Difference of Gaussians (DoG). DoG is approximately equal to the scale-normalised LoG by the heat equation, hence this approach allows for LoG estimation without the need for derivatives to be computed. However, the DoG response is large for edge-like structures, so SIFT subsequently culls edge responses using the ratio of eigenvalues of the Hessian. The traditional Gaussian scale-space approach has its limitations since it blurs both noise and fine detail (e.g. edges); this has been addressed by Alcantarilla et al. (2012) who use a non-linear scale-space that respects the natural boundaries of the image.

A secondary category of point feature detectors are those that are *intensity-based*. These detectors typically operate over a neighbouring set of pixels, but disregard the derivative of the image. As such, they are often more robust to noise (particularly salt-and-pepper noise) than derivative-based feature detectors. An early intensity-based approach is the SUSAN detector (Smith and Brady, 1997); it defines a Univalue Segment Assimilating Nucleus (USAN) as a set of neighbouring pixels that have a similar intensity value to a centre pixel. Corners are subsequently defined where the number of pixels in the USAN is small. Region detectors typically fall into the intensity-based approaches category; for example, the MSER detector (Matas et al., 2002) detects regions where pixel intensities inside the region are either higher or lower than those on its boundary.

A subset of intensity-based approaches are the *histogram-based* feature detectors that detect feature points via histogram construction. The Kadir-Brady saliency detector (Kadir and Brady, 2001) is an example of this; it constructs a histogram of pixel intensities in a neighbourhood of a point, salient points are detected where the distribution of pixel intensities has a high entropy at a particular scale. It will be discussed in greater detail in the next section, where it forms the basis of the proposed 2D-3D point feature detector.

Using the histogram-based approach, a keypoint may be detected based on the idea of *self-similarity*, (or lack of it) to its neighbours. Maver (Maver, 2010) looks for similar histograms of pixel intensities in radial and tangential regions so as to detect keypoints that exhibit different types of symmetry. Conversely, Lee and Chen (2009) look for a point whose histogram is significantly dissimilar from its immediate neighbours. Tombari and di Stefano (2014) use a similar idea, but where histogram comparison is only performed on the *k*-nearest neighbours and a computationally efficient implementation is proposed. The notion of self-similarity is very useful for multi-modal registration, since scenes may often exhibit a similar structure between modalities but lack similar finer features. Tombari and di Stefano (2014) show their approach to be of potential use for cross-spectral image registration, and Shechtman and Irani (2007) construct a self-similarity descriptor for cross-spectral imagery and sketch-based retrieval.

The majority of 2D point feature detectors are focused purely within the 2D domain. There is evidence to suggest that histogram-based approaches are a promising avenue for multi-modal feature detection due to their general formulation. However, this has never been applied in a 2D-3D context, where the histogram construction process may more generally result in feature detection based on both the geometry and texture of the 3D data.

2.2. 3D point feature detection

Approaches to point feature detection in 3D vary depending upon the type of data being used. For volumetric 3D data many 2D feature detectors may be naturally extended, e.g. 3D SIFT (Flitton et al., 2010). Indeed, a performance evaluation of volumetric 3D feature detectors (Yu et al., 2013) show extensions of familiar 2D feature detectors (Harris, Hessian, MSER, etc). However, other representations of 3D data (point cloud or mesh) create difficulties since points are non-uniformly sampled, points may or may not be textured, and a scale-space may not be so naturally constructed. Point cloud representations are however the subject of this paper and as such feature detection for this representation will be reviewed here.

Similarly to 2D feature detection, the Harris corner detector has been naturally extended to operate on 3D data (Sipiran and Bustos, 2010). For each point, a best fit tangent plane is first determined. Each neighbouring point is projected onto the plane and assigned an 'intensity' value for each point as its distance to the plane. The 2D Harris corner detector may be applied to this set of intensity values, resulting in the 3D Harris corner detector.

Second derivative-based approaches in 3D typically manifest themselves through curvature-based approaches, while avoiding any mention of a Hessian matrix. For example, Chen and Bhanu (2007) propose an approach that locally estimates a quadratic surface around each vertex and uses this to obtain the principal curvatures. They then assign a Shape Index (SI) to each vertex based on the maximum and minimum principal curvatures. Points are detected based upon whether its SI is significantly bigger or smaller than the mean of a neighbourhood of SIs.

Alternative approaches may not be derivative-based at all, taking advantage of the unordered point cloud representation of the data. For example, Zhong (2009) proposes Intrinsic Shape Signatures (ISS), based on the eigenvalue decomposition of the 3 $\times$ 3 covariance matrix around a point. They subsequently cull points whose ratio between successive eigenvalues are similar, then rank feature points in proportion to the smallest eigenvalue. Learning-based approaches have also been proposed, for example by Teran and Mordohai (2014), who learn across a set of geometric attributes using a random forest. The approach allows for specific point detection to match the criteria observed during the training phase, resulting in a more flexible approach.

Scale-space approaches to 3D feature detection have been proposed in a number of ways. Castellani et al. (2008) propose to detect point features by using the Difference of Gaussians (DoG) on the set of 3D points, determining a point's saliency by how far it moves along its normal under the DoG operator. However, this type of approach has been criticised since it obtains a scale-space representation by altering the geometry of the scene. Alternatively, a scale-space may be constructed by convolving other attributes of the 3D data. Such an approach is taken by Zaharescu et al. (2012): they detect keypoints in a generic way that is applicable to scalar functions of 2D manifolds, e.g. mean curvature, or the intensity (if the data is textured). However, it cannot detect keypoints based jointly on geometry and texture. Their approach is similar to SIFT, computing a scalar function at each point, using a DoG operator on the scalar function and rejecting keypoints for which the ratio of the eigenvalues of the Hessian are large.

An approach that is very similar to SIFT is the Viewpoint Invariant Patches approach of Wu et al. (2008a), that is only applicable to textured 3D models. They propose to compute a local tangent plane to each 3D point, onto which a neighbouring texture patch may be orthographically projected. The 2D SIFT detector and descriptor may be subsequently applied on the texture patch to allow a framework for 3D-3D registration. Wu et al. furthermore apply their approach in a 2D-3D scenario (Wu et al., 2008b), where SIFT features are detected in both 2D and 3D data. They determine putative feature matches that are refined by warping the 2D SIFT features such that they approximately match the same form of the orthographic VIP SIFT features.

A histogram-based approach to 3D point feature detection was prosed by Fiolka et al. (2012), who extend the KB saliency detector (Kadir and Brady, 2001) and construct a histogram based on the distribution of normals. However, their approach only detects salient features based on the geometry of the scene and does not detect those based on any available texture; as a result it does not provide a unified approach to salient point detection in 3D. An earlier version of this work was published in Brown et al. (2014) based on the mean curvature, however this was a purely geometry-based KB saliency detector. In this paper we a) propose a derivative-based 2D KB saliency detector, and b) in contrast to both Fiolka et al. (2012) and Brown et al. (2014), we consider both the geometry and texture of the scene, allowing for salient point detection based on both attributes of the data simultaneously. Our framework for generalisable salient point detection is evaluated between 2D and 3D on a range of synthetic and real data.

3. The Kadir-Brady saliency detector

Here an outline of the Kadir-Brady (KB) saliency detector (Kadir and Brady, 2001) and its extensions and various implementations (Kadir et al., 2004; Shao and Brady, 2006; Shao et al., 2007) are given.

The KB saliency detector (Kadir and Brady, 2001) is originally based on the principle that the parts on an image that are highly complex are salient. Scale-invariance is achieved by measuring the complexity across a range of scales and only selecting points whose complexity is peaked with respect to their scale. To further localise its scale, it is required that the point is statistically dissimilar across its neighbouring scales, known as inter-scale saliency. The saliency of a point is therefore the product of two terms: its complexity and its inter-scale saliency. Finally, salient points are clustered into salient regions so as to be more robust to noise. These three stages of the KB saliency detector (complexity estimation, inter-scale saliency, and clustering) are now described in more detail:

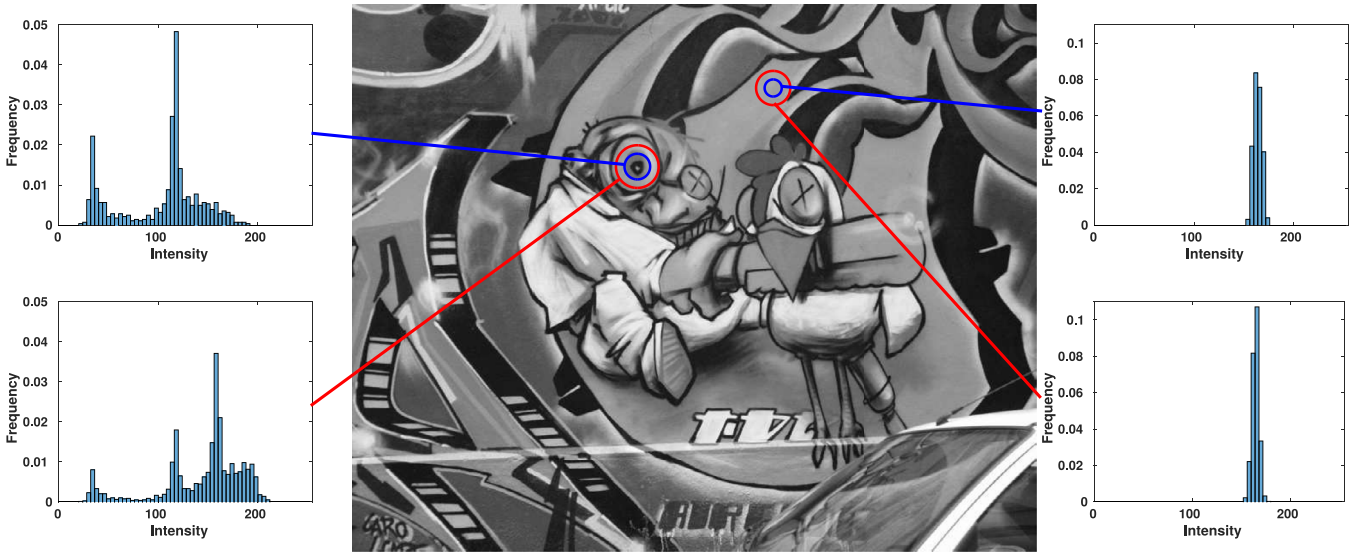

Fig. 1. An example of four distributions of pixel intensities from the image. The distributions on the left have a relatively large entropy and change significantly over scale. The distributions on the right lie in an approximately uniform part of the image, having low entropy and not changing over scale, hence will not be deemed salient by the approach. Image taken from (Mikolajczyk et al., 2005).

Stage I: Complexity estimation. The complexity of a given point ($\mathbf{p}$) at a particular scale (σ_s) is determined by its *entropy*. Entropy is, however, defined for a probability mass function (pmf) P taking one of K values (i.e. $P = \{p_1, \ldots, p_K\}$, $p_i \geq 0\ \forall i$, $\sum_{i=1}^{K} p_i = 1$), and is defined as:

$$H(P) = -\sum_{i=1}^{K} p_i \ln p_i \tag{1}$$

Informally, the entropy of a pmf gives a measure of how 'spread out' it is: it is maximised for the uniform distribution and minimised when the pmf is 1 for one bin and zero for all other bins (Shannon, 1948). We take $0 \ln 0 = 0$ (since $\lim_{x \to 0} x \ln x = 0$).

To meaningfully apply the concept of entropy to a point $\mathbf{p}$ at scale σ_s, a histogram of pixel intensities is first constructed from all pixels within a distance σ_s from $\mathbf{p}$; denoted $\{v_{1,\sigma_s}, \ldots, v_{K,\sigma_s}\}$. The histogram is normalised to obtain a (frequentist) pmf, denoted $\{\hat{v}_{1,\sigma_s}, \ldots, \hat{v}_{K,\sigma_s}\}$, i.e. $\sum_{i=1}^{K} \hat{v}_{i,\sigma_s} = 1$. Then the entropy of point $\mathbf{p}$ at scale σ_s is defined as the entropy of the frequentist pmf:

$$H(\mathbf{p}, \sigma_s) = -\sum_{i=1}^{K} \hat{v}_{i,\sigma_s} \log\left(\hat{v}_{i,\sigma_s}\right) \tag{2}$$

Stage II: Inter-scale saliency. Similarly to other feature detectors, only features whose response value is peaked in scale-space are sought-after; i.e. only features whose entropy is peaked in scale-space are kept. Furthermore, it is necessary for the feature to be statistically dissimilar across scale. Based on this, the pmf is compared to the pmfs of the neighbouring scales, and the saliency is weighted by how dissimilar the pmfs are. Thus the weighting function is constructed as:

$$W(\mathbf{p}, \sigma_s) = \frac{\sigma_s^2}{\sigma_s^2 - \sigma_{s-1}^2} \sum_{i=1}^{K} \left| \hat{v}_{i,\sigma_s} - \hat{v}_{i,\sigma_{s-1}} \right| \tag{3}$$

The coefficient $\frac{\sigma_s^2}{\sigma_s^2 - \sigma_{s-1}^2}$ is used so as to be scale-invariant.

From these two stages, a set of keypoints - those whose entropy is peaked in scale-space - are obtained. They have a saliency value of $H(\mathbf{p}, \sigma_s) \times W(\mathbf{p}, \sigma_s)$. An example of histograms obtained for the first two stages is given in Fig. 1, where the advantages of determining salient points as those with a high entropy and dissimilarity across scale are demonstrated.

Stage III: Salient regions. From the previous stage a great deal of salient points are returned by the detector (typically hundreds of thousands); far too many to be of use in any practical application. Hence, a simple clustering algorithm is proposed. In the original paper (Kadir and Brady, 2001) a rather complicated clustering algorithm, dependent upon two user-defined parameters, is proposed. However, code provided on the author's webpage uses a greedy clustering algorithm: it iteratively takes the point with the highest saliency value and removes all other points within its scale, continuing in this fashion until no points are left. We have found the greedy clustering algorithm to be better in practice, as well as more general since it is parameter free.

A deficiency in the above approach is that it is not affine-invariant: histograms are computed in a circular region around a point, rather than the full range of potential elliptical regions. This was addressed in Kadir et al. (2004) where a full, time-consuming search over all ellipses in the image is implemented. Alternatively, in Shao and Brady (2006), the authors propose to first detect affine-covariant salient regions using the original KB saliency detector, then adapt these to make them affine-invariant.

In (Shao et al., 2007) Shao et al. provide a number of improvements to the algorithm that significantly increase its robustness. They do not change any fundamental aspect of the approach, instead computing desired quantities in a more accurate and principled manner. Specifically,

i) The weighting $W(\mathbf{p}, \sigma_s)$ is more accurately computed, reflecting the ratios of the number of pixels at each scale. Let there be N_s pixels within σ_s from $\mathbf{p}$. Then the weighting is determined as:

$$W(\mathbf{p}, \sigma_s) = \frac{N_s}{N_s - N_{s-1}} \sum_{i=1}^{K} \left| \hat{v}_{i,\sigma_s} - \hat{v}_{i,\sigma_{s-1}} \right| + \frac{N_{s+1}}{N_{s+1} - N_s} \sum_{i=1}^{K} \left| \hat{v}_{i,\sigma_{s+1}} - \hat{v}_{i,\sigma_s} \right| \tag{4}$$

ii) The histogram is sampled differently so as to weight pixels towards the centre of the circle more than those towards the

edge. A Gaussian weighting is initially suggested; instead a computationally inexpensive alternative is proposed where a pixel is weighted twice as much if it is within σ_{s-1} and three times as much if within σ_{s-2}.

iii) Partial volume estimation: some pixels are only partly within the circle. In this case, they contribute to the histogram in proportion to how much of the pixel is inside the circle.

iv) Parzen windowing: the histogram is convolved with a Gaussian to obtain a smoother pdf. Bilinear interpolation is suggested as a computationally inexpensive alternative.

The proposed modifications of Shao et al. (2007) result in some improvement to the performance of the KB detector, as evaluated on the dataset of Mikolajczyk et al. (2005). Hence, Shao et al. demonstrate the potential of the approach as a repeatable feature detector, but do not demonstrate its broad applicability. In the next section, we generalise the KB detector and show how it may be broadly applied across different modalities.

4. The generalised Kadir-Brady saliency detector

The original KB saliency detector was limited in its construction and as such was only applicable to images. In this section we propose a much more general formulation that allows it to be applicable in a multi-modal manner. Subsequently, we propose a derivative-based reformulation in the 2D domain, and a 3D formulation that naturally accounts for both the geometry and texture of the scene.

To generalise the KB saliency detector, we observe that much of its construction is based on a very general concept: points whose entropy is peaked across scale are regarded as salient. To illustrate how widely this concept may be applied, we shall formulate the KB saliency detector in a more general manner for points lying in a metric space.

To this end, let $\mathcal{M}$ be a set and $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ be a metric, i.e. let $(\mathcal{M}, d)$ be a metric space. Define a ball of radius σ centred at $\mathbf{p} \in \mathcal{M}$ as

$$B_\sigma(\mathbf{p}) := \{\mathbf{x} \in \mathcal{M} : d(\mathbf{p}, \mathbf{x}) \leq \sigma\} \tag{5}$$

representing the set of elements of $\mathcal{M}$ within σ of $\mathbf{p}$. Finally, assume a mapping $\mathbf{F}$ may be constructed from each element of $\mathcal{M}$ to an K-dimensional positive vector, i.e. $\mathbf{F} : \mathcal{M} \to \mathbb{R}^{+K}$. Constructing $\mathbf{F}$ as a specifically vector-valued function will allow for broader applicability where multiple attributes of the data are taken into account (e.g. geometry and texture).

From the above constructions the key components of the KB detector may be defined, allowing for generalised KB saliency detection in $(\mathcal{M}, d)$. The probability mass function for an element $\mathbf{p} \in \mathcal{M}$ at scale σ_s is determined by computing a weighted sum over mappings ($\mathbf{F}$) from all points in ball $B_{\sigma_s}(\mathbf{p})$ and normalising: explicitly, the pmf is $\{\hat{v}_{1,\sigma_s}, \ldots, \hat{v}_{K,\sigma_s}\}$, where

$$\hat{v}_{i,\sigma_s} = \frac{\sum_{\mathbf{q} \in B_{\sigma_s}(\mathbf{p})} w(\mathbf{q}, \mathbf{p}) F_i(\mathbf{q})}{\sum_{j=1}^{K} \sum_{\mathbf{q} \in B_{\sigma_s}(\mathbf{p})} w(\mathbf{q}, \mathbf{p}) F_j(\mathbf{q})} \tag{6}$$

where the weighting $w(\mathbf{q}, \mathbf{p})$ is constructed to favour points closer to $\mathbf{p}$. A Gaussian weighting is originally suggested by Shao et al. (2007) but discarded due to considerations of computational efficiency. However, this consideration does not necessarily hold since the weightings may be precomputed, and relative gains in efficiency are always application dependent. In this paper, we use a Gaussian weighting since it leads to a more principled and robust approach:

$$w(\mathbf{q}, \mathbf{p}) = e^{\frac{-\|\mathbf{q}-\mathbf{p}\|}{\sigma_s^2}} \tag{7}$$

With the construction of the pmf (Eq. (6)), the entropy of a point $\mathbf{p} \in \mathcal{M}$ at scale σ_s is well defined, and is the same as Eq. (2):

$$H(\mathbf{p}, \sigma_s) = -\sum_{i=1}^{K} \hat{v}_{i,\sigma_s} \log\left(\hat{v}_{i,\sigma_s}\right) \tag{8}$$

Subsequently the inter-scale saliency, $W(\mathbf{p}, \sigma_s)$, is defined as in Eq. (4). Finally, the saliency of a point $\mathbf{p} \in \mathcal{M}$ at scale σ_s is defined as the product of $H(\mathbf{p}, \sigma_s)$ and $W(\mathbf{p}, \sigma_s)$. Salient points are subsequently clustered by iteratively taking the point with the highest saliency value ($\mathbf{p}_H$) and removing all other points within $B_{\sigma_s}(\mathbf{p}_H)$.

As an example, for the 2D KB saliency detector, the metric space is $(\mathbb{R}^2, L^2)$, representing the image plane under the Euclidean norm. A ball $B_{\sigma_s}(\mathbf{p})$ is simply a circle of radius σ_s centred at $\mathbf{p}$. The mapping $\mathbf{F}$ takes the intensity of a pixel and maps it to the index of the histogram bin (i.e. if the intensity of pixel $\mathbf{p}$ is $I(\mathbf{p})$ then $\mathbf{F}(\mathbf{p}) = (0, \ldots 0, 1, 0, \ldots, 0)$, with a 1 in the $I(\mathbf{p})$th element of the vector). However, the more general construction where $\mathbf{F}$ is a multi-valued function allows for pixels to contribute to multiple bins. This not only extends the KB saliency detector to other modalities but provides additional advantages, e.g. for bilinear interpolation, or where points have multiple attributes (such as where 3D points contain information regarding geometry and texture).

Based on the above formulation, the generalised KB saliency detector may be applied to a range of multi-modal data. In the next two subsections, we construct a derivative-based 2D KB saliency detector, as well as a 3D KB saliency detector that naturally operates on both the geometry and texture of the scene. In both cases, the approaches are elegantly incorporated within the generalised KB saliency framework by simply defining the metric space and constructing the mapping $\mathbf{F}$.

5. Derivative-based 2D Kadir-Brady saliency detector

The original 2D KB saliency detector was constructed based on the distribution of pixel intensities in a neighbourhood of a point. Whilst this gives some indication of some of the more complex, salient parts of the image, it fails to detect the *geometrically salient* aspects. In particular, it rarely detects corners, for which the neighbouring complexity of pixel intensities varies little with scale. As a result, the original 2D KB saliency detector fails to detect repeatable features between 2D and 3D (see the results in Section 7.5); focusing more on the texture of the scene rather than the geometry.

In light of this limitation for the original KB saliency detector and based on the preceding generalisation, in this section we propose a derivative-based KB saliency detector. Specifically, the histogram mapping $\mathbf{F}$ is modified to be a function of the derivative of the image at any given pixel. This allows for high-derivative points within a low-derivative neighbourhood (e.g. corners) to be deemed salient; an important outcome in low-textured scenes. However, it is more general than a typical corner detector, determining salient points wherever a change in image derivative with respect to scale occurs, and avoiding noisy or repetitive parts of the scene.

The derivative-based KB saliency detector is formulated as follows: the metric space is $(\mathbb{R}^2, L^2)$, the same as the original KB saliency detector. The mapping $\mathbf{F}$ is a function of the derivative of the image (specifically, the second moment matrix). Denote the intensity of a pixel $\mathbf{p}$ as $I(\mathbf{p})$ and its derivatives in the x and y directions as $I(\mathbf{p})_x$ and $I(\mathbf{p})_y$ respectively. For a fixed scale σ, construct the second moment matrix (Harris and Stephens, 1988) centred at

Fig. 2. An example of four distributions of second moment matrix eigenvalues from the image. The distributions on the left have a relatively large entropy and change significantly with scale, and are likely to have a high saliency value. Conversely, the distributions on the right, while having a relatively large entropy, do not change significantly over scale, and are likely to have a lower saliency value.

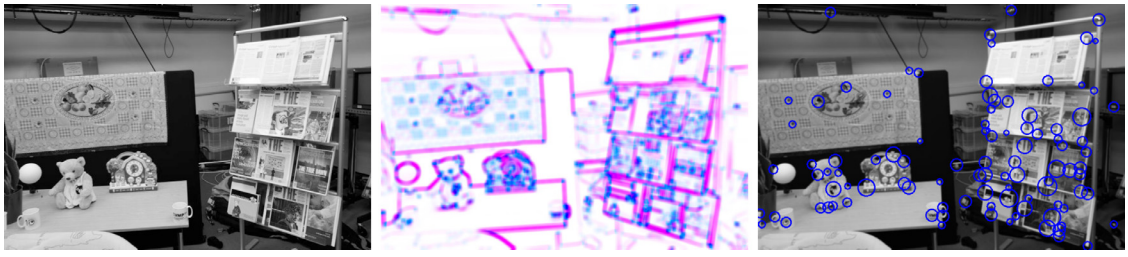

Fig. 3. Example output of the proposed derivative-based KB saliency detector. **Left**: Input image. **Middle**: A heatmap indicating the magnitude of the eigenvalues of $\mathsf{M}(\mathbf{p})$. The intensity of magenta represents the relative magnitude of the first eigenvalue, with blue representing the second eigenvalue. **Right**: Salient points detected based on a histogram of the eigenvalues. The size of the circle represents its scale.

$\mathbf{p}$ as:

$$\mathsf{M}(\mathbf{p}) = \left(\sum_{\mathbf{q} \in B_\sigma(\mathbf{p})} w(\mathbf{q}, \mathbf{p}) \right)^{-1} \times \sum_{\mathbf{q} \in B_\sigma(\mathbf{p})} w(\mathbf{q}, \mathbf{p}) \begin{pmatrix} I(\mathbf{q})_x^2 & I(\mathbf{q})_x I(\mathbf{q})_y \\ I(\mathbf{q})_x I(\mathbf{q})_y & I(\mathbf{q})_y{}^2 \end{pmatrix} \tag{9}$$

where $w(\mathbf{q}, \mathbf{p})$ is a Gaussian weighting function designed to favour points closer to $\mathbf{p}$, e.g. $w(\mathbf{q}, \mathbf{p}) = e^{\frac{-\|\mathbf{q}-\mathbf{p}\|^2}{2\sigma^2}}$. In constructing the matrix, we cap the derivatives at 50 pixels to give a more perceptually meaningful approach that favours all large changes in image derivative to the same extent.

For constructing the derivative-based KB saliency detector, we are interested in the eigenvalues λ_1 and λ_2 of $\mathsf{M}(\mathbf{p})$ that describe the derivative of the image. In qualitative terms, when λ_1 and λ_2 are both large, $\mathbf{p}$ is a corner; when $\lambda_1 \gg \lambda_2$, $\mathbf{p}$ is an edge; and otherwise $\mathbf{p}$ has little change in derivative in any direction. To construct the histogram mapping $\mathbf{F}$, the eigenvalues of $\mathsf{M}(\mathbf{p})$ of all pixels on the image are normalised and discretised to lie in a $r_D \times r_D$ histogram. Subsequently, $\mathbf{F}$ maps the eigenvalues of $\mathsf{M}(\mathbf{p})$ to the bins of the $r_D \times r_D$ histogram (hence, the codomain of $\mathbf{F}$ is $\mathbb{R}^{+r_D^2}$). Bilinear interpolation is performed, meaning at most four elements of $\mathbf{F}$ will be non-zero.

An example of histograms constructed using the proposed derivative-based 2D KB saliency detector is given in Fig. 2, and a heatmap of the relative magnitudes of the eigenvalues of $\mathsf{M}(\mathbf{p})$ alongside the output of the proposed detector is given in Fig. 3. It can be seen that the approach detects salient points where the histogram of eigenvalues changes with respect to scale. This allows it to detect a range of derivative-based structures within the scene while naturally avoiding the repetitive areas.

6. The 3D Kadir-Brady saliency detector

For 3D KB saliency detection, we shall define the metric space and histogram construction from Section 4. Such a general formulation allows for a large range of potential implementations; of note is its applicability to both textureless and textured 3D data within the same framework. More concretely, we may use a histogram mapping $\mathbf{F}$ that describes both the geometry and the texture of 3D data, rendering it equally applicable regardless of whether the 3D data is textured. In this section, we describe the histogram construction based purely on geometry (Section 6.1), on texture (Section 6.2), and on both (Section 6.3). An example of histograms constructed using each approach is shown in Fig. 5.

Regardless of histogram construction, the metric space used here is simply $(\mathbb{R}^3, L^2)$, i.e. consider all points to lie in 3D space under the Euclidean norm. If the 3D data were a mesh the geodesic distance may be used instead, however this is slower to compute and not as widely applicable.

6.1. Geometry-based 3D KB saliency detector

Initially, we describe the approach taken based purely on the geometry of the 3D data. To do so, we project the local surface of the 3D data to an image and apply the same techniques as performed previously (construction of the second moment matrix); a similar approach has been taken for the construction of the 3D Harris corner detector (Sipiran and Bustos, 2010). The image is taken to be a tangent plane to the 3D data, and the 'intensity' value of the image represents the distance of the 3D data to the plane. We take a purely derivative-based approach in this subsection; an intensity-based geometric KB detector may not be constructed since the 'intensity' value of every point onto its own tangent plane is always zero.

Our derivative-based geometric KB detector is more formally constructed as follows: for a point $\mathbf{p} \in \mathbb{R}^3$, first determine a least-square tangent plane at $\mathbf{p}$. Construct an orthonormal frame for the tangent plane as $\{\mathbf{t}_1, \mathbf{t}_2, \mathbf{n}\}$, where $\mathbf{n}$ is the normal to the plane. Then, for a fixed scale σ, consider the neighbouring set of points $\{\mathbf{q} \in B_\sigma(\mathbf{p})\}$. Project each point onto the plane, yielding local (u, v) coordinates $((\mathbf{q}-\mathbf{p})\cdot\mathbf{t}_1, (\mathbf{q}-\mathbf{p})\cdot\mathbf{t}_2)$ and define its 'intensity' value $I(\mathbf{q})$ as the directional distance from $\mathbf{q}$ to the plane, computed as $(\mathbf{q}-\mathbf{p})\cdot\mathbf{n}$. The second moment matrix may thus be constructed in the same way as Section 5 as:

$$\mathsf{N}(\mathbf{p}) = \left(\sum_{\mathbf{q}\in B_\sigma(\mathbf{p})} w(\mathbf{q},\mathbf{p})\right)^{-1} \times \sum_{\mathbf{q}\in B_\sigma(\mathbf{p})} w(\mathbf{q},\mathbf{p}) \begin{pmatrix} I(\mathbf{q})_u^2 & I(\mathbf{q})_u I(\mathbf{q})_v \\ I(\mathbf{q})_u I(\mathbf{q})_v & I(\mathbf{q})_v{}^2 \end{pmatrix} \tag{10}$$

where, similarly to Section 5, $w(\mathbf{q},\mathbf{p}) = e^{\frac{-||\mathbf{q}-\mathbf{p}||^2}{2\sigma^2}}$. The eigenvalues of $\mathsf{N}(\mathbf{p})$ are subsequently used in the histogram mapping $\mathbf{F}$, in the same manner as performed previously. Note that the orthonormal frame $\{\mathbf{t}_1, \mathbf{t}_2, \mathbf{n}\}$ is not unique - there is ambiguity in the directions of $\mathbf{t}_1$ and $\mathbf{t}_2$. However, the eigenvalues of $\mathsf{N}(\mathbf{p})$ are rotationally invariant and therefore this ambiguity will not affect the desired outcome. Hence, we have avoided the need to construct a unique and unambiguous orthonormal reference frame that often plagues 3D feature detectors (Guo et al., 2013; Petrelli and di Stefano, 2012).

However, the derivatives $I(\mathbf{q})_u$ and $I(\mathbf{q})_v$ required in Eq. (10) may not be estimated as easily as for the 2D detector, where the intensity values of a pixel's immediate neighbours may be used to determine the derivative. Instead, we compute a Gaussian weighted average from a set of neighbouring points, similarly to Zaharescu et al. (2012). To compute the derivatives $I(\mathbf{q})_u$ and $I(\mathbf{q})_v$ from a non-uniformly sampled set of 2D points $\{\mathbf{r} \in B_\sigma(\mathbf{q})\}$ each with intensity $I(\mathbf{r})$; firstly, denote the derivative for the 2D point $\mathbf{q}$ as $\mathbf{g} := (I(\mathbf{q})_u, I(\mathbf{q})_v)$. Then note that, for a point $\mathbf{r}$ lying sufficiently close to $\mathbf{q}$, the following relationship holds by definition of the derivative:

$$\mathbf{g}^T(\mathbf{q}-\mathbf{r}) \approx I(\mathbf{q}) - I(\mathbf{r}) \tag{11}$$

We may use Eq. (11) to determine $\mathbf{g}$ by solving the weighted least-squares equation:

$$\arg\min_{\mathbf{g}} \sum_{\mathbf{r}\in B_\sigma(\mathbf{q})} w(\mathbf{r},\mathbf{q})\left(I(\mathbf{q}) - I(\mathbf{r}) - \mathbf{g}^T(\mathbf{r}-\mathbf{q})\right)^2 \tag{12}$$

where $w(\mathbf{r}, \mathbf{q})$ is a Gaussian of small variance, e.g. $w(\mathbf{r},\mathbf{q}) = e^{\frac{-||\mathbf{r}-\mathbf{q}||^2}{2\left(\frac{\sigma}{2}\right)^2}}$ so that the local derivative estimates of $I(\mathbf{q})$ are computed over a tighter region than that from which $\mathsf{N}(\mathbf{p})$ is constructed. Eq. (12) is solved by 'stacking' each weighted equality in (11) to form an over-determined system of the form $\mathsf{A}\mathbf{g} = \mathbf{b}$, from which the least-squares solution to (12) is given by $\mathbf{g} = (\mathsf{A}^\mathsf{T}\mathsf{A})^{-1}\mathsf{A}^\mathsf{T}\mathbf{b}$.

Subsequently, computing the gradient $(I(\mathbf{q})_u, I(\mathbf{q})_v)$ for every neighbouring point projected onto the tangent plane allows for the matrix $\mathsf{N}(\mathbf{p})$ to be constructed and its eigenvalues to be computed. To construct the mapping $\mathbf{F}$, the eigenvalues of $\mathsf{N}(\mathbf{p})$ of all points in the data are normalised and discretised to lie in a $r_G \times r_G$ histogram, where bilinear interpolation is performed. An example of the proposed geometry-based KB saliency detector is shown in Fig. 4 alongside a heatmap of the eigenvalues of $\mathsf{N}(\mathbf{p})$. The approach detects a range of geometrically significant structures in a scale-invariant manner, while avoiding the more repetitive areas of the model.

6.2. Texture-based 3D KB saliency detector

We propose two texture-based 3D KB detectors: an intensity-based approach and a derivative-based approach, both of which will be evaluated in Section 7.5. For the intensity-based approach, the mapping $\mathbf{F}$ is exactly the same as in the original 2D KB implementation: taking the intensity of a point to its histogram bin while applying bilinear interpolation. Where the 3D data is coloured, the greyscale value is computed via the equation $I = 0.299R + 0.587G + 0.114B$. The histogram is assumed to be of the same size (K) as the original intensity-based 2D KB implementation.

To obtain the mapping $\mathbf{F}$ for the derivative texture-based 3D KB saliency detector, we adopt essentially the same approach as the geometry-based 3D KB saliency detector in the previous section. The local surface of the 3D data is projected onto a tangent plane, and the second-moment matrix (Eq. (10)) may be constructed again. However, rather than using the intensity value of a projected point $I(\mathbf{q})$ as the directed distance between $\mathbf{q}$ and the tangent plane, the greyscale value of the point $\mathbf{q}$ is used instead. The intensity differences $(I(\mathbf{r}) - I(\mathbf{q}))$ in Eq. (12) are capped between -50 and 50 pixels, similarly to the 2D approach in Section 5, so as to give a more perceptually meaningful distance. The eigenvalues of $\mathsf{N}(\mathbf{p})$ (where $I(\mathbf{q})$ represents the intensity of point $\mathbf{q}$) are subsequently normalised to lie in a ${r_D}^2$ histogram.

6.3. Geometry and texture based 3D KB saliency detector

Our framework naturally allows for the extension to detect salient points based on both the geometry and texture. Given that the two histograms may be constructed based on the geometry or the texture, their joint histogram may be constructed. The intensity texture-based KB detector may be combined with the geometry-based KB detector to produce a $K{r_G}^2$ histogram. Alternatively, the derivative texture-based KB detector may be combined with the geometry-based KB detector, to produce a ${r_D}^2{r_G}^2$ histogram. Bilinear interpolation is again performed in these histograms.

An example of histograms constructed based on the geometry, derivative-based texture, and both, is shown in Fig. 5. The histograms based on both are the joint histogram of the geometry and the derivative-based texture histograms. They are relatively large and, in general, sparse; exhibiting a very high entropy only when caused by both the geometry and texture. However, this approach is able to detect salient points based on either the geometry and texture, since in either case a relatively high entropy is observed at a particular scale.

7. Experimental evaluation

In this section we evaluate the performance of our proposed generalised salient point detector against other approaches, with both 2D and 3D data. Qualitative and quantitative results are given, where the final aim is to detect highly repeatable, sparse features

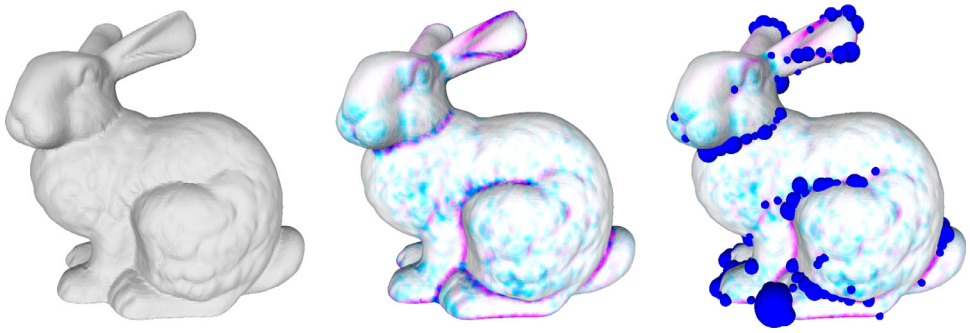

Fig. 4. Example output of the proposed derivative-based KB saliency detector. **Left**: Input 3D data. **Middle**: A heatmap indicating the magnitude of the eigenvalues of $\mathrm{N}(\mathbf{p})$. The intensity of magenta represents the relative magnitude of the first eigenvalue, with blue representing the second eigenvalue. **Right**: Salient points detected based on a histogram of the eigenvalues. The size of the sphere represents its scale.

Fig. 5. An example of the derivative-based histogram distributions from 3D data when considering geometry, texture, and both. The point on the right has a large distribution of eigenvalues based on texture but not based on geometry, whereas the point on the left has a relatively larger distribution of eigenvalues based on geometry (as well as texture). In both cases, the resulting joint histogram (based on geometry and texture) is relatively sparse.

between 2D and 3D, that may be of use in the subsequent registration stage. For comparison against our approaches, there exist a large number of feature detectors in both 2D and 3D (Guo et al., 2014; Tuytelaars and Mikolajczyk, 2008), however we focus specifically on comparing against feature detectors that may be meaningfully constructed in both 2D and 3D. We shall first introduce the detectors in each modality before describing how they are evaluated: firstly between 2D and 2D, and secondly between 2D and 3D.

In 2D, we consider five detectors. Firstly, the traditional *Harris* corner detector (Mikolajczyk and Schmid, 2004). However, it is observed that, for small numbers of features, *Harris* does not detect a suitable spread of features, with many corners detected in the same area (see Fig. 9). Therefore, we secondly evaluate the Good Features to Track algorithm (*GFT*) Shi and Tomasi (1994) to obtain a better, more representative set of corners. Thirdly, we evaluate against the state-of-the-art *SIFT* detector (Lowe, 2004). The final two detectors evaluated are the proposed derivative-based KB detector (Section 5), referred to as *KBD*, and the original intensity-based KB detector (Shao et al., 2007) (referred to as *KBI*) so as to experimentally justify the construction of the proposed *KBD* detector formulated in Section 5.

In 3D, there are optional detectors available to compare against depending upon if the texture of the data is used. For untextured 3D data, we consider four detectors: *Harris* (Sipiran and Bustos, 2010), *SIFT*, *SURE*[1](Fiolka et al., 2012) and the proposed derivative-based geometric KB detector (Section 6.1), referred to as *KB-G*. In 3D, *Harris* is not scale-invariant and performs non-maxima suppression, therefore typically detects a better spread of corners in 3D than its 2D counterpart; hence there is no need for a 3D Good Features to Track detector. For untextured 3D data, *SIFT* detects keypoints based upon the mean curvature, and will be referred to as *SIFT-G*. Both *Harris* and *SIFT-G* are implemented in Point Cloud Library.[2] Harris is extended to 3D (Filipe and Alexandre, 2014) by replacing image gradients by surface normals from which a 3D covariance matrix is constructed. The response value is then a function of the determinant and trace of the covariance matrix (similar to 2D). *SIFT* is extended to 3D (Hänsch et al., 2014) using either the curvature of a point or the intensity (if the 3D

[1] Code available from https://github.com/torstenfiolka/sure3d
[2] http://pointclouds.org/

point cloud is textured). A Difference-of-Gaussians (DoG) may be applied solely on this attribute of the point cloud (curvature or intensity) that does not change the position of the points. Local maxima and minima may then be found by comparing to a point's k-nearest neighbours, subsequently points with low curvature are rejected as they are deemed unstable.

For textured 3D data, there are additional detectors that may be evaluated against. *SIFT* may detect features on textured data based on the intensity (referred to as *SIFT-T*). Alternatively, the KB approaches may be used to detect features based purely on the texture, with the intensity-based KB detector referred as *KBI-T* and the derivative-based KB detector for textured 3D data referred to as *KBD-T*. Only the KB approaches allow for both the texture and geometry to be combined (Section 6.3), referred to as *KBI-B* and *KBD-B*.

From the above 2D feature detectors (*Harris, GFT, SIFT, KBI*, and *KBD*) we firstly evaluate their repeatability in a 2D-2D scenario (Section 7.4). Subsequently, alongside the 3D feature detectors (*Harris, SIFT-G, KB-G, SURE, SIFT-T, KBI-T, KBD-T, KBI-B*, and *KBD-B*) we evaluate their repeatability between 2D and 3D. For untextured 3D data, we use six 2D-3D point combinations: *Harris-Harris, GFT-Harris, SIFT-SIFT-G, KBI-KB-G, KBD-SURE* and *KBD-KB-G*. For textured data there are a further five 2D-3D combinations: *SIFT-SIFT-T, KBI-KBI-T, KBD-KBD-T*; and where both geometry and texture are considered by KB: *KBI-KBI-B* and *KBD-KBD-B*. Thus, where the 3D data is textured, a total of 11 2D-3D feature detector combinations will be evaluated, to compare the effects of considering the geometry, texture, or both, of the textured 3D data.

7.1. Implementation details

For the proposed *KB* detectors two parameters are user-defined: the number of bins for the mapping $\mathbf{F}$ (K, r_D and r_G), and the number and range of scales (σ_s). For the number of bins of *KBI* we take $K = 16$ in both 2D and 3D. For the proposed derivative-based approaches (*KBD*) we use $r_D = r_G = 4$; hence, both *KBI-B* and *KBD-B* have the same total number of bins of 256. The number of scales is 12 in all cases. For the range of scales in 2D we take $\sigma_1 = 3$ with $\sigma_s = 3 + \sigma_{s-1}$. This is similar to the parameters of Shao et al. (2007) whose experiments show that a gap of 3 pixels between scales performed the best. In 3D, the scale is defined in proportion to the size of the model. First, denote the length of the diagonal of the bounding box of the model as L. Then, for the synthetic data, $\sigma_1 = 0.004L$ whereas $\sigma_1 = 0.003L$ for real data (since features are relatively smaller for the more complex real data). Subsequent scales are defined by $\sigma_s = s\sigma_1$, the same as the mesh saliency approach by Lee et al. (2005). In determining the parameter σ_1 in both the 2D and 3D case, we run experiments to justify our choice of parameters (shown in the appendix). For the construction of matrices $\mathsf{M}(\mathbf{p})$ and $\mathsf{N}(\mathbf{p})$ in Eqs. (9) and (10), the size of the ball $B_\sigma(\mathbf{p})$ is taken to be $\sigma = 5$.

For a fair comparison, the other approaches (*SIFT, GFT, Harris*, and *SURE*) are altered, where possible, to align with these user-defined parameters. For *SIFT* in 2D the parameters provided by Vedaldi and Fulkerson (2008) are used and by Mikolajczyk et al. (2005) for *Harris*; and the parameter for *GFT* is defined such that no two corners are within 16 pixels of each other. In 3D, the fixed scale of *Harris* is set to σ_1, and for *SIFT-G, SIFT-T*, and *SURE*, 12 scales are used, with the smallest set to σ_1.

7.2. Datasets

Three datasets are used: a 2D-2D dataset from Mikolajczyk et al. (2005) (shown in Fig. 6); a synthetic 2D-3D dataset (shown in Fig. 7); and a real 2D-3D dataset (shown in Fig. 8).

The 2D dataset is taken from Mikolajczyk et al. (2005). It is a set of six groups of six images, with the known homography between each image in a group provided. Each group of images has undergone a certain transformation (blurring, scale, JPEG compression, lighting, and viewpoint (twice)), from small to large transformations. The first and last images in each group are shown in Fig. 6.

For synthetic data, we use six untextured 3D models. The first four models in Fig. 7 are from the Stanford 3D Scanning Repository.[3] For each of these four models, 50 images were rendered using POV-Ray using a random rotation matrix (Arvo, 1992) and translation such that the model is centred in the image, using a point light source at the same location as the camera. The latter two models are the 3D reconstruction provided by Guillemaut and Hilton (2011) of the dinosaur and temple from Middlebury's multi-view reconstruction dataset (Seitz et al., 2006). In this case, 50 images with their known projection matrix from the model are provided as part of the dataset, so there is no need for rendering using POV-Ray.

For real data (Fig. 8), we use five textured 3D models, obtained by a colour LiDAR scanner. All have been obtained from Kim (2014) with the exception of *room*, which is from Klaudiny et al. (2014). The number of points and the dimensions of the 3D models is tabulated below (Table 1):

For each model, a set of between 7 and 11 images have been taken of the scene and manually aligned. This has been achieved by picking pairs of image and scene points, and using the approach by Penate-Sanchez et al. (2013) to determine the pose and focal length of the camera. An example image of each model is shown at the bottom of Fig. 8. Note that for certain models this does not encapsulate much of the scene (e.g. *courtyard*), making 2D-3D point detection more difficult.

7.3. Evaluation measure

The performance of a point detector (either in 2D-2D, or in 2D-3D) is measured by its *relative repeatability*. To define this, we shall first define the repeatability between two sets of points (2D-2D or 2D-3D) as follows: first apply the known transformation (homography, or projection matrix) to one set of points, discarding any that do not lie within the image boundary of the other set of points. For 2D-3D evaluation, occlusions may be handled in the case of the synthetic 2D-3D dataset, the 3D mesh is known and hence occluded points may be discarded; however often real data is in the form of a point cloud and this is not possible. From one set of 2D points $\{\mathbf{p}_i \in \mathbb{R}^2\}_{i=1}^N$ and the other set of transformed points $\{\mathbf{q}_i \in \mathbb{R}^2\}_{i=1}^M$ (transformed under a homography, or a projection matrix), and given an inlier threshold t, define an *inlier* as a point pair $(\mathbf{p}, \mathbf{q})$ for which i) the nearest neighbour to $\mathbf{p}$ from the set $\{\mathbf{q}_i\}_{i=1}^M$ is $\mathbf{q}$ and vice-versa; and ii) $||\mathbf{p} - \mathbf{q}|| < t$. The repeatability is subsequently defined as the number of inliers divided by $min(N, M)$.

It has been observed in the literature (e.g. Hauagge and Snavely, 2012; Tombari et al., 2013b) that the repeatability measure is biased towards detectors that produce a lot of features, and a measure that is invariant to the number of points detected is proposed. Therefore, we compute the *relative repeatability*: for each set of points, order them in decreasing value of their response value. Then, the repeatability may be determined from the top-k points, and a graph may be plotted of repeatability against the k most responsive features in each set. Furthermore, this is a more useful measure for the purposes of sparse 2D-3D registration, where large numbers of features will not be of use due to the computational complexity of such a registration problem.

[3] http://graphics.stanford.edu/data/3Dscanrep/

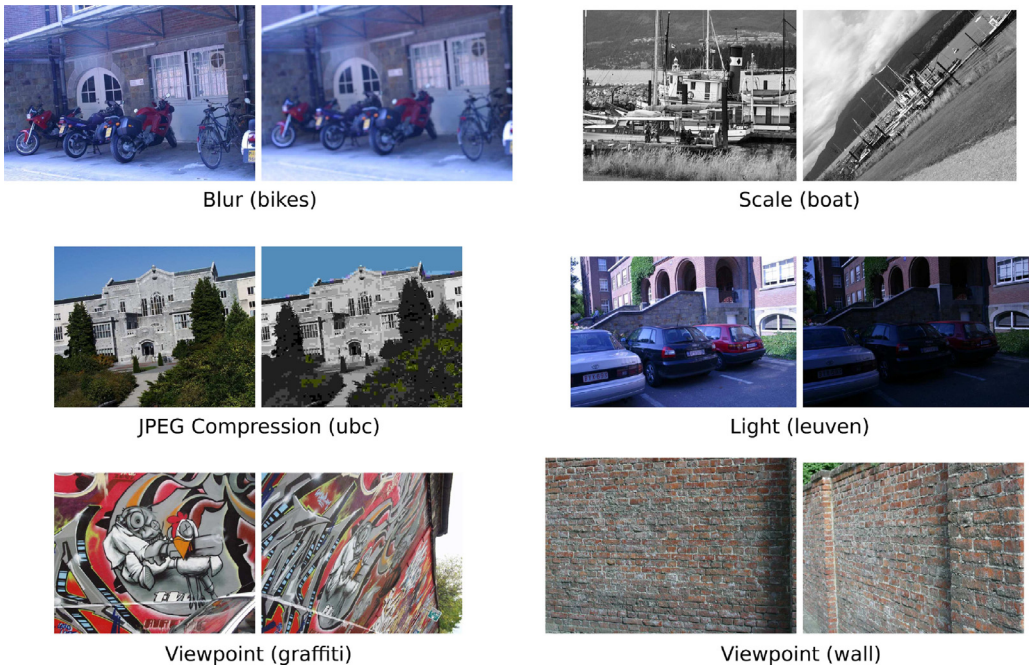

Fig. 6. Examples in the 2D-2D dataset from six groups of image transformations. For each group, there are six images in the dataset ranging from small to large transformations, with the first and last images in each group shown here.

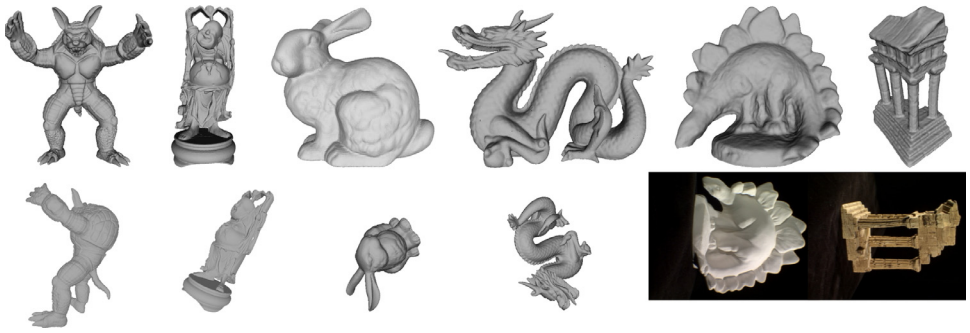

Fig. 7. Top: The 3D models used in the synthetic 2D-3D dataset. **Bottom**: An example image from each synthetic model used in the dataset. From left to right: *armadillo, buddha, bunny, dragon, dino, temple*.

Fig. 8. Top: The 3D models used in the real 2D-3D dataset. **Bottom**: An example image from each model used in the real dataset. From left to right: *cathedral, courtyard, reception, room, studio*.

Table 1
3D models information.

	cathedral	*courtyard*	*reception*	*room*	*studio*
Number of vertices	522,018	672,342	772,536	524,873	348,592
Bounding box diameter (m)	67.2	27.9	17.6	5.34	7.80

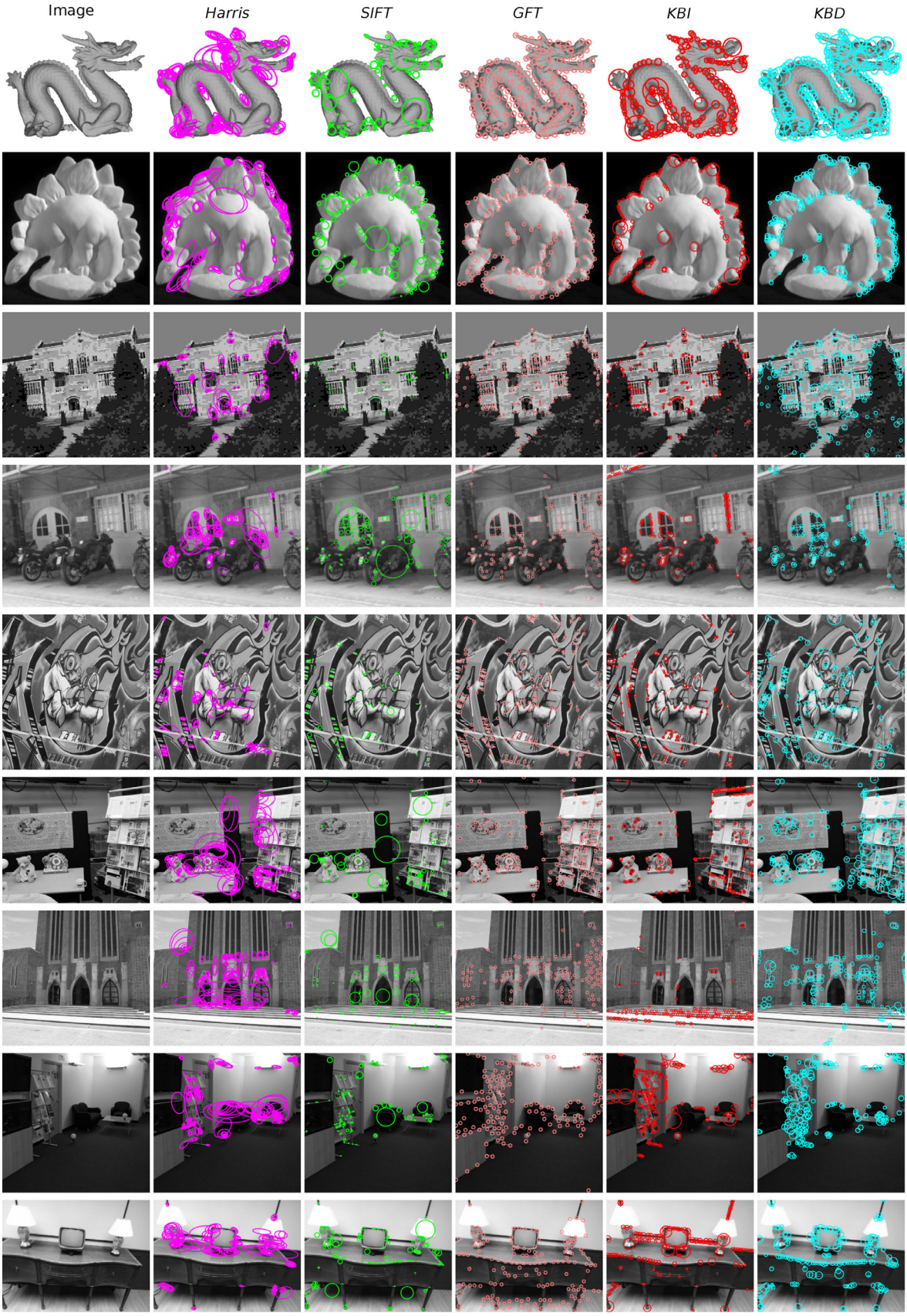

Fig. 9. Qualitative 2D results. The top-150 features are shown in each case. The top two images are from the synthetic 2D-3D dataset, third to fifth from the 2D-2D dataset (Mikolajczyk et al., 2005), with the bottom four from the real 2D-3D dataset. Many images were cropped from their original dataset for ease of presentation in this figure.

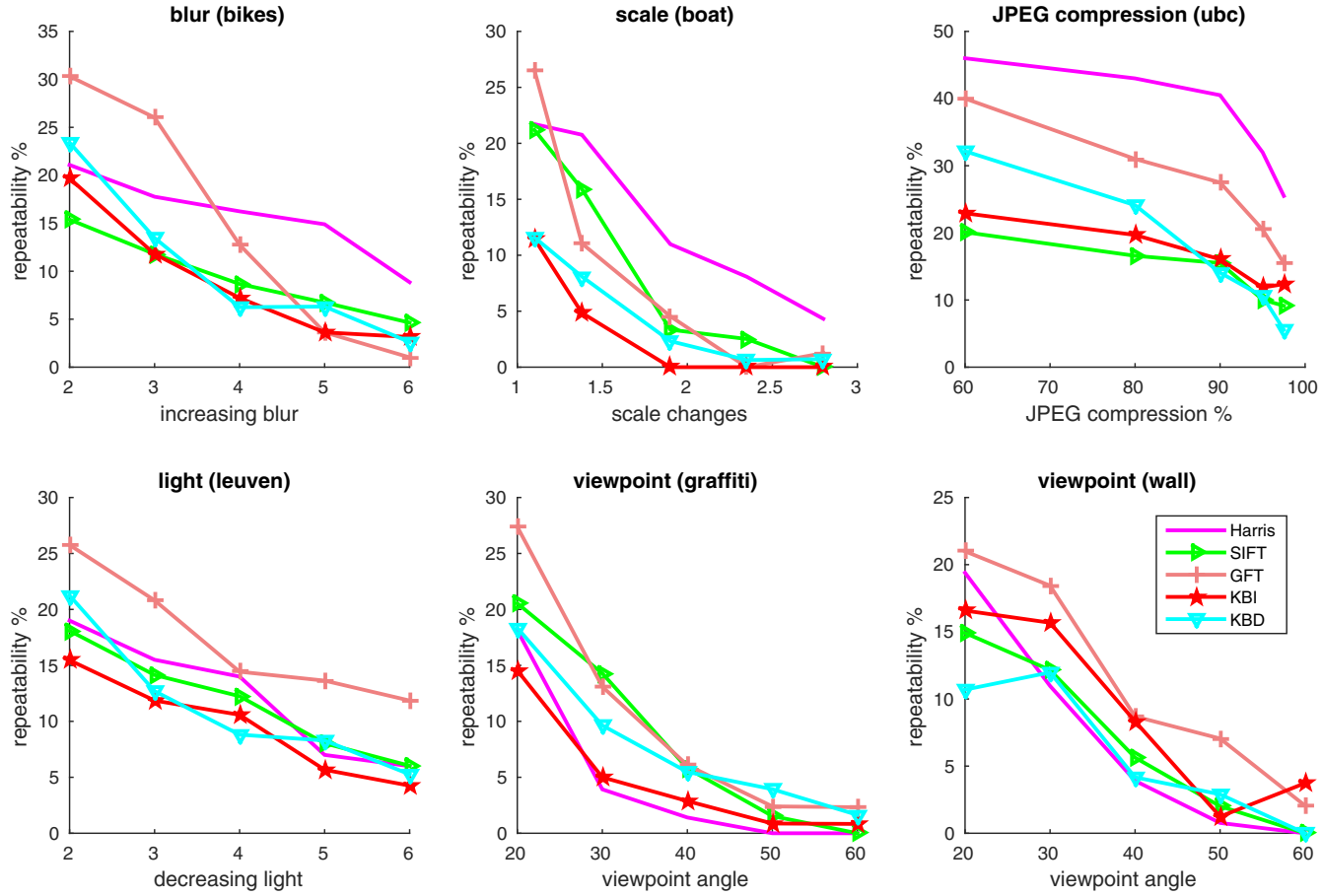

Fig. 10. Quantitative 2D-2D results across a range of image transformations. The relative repeatability is measured for the top-100 point features in each case. An inlier threshold of 3 pixels is used. Example images from this dataset are shown in Fig. 9.

7.4. 2D point detection

Qualitative results for the set of five 2D point detectors are shown in Fig. 9, for a selection of images across the three datasets used. It is immediately noticeable, by the size and shape of the features, that *Harris* is affine- and scale-invariant; *SIFT, KBI* and *KBD* are scale-invariant, and *GFT* is neither, being a very parameter-dependent approach. *SIFT*, and in particular *Harris*, evidently have a tendency to detect the same feature at multiple scales and very similar locations: this motivated the use of *GFT* to obtain a better spread of features (Section 7). *KBI* and *KBD* naturally detect a better spread of points than *Harris* and *SIFT*, while retaining a parameter-free approach to scale selection.

As a qualitative comparison between the KB approaches; *KBD* detects more corners than *KBI* (e.g. on the cathedral) while still detecting blob-like structures (e.g. windows in the third from top image) due to the necessary change in derivative present in such features. In contrast, *KBI* does not detect as wide a range of point feature types as *KBD* and often detects many edges (e.g. the cathedral). While edges may be regarded as salient, a point on an edge is poorly localised along the edge and is not useful for registration purposes.

Quantitative results for 2D point feature detectors are given in Fig. 10 for the 2D-2D dataset (Fig. 6). The top-100 features are detected in each image, and an inlier threshold of 3 pixels is used. It is observed that no feature detector performs the best across all transformations. *Harris* performs particularly well for scale and JPEG compression changes, but very poorly across a change in viewpoint. *GFT* generally performs very well across the range of transformations. Importantly, *KBD* outperforms *KBI* across a number of transformations, justifying our proposed reformulation of the 2D KB detector.

7.5. 3D point detection

7.5.1. Qualitative results

Qualitative results for the 3D feature detectors are shown in Fig. 11 for synthetic data and Fig. 12 for real data.

For the untextured synthetic data, *Harris, SIFT-G, KB-G*, and *SURE* may be used. In Fig. 11, the scale-covariant *Harris* detector successfully detects a number of small-scale corners but often in repetitive places (e.g. the leg of the *armadillo*). *KB-G* is more robust than *SIFT-G*, detecting a wider range of points, e.g. on the *armadillo* and *dino*. By contrast, *SIFT-G* has a tendency to detect smaller, less meaningful features, e.g on the *bunny*. *SURE* typically detects corner-like structures where there is a wide distribution of normals, however it often detects large features and misses smaller corners e.g. on the *dragon*. As a comparison between features detected in 3D and the qualitative 2D results (Fig. 9); 3D *Harris* correlates quite well with 2D *GFT*, however it is clear the scale-covariance of *GFT* is an issue on the *dragon*. *SIFT* and *SIFT-G* often do not detect geometrically meaningful entities, with some 2D *SIFT* features detected off the model. *KBI* and *KBD* have some qualitative correlation with *KB-G*, but *KBI* often detects edges and avoids corner-like structures (particularly so on the *dino*).

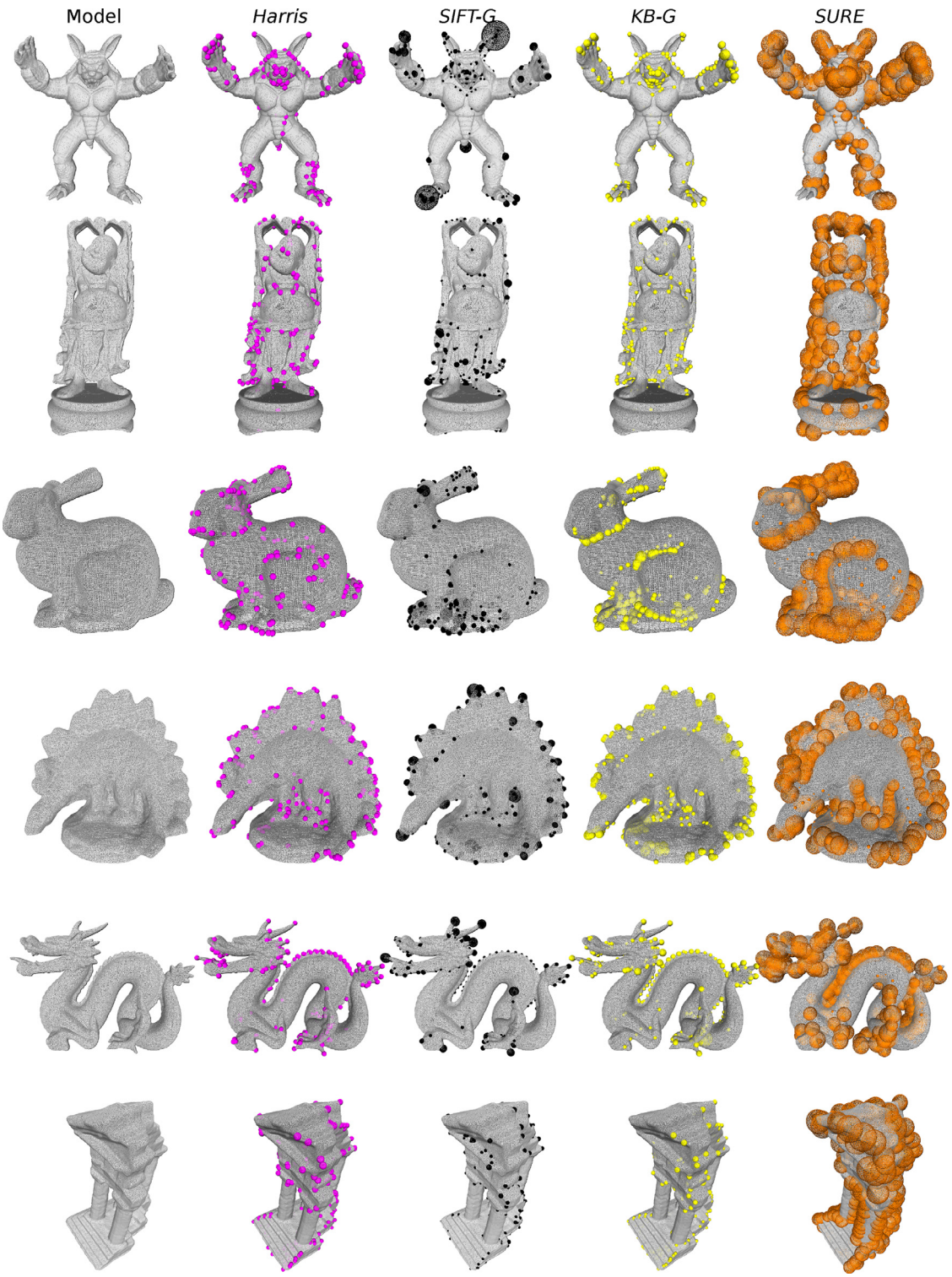

Fig. 11. Qualitative 3D results for all models from the synthetic dataset. The top-200 points are shown in each case. The size of the sphere indicates the scale of the point.

Qualitative results for real data are given in Fig. 12, where points are detected based on geometry (*Harris, SIFT-G, KB-G*), texture (*SIFT-T, KBI-T* and *KBD-T*), or both (*KBI-B* and *KBD-B*). Similar conclusions may be drawn from the geometry-based approaches as for the synthetic results (Fig. 11): *Harris* is limited by its scale-covariance, *KB-G* is generally more robust than *SIFT-G*, and *SURE* typically detects larger features and misses the finer detail. For texture-based detectors, few qualitative distinctions can be made between *SIFT-T* and *KBD-T*, however *KBD-T* detects more textural corner-like structures than *SIFT-T* (the same as in 2D in Fig. 9). Similarly to the 2D results, *KBI-T* detects more edge-like structures - particularly on the pavement on the *cathedral*. Interestingly, texture-based feature detectors often detect geometrically-significant features (e.g. corners on the *cathedral*, and the table-leg in the *room*) due to a natural change in colour on the model surface, or the lighting conditions. Finally, it is clear that both *KBI-B*

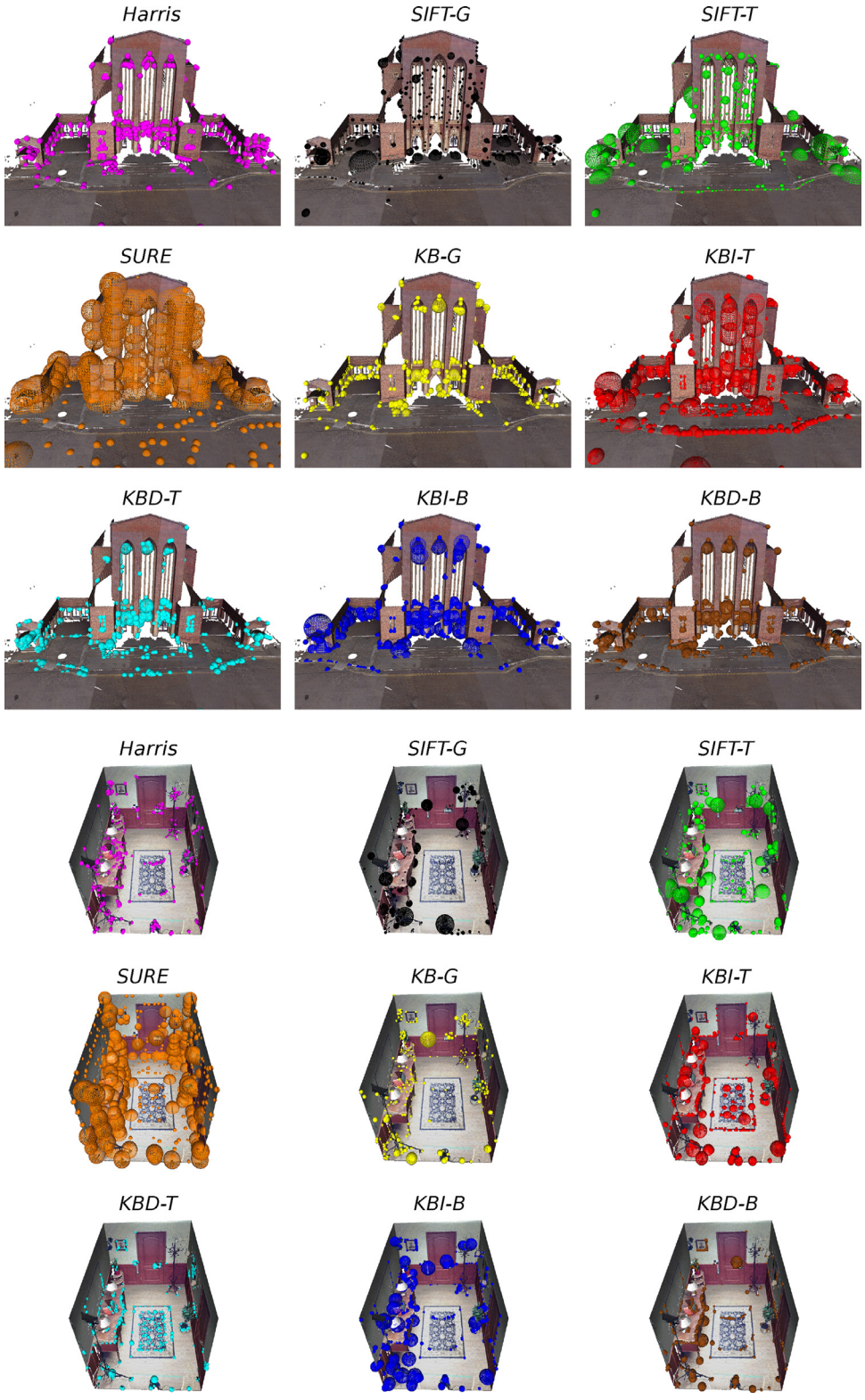

Fig. 12. Qualitative 3D results for *cathedral* and *room* from the synthetic dataset. The top-400 points are shown in each case. The size of the sphere indicates the scale of the point.

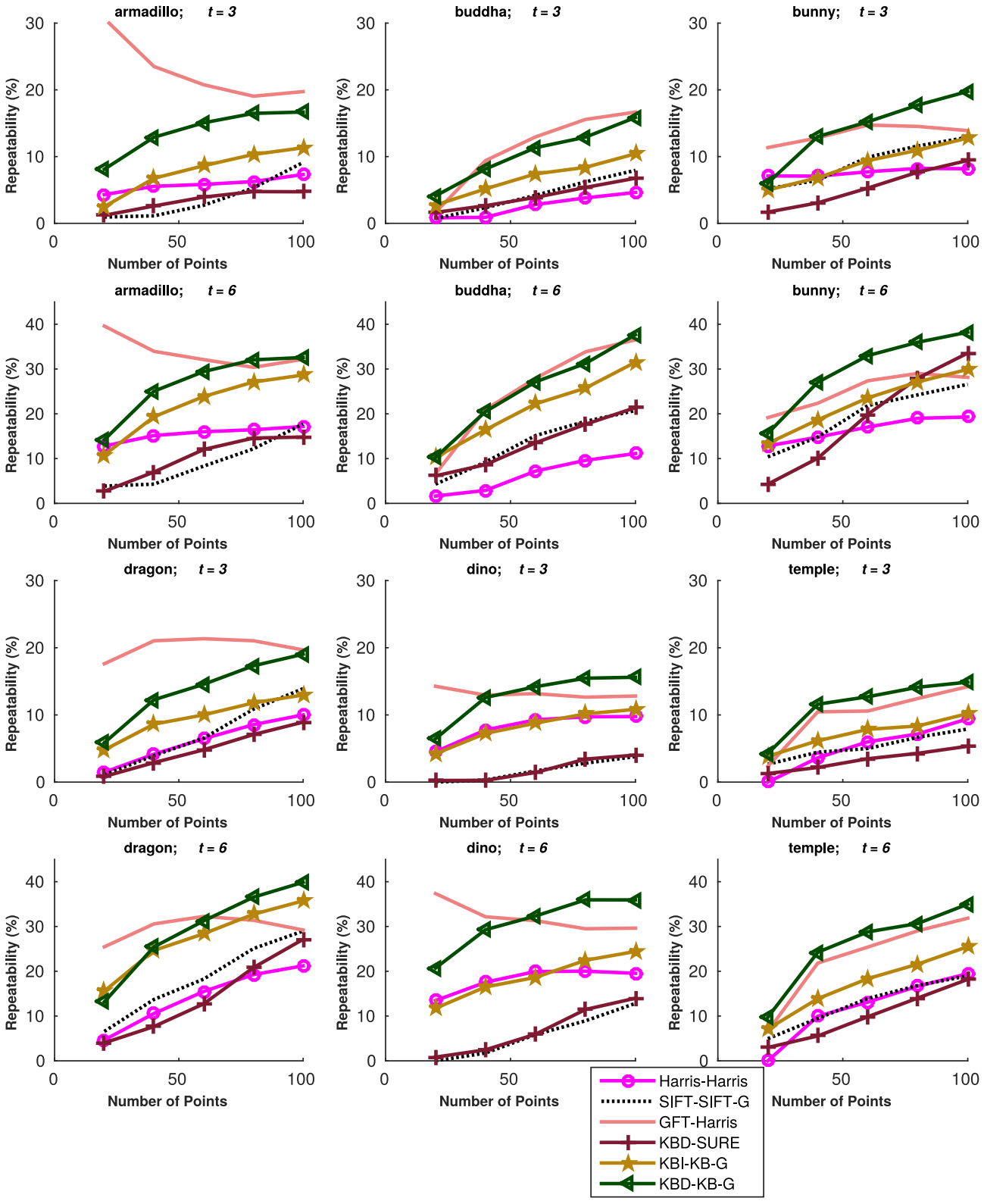

Fig. 13. Results on the untextured synthetic dataset. Each graph shows the relative repeatability of the detectors for each dataset, for $k = 20, 40, 60, 80, 100$. The graphs are ordered such that a graph of inlier threshold 3 pixels is shown above that of inlier threshold 6 pixels.

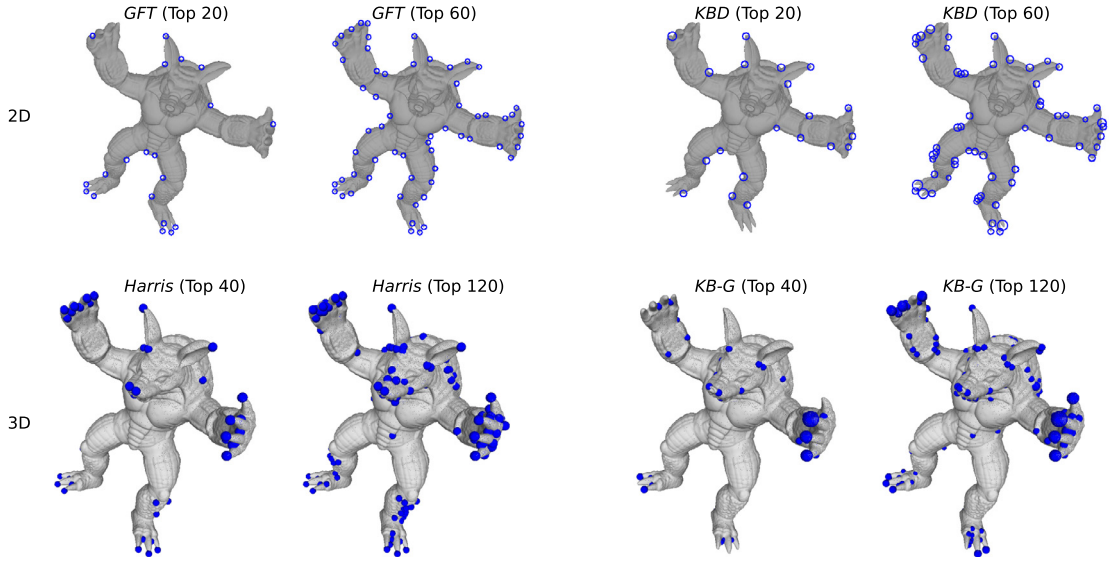

Fig. 14. Qualitative 3D results for varying quantities of features on the *armadillo* model. The left shows results for *GFT* and *Harris*, with *KBD* and *KB-G* on the right.

and *KBD-B* detect points based on both the geometry (corners of the *cathedral*) and texture (carpet and picture in *room*).

7.5.2. Quantitative results

Quantitative results for the synthetic dataset are presented first. For each model-image pair, the relative repeatability is computed using the top-k 2D points and the top-2k 3D points (since it is expected half the 3D points will be occluded by the model), for k varying between 20 and 100. It is computed for inlier thresholds (t) of 3 and 6 pixels and averaged across all images of the model. Results are given in Fig. (13), where, given the 3D data is untextured, a comparison is made between *Harris-Harris, SIFT-SIFT-G, GFT-Harris, KBD-SURE, KBI-KB-G*, and *KBD-KB-G*.

It is observed that, in general, *GFT-Harris* and *KBD-KB-G* perform the best; between them having the highest repeatability across all six models. Both have repeatabilities of at least 30% for (relatively) large numbers of points; sufficiently high for subsequent 2D-3D registration. *KBI-KB-G* performs quite well, but never as well as *KBD-KB-G*. This is perhaps surprsing in comparison to the results of *KBI* on the 2D-2D evaluation (Fig. 10) - the derivative-based KB formulation is evidently more indicative of geometry rather than texture based on these results. *Harris-Harris, SIFT-SIFT-G, KBD-SURE*, and *KBI-KB-G* perform similarly poorly, rarely obtaining a repeatability of above 20%. Comparing between 3 pixels and 6 pixels as the inlier threshold; *GFT-Harris* performs slightly better than *KBD-KB-G* for the smaller threshold, the reverse is true of the larger threshold. However, the increase in inlier threshold from 3 to 6 typically results in a repeatability increase by a factor of around 2, regardless of detector or dataset.

Fig. 13 shows that, in general, the repeatability increases with respect to the number of points detected. However, this is not the case with *GFT-Harris* which, in some circumstances, shows a decrease in repeatability for increasing numbers of points - particularly so on the *armadillo*, and to a lesser extent on the *dino* and *dragon*. Fig. 14 shows qualitative results on the *armadillo* for *GFT-Harris* and *KBD-KBG* for smaller quantities of points. For very small quantities of points (20 in 2D and 40 in 3D) *GFT-Harris* has a high correlation due to the relatively small number of well-defined corners on the model (toes, fingers, and ears) and hence the relative ease at which they are detected by a corner detector. For a higher quantity of features (60 in 2D and 120 in 3D) there are insufficient corners in the scene and so it becomes unclear why certain features should be detected by the corner detectors. By contrast, our saliency-based approach is more broadly defined than a corner detector allowing *KBD* and *KBG* to admit a wider range of features. As a result, it is relatively unlikely our approach will have a higher repeatability for small numbers of features (since salient points are not as narrowly defined as corner points) but conversely the definition of saliency extends to larger numbers of features.

Next, quantitative results for the real dataset are presented. For each model-image pair, the relative repeatability is computed using the top-k 2D points and the top-2k 3D points, with the exception of the larger *courtyard* and *reception* datasets where the top-4k 3D points are used, since here it is expected the majority of the 3D points will not be projected onto the image. k is varied up from 20 to 200. Similarly to the synthetic dataset, the relative repeatability is computed for inlier thresholds of 3 and 6 pixels.

Results are presented in Fig. 15, where a comparison is made between all 11 approaches (as described at the beginning of Section 7). Between the different models, the best results are obtained on *reception* and *room*, with repeatability rates of over 30% in some cases. However, the other three models only obtain repeatability rates of between 15% and 25%. Between the different point detectors, *KBD-KBD-T* and *KBD-KBD-B* generally perform the best across all models. *GFT-Harris* performs nearly as well except on the more textured models *room* and *studio*. *KBI-KBI-T* more often outperforms *KBI-KBI-B*, further demonstrating that *KBI* does not detect geometrically significant features in 2D. Similarly to the synthetic dataset, *SIFT-SIFT-G Harris-Harris*, and *KBD-SURE* do not perform well in general.

As a comparison between the methods proposed here (*KBD-KB-G, KBD-KB-T*, and *KBD-KB-B*), *KBD-KB-G* generally does not perform as well except on the *cathedral* model. It is perhaps surprising that *KBD-KB-T* consistently performs well, particularly on *courtyard* and *reception* where there is little discriminating texture; however as observed in the qualitative results, geometric features are often accompanied by a change in texture. Furthermore, the scale selection process within the *KB* detector allows it to naturally avoid repetitive parts of a scene. *KBD-B* consistently performs well regardless of the scene, outperforming the other approaches on the *cathedral* and *studio*.

8. Conclusions and future work

In this paper we have presented a general approach to 2D-3D salient point feature detection, based on the information-

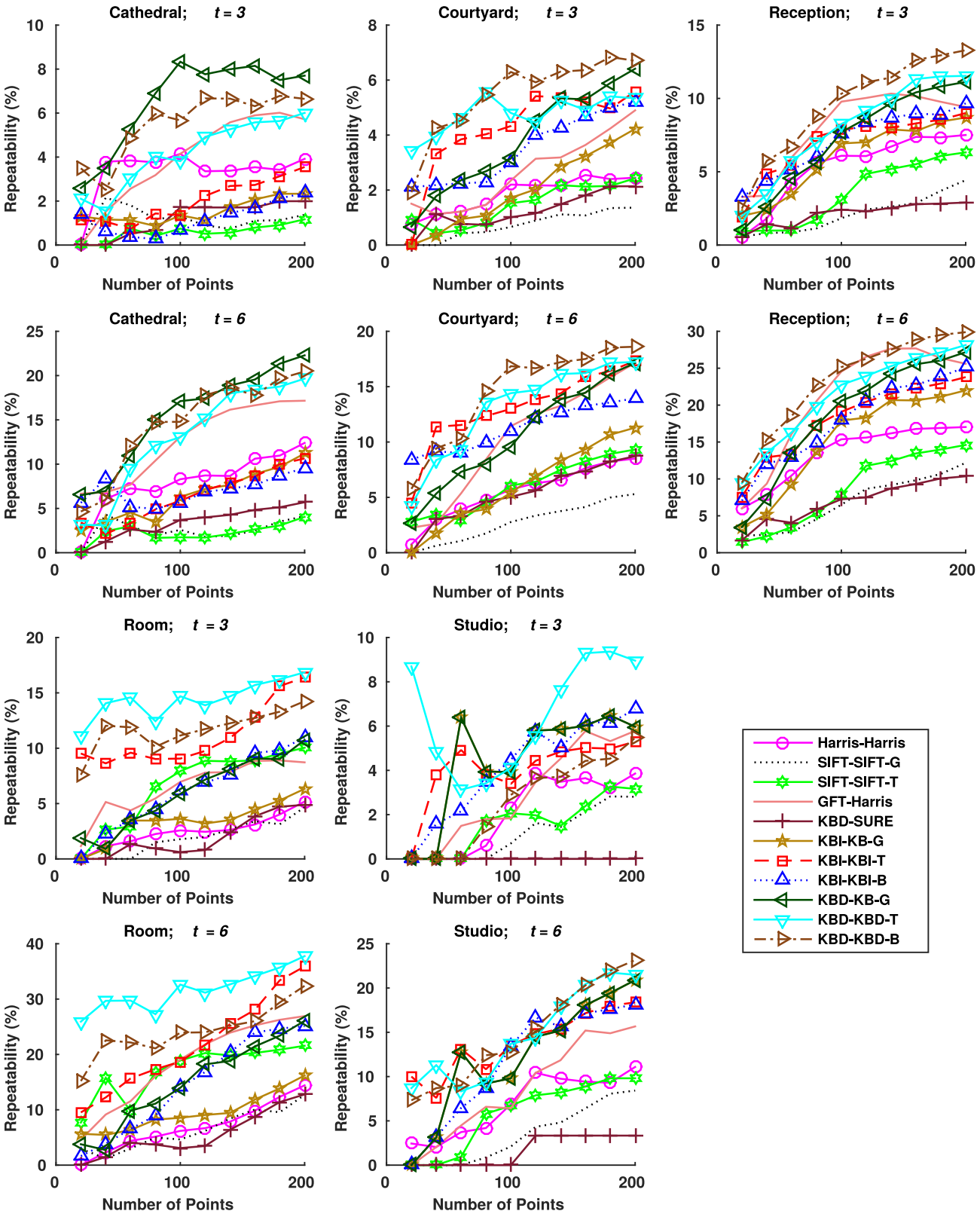

Fig. 15. Results on the real dataset. On the left shows the relative repeatability of the detectors for an inlier threshold of 3 pixels; on the right an inlier threshold of 6 pixels is used. k varies between 20 and 200. The graphs are ordered such that a graph of inlier threshold 3 pixels is shown above that of inlier threshold 6 pixels.

theoretic Kadir-Brady saliency detector (Kadir and Brady, 2001). The histogram-based framework developed allows for a unified approach to feature detection in 2D, and both textured and untextured 3D data. Intensity-based and derivative-based approaches were proposed, where the derivative-based approaches were shown to be superior since image derivatives are more indicative of the underlying geometry of the scene. The results also show the proposed approach to be more repeatable than existing feature detectors that have 2D and 3D implementations (Harris and SIFT) across a range of image and LiDAR data, from both indoor and outdoor scenes. Furthermore, its ability to naturally operate on textured or untextured 3D data allow the approach to detect features based on both attributes simultaneously, increasing its robustness and widening its applicability.

There is scope for improvement in our method; in particular, the qualitative results show our approach to occasionally detect edges as salient. While there may be some salient properties regarding the edges, a point on an edge is not well localised along the edge and may not be as useful for geometry estimation. This could be addressed in a similar manner to Tombari and di Stefano (2014) where histograms are compared between neighbouring points, rather than between neighbouring scales. Alternatively, one may consider other attributes to construct a histogram from, other than the first derivatives of the image. However, while the second derivatives of the image have had considerable success in feature detection via SIFT (Lowe, 2004), the blob-like features they detect are generally more indicative of texture rather than geometry.

Future work will include the registration of points between images and 3D LiDAR data. In many cases, correspondences between features cannot be automatically determined, and need to be established alongside registration parameters. It is a computationally expensive problem (Moreno-Noguer et al., 2008), so any method that has a high repeatability for a smaller number of points will be more suited to this kind of problem. We furthermore plan to integrate our approach with line features (Brown et al., 2015) detected in both 2D and 3D, so as to obtain a more complete scene description and make the subsequent registration process more robust due to the complementarity of these features.

Research data

The authors confirm that the indoor and outdoor 2D-3D datasets generated as part of this research are freely available under the terms and conditions detailed in the license agreement enclosed in the data repositories. Details of the data and how to obtain access are available for the *Room* dataset at Klaudiny et al. (2014); and for the *Cathedral, Courtyard, Reception*, and *Studio* datasets at (Kim, 2014).

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (grant number EP/K503186/1) and the European Commission FP7 IM-PART pro-ject (grant number 316564).

Appendix A. Effect of scale parameter setting on repeatability

Here we present repeatability results when varying the choice of σ_1 in both 2D and 3D. The results in 2D are shown in Fig. 16 comparing results for *KBO* and *KBD* on the 2D-3D synthetic dataset. The results for *KBO* show some variability depending on the choice of σ_1, with better results observed on the *buddha* and the *dragon* with $\sigma_1 = 4$ but this choice of parameter gave worse results on the *dino*. The choice of σ_1 makes very little difference on *KBD* however.

The results for varying σ_1 in 3D are given in Fig. 17. The choice of σ_1 affects the different approaches in a very similar way, with $\sigma_1 = 0.3\%$ the diameter of the bounding box giving the poorest results and $\sigma_1 = 0.5\%$ giving slightly stronger results than $\sigma_1 = 0.4\%$.

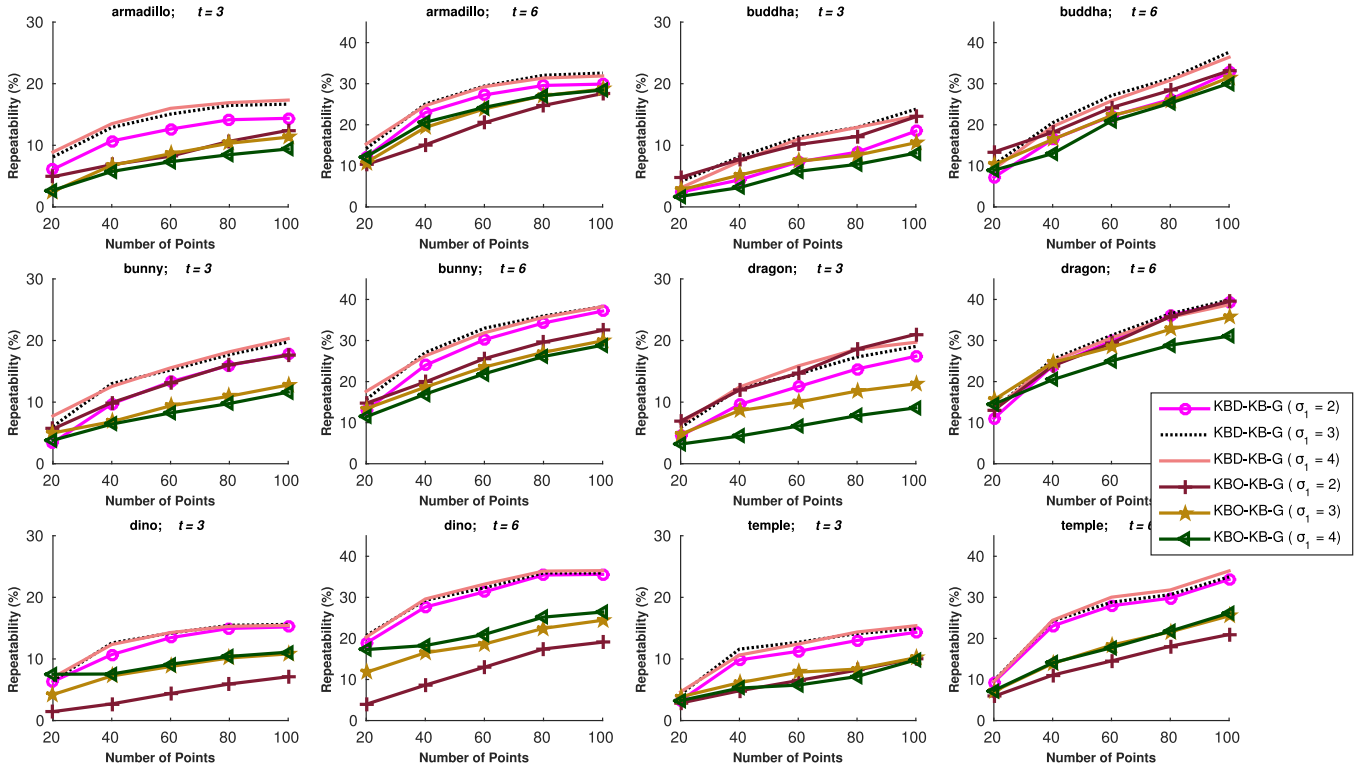

Fig. 16. 2D-3D repeatability results where σ_1 is varied between 2, 3, and 4 pixels in 2D. Only *KBO* and *KBD* are shown here because the other 2D feature detectors use a different approach to scale selection. The default parameter is used for scale selection in 3D (Section 7.1) in these experiments.

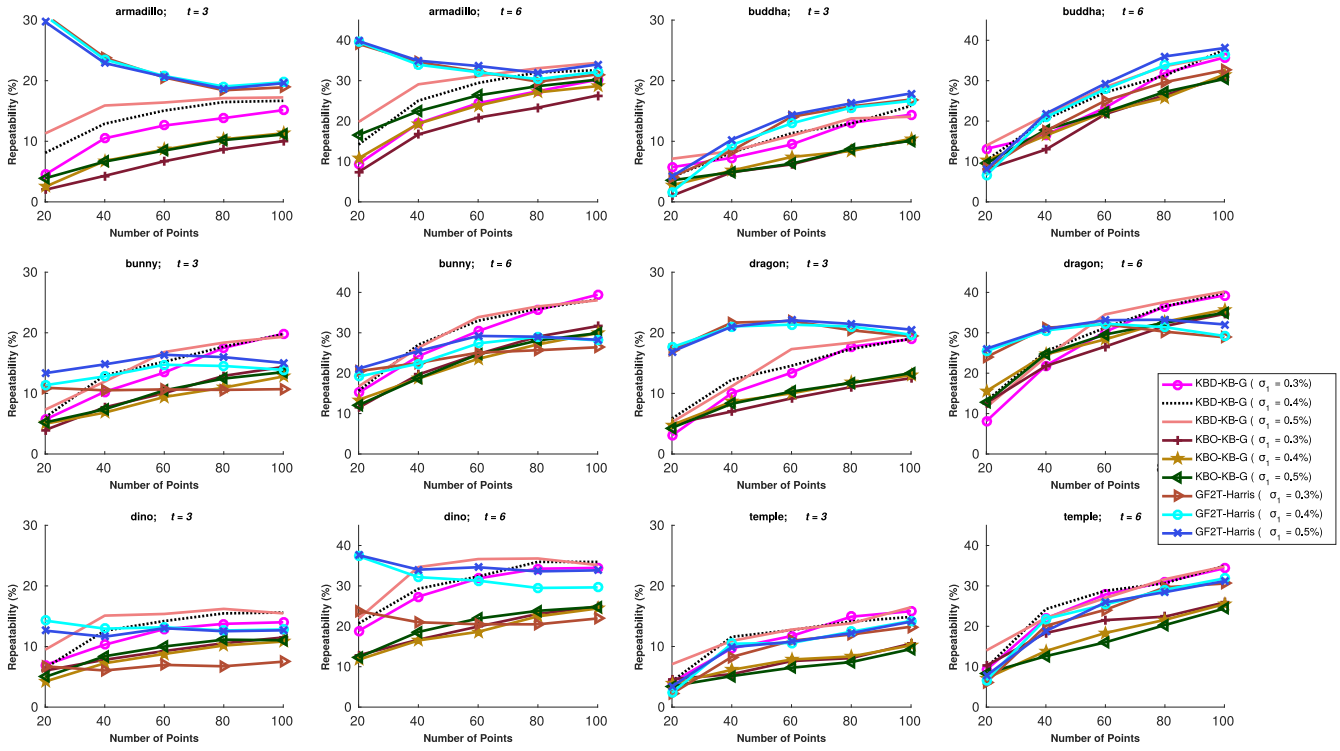

Fig. 17. 2D-3D repeatability results where σ_1 is varied between 0.3% and 0.5% of the diameter of the bounding box in 3D. The default parameter is used for scale selection in 2D (Section 7.1) in these experiments.

These results demonstrate that our choice of σ_1, while not optimised per dataset, gives a relative indication of the performance of the approaches and hence supports the overall conclusions of this paper.

References

Aanæs, H., Dahl, A.L., Pedersen, K.S., 2012. Interesting interest points - a comparative study of interest point performance on a unique data set. Int. J. Comput. Vision 97 (1), 18–35.

Alcantarilla, P.F., Bartoli, A., Davison, A.J., 2012. Kaze features. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part VI, pp. 214–227.

Arvo, J., 1992. Fast random rotation matrices. pp. 117–120.

Beaudet, P.R., 1978. Rotationally invariant image operators. In: Proceedings of the 4th International Joint Conference on Pattern Recognition, pp. 579–583.

Brown, M., Guillemaut, J.-Y., Windridge, D., 2014. A saliency-based approach to 2d-3d registration. In: Proc. International Conference on Computer Vision Theory and Applications (VISAPP). Available at http://personal.ee.surrey.ac.uk/Personal/J.Guillemaut/publications/14/BrownVISAPP14.pdf.

Brown, M., Windridge, D., Guillemaut, J.-Y., 2015. A generalisable framework for saliency-based line segment detection. Pattern Recognit. 48 (12), 3993–4011.

Castellani, U., Cristani, M., Fantoni, S., Murino, V., 2008. Sparse points matching by combining 3d mesh saliency with statistical descriptors. Comput. Graphics Forum 27 (2), 643–652.

Chen, H., Bhanu, B., 2007. 3D free-form object recognition in range images using local surface patches. Pattern Recognit. Lett. 28 (10), 1252–1262.

Filipe, S., Alexandre, L.A., 2014. A comparative evaluation of 3d keypoint detectors in a RGB-D object dataset. In: VISAPP Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 1, pp. 476–483.

Fiolka, T., Stückler, J., Klein, D.A., Schulz, D., Behnke, S., 2012. Sure: Surface entropy for distinctive 3d features. In: Spatial Cognition VIII. In: Lecture Notes in Computer Science, 7463, pp. 74–93.

Flitton, G., Breckon, T., Megherbi Bouallagu, N., 2010. Object recognition using 3d sift in complex ct volumes. In: Proceedings of the British Machine Vision Conference, pp. 11.1–11.12.

Guillemaut, J.-Y., Hilton, A., 2011. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. Int. J. Comput. Vision 93 (1), 73–100.

Guo, Y., Bennamoun, M., Sohel, F.A., Lu, M., Wan, J., 2014. 3D object recognition in cluttered scenes with local surface features: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 36 (11), 2270–2287.

Guo, Y., Sohel, F.A., Bennamoun, M., Lu, M., Wan, J., 2013. Rotational projection statistics for 3d local surface description and object recognition. Int. J. Comput. Vision 105 (1), 63–86.

Hänsch, R., Weber, T., Hellwich, O., 2014. Comparison of 3D interest point detectors and descriptors for point cloud fusion. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. Vol. II-3, 57–64.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, pp. 147–151.

Hauagge, D.C., Snavely, N., 2012. Image matching using local symmetry features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 206–213.

Kadir, T., Brady, M., 2001. Saliency, scale and image description. Int. J. Comput. Vision 45 (2), 83–105.

Kadir, T., Zisserman, A., Brady, M., 2004. An affine invariant salient region detector. In: Proceedings of the 8th European Conference on Computer Vision. Springer, pp. 228–241.

Kim, H., 2014. Impart multi-modal dataset. http://dx.doi.org/10.15126/surreydata.00807707.

Klauduiny, M., Tejera, M., Malleson, C., Guillemaut, J.-Y., Hilton, A., 2014. Scene digital cinema datasets. http://dx.doi.org/10.15126/surreydata.00807665.

Lee, C.H., Varshney, A., Jacobs, D.W., 2005. Mesh saliency. ACM Trans. Graph. 24 (3), 659–666.

Lee, W.-T., Chen, H.-T., 2009. Histogram-based interest point detectors. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1590–1596.

Li, Y., Snavely, N., Huttenlocher, D.P., 2010. Location recognition using prioritized feature matching. In: Proceedings of the 11th European Conference on Computer Vision: Part II, pp. 791–804.

Li, Y., Wang, S., Tian, Q., Ding, X., 2015. A survey of recent advances in visual feature detection. Neurocomputing 149 (Part B), 736–751.

Liu, L., Stamos, I., 2012. A systematic approach for 2d-image to 3d-range registration in urban environments. Comput. Vision Image Understanding 116 (1), 25–37.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60 (2), 91–110.

Mastin, A., Kepner, J., Fisher III, J.W., 2009. Automatic registration of lidar and optical images of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition.

Matas, J., Chum, O., Urban, M., Pajdla, T., 2002. Robust wide baseline stereo from maximally stable extremal regions. In: British Machine Vision Conference, pp. 36.1–36.10.

Maver, J., 2010. Self-similarity and points of interest. IEEE Trans. Pattern Anal. Mach. Intell. 32 (7), 1211–1226.

Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. Int. J. Comput. Vision 60 (1), 63–86.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V., 2005. A comparison of affine region detectors. Int. J. Comput. Vision 65 (1–2), 43–72.

Moreno-Noguer, F., Lepetit, V., Fua, P., 2008. Pose priors for simultaneously solving alignment and correspondence. In: Proceedings of the 10th European Conference on Computer Vision: Part II, pp. 405–418.

Penate-Sanchez, A., Andrade-Cetto, J., Moreno-Noguer, F., 2013. Exhaustive linearization for robust camera pose and focal length estimation. IEEE Trans. Pattern Anal. Mach. Intell. 35 (10), 2387–2400.
Petrelli, A., di Stefano, L., 2012. A repeatable and efficient canonical reference for surface matching. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, pp. 403–410.
Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 519–528.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Shao, L., Brady, M., 2006. Invariant salient regions based image retrieval under viewpoint and illumination variations. J. Visual Commun. Image Represent. 17 (6), 1256–1272.
Shao, L., Kadir, T., Brady, M., 2007. Geometric and photometric invariant distinctive regions detection. Inf. Sci. 177 (4), 1088–1122.
Shechtman, E., Irani, M., 2007. Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition.
Shi, J., Tomasi, C., 1994. Good features to track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600.
Sipiran, I., Bustos, B., 2010. A robust 3d interest points detector based on harris operator. In: Proceedings of the 3rd Eurographics Conference on 3D Object Retrieval, pp. 7–14.
Smith, S.M., Brady, J.M., 1997. Susan -a new approach to low level image processing. Int. J. Comput. Vision 23 (1), 45–78.
Teran, L., Mordohai, P., 2014. 3d interest point detection via discriminative learning. In: Proceedings of the 13th European Conference on Computer Vision: Part II. In: Lecture Notes in Computer Science, 8689, pp. 159–173.
Tombari, F., Franchi, A., di Stefano, L., 2013a. BOLD features to detect texture–less objects. In: IEEE International Conference on Computer Vision, ICCV, pp. 1265–1272.
Tombari, F., Salti, S., Di Stefano, L., 2013b. Performance evaluation of 3d keypoint detectors. Int. J. Comput. Vision 102 (1–3), 198–220.
Tombari, F., di Stefano, L., 2014. Interest points via maximal self-dissimilarities. In: Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Part II. In: Lecture Notes in Computer Science, 9004, pp. 586–600.
Tuytelaars, T., Mikolajczyk, K., 2008. Local invariant feature detectors: A survey. Found. Trends Comput. Graph. Vision 3 (3), 177–280.
Vedaldi, A., Fulkerson, B., 2008. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/.
Wang, L., Neumann, U., 2009. A robust approach for automatic registration of aerial images with untextured aerial lidar data. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2623–2630.
Wu, C., Clipp, B., Li, X., Frahm, J.-M., Pollefeys, M., 2008a. 3d model matching with viewpoint-invariant patches (vip). In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.
Wu, C., Fraundorfer, F., Frahm, J.-M., Pollefeys, M., 2013b. 3d model search and pose estimation from single images using vip features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008., pp. 1–8.
Yu, T.-H., Woodford, O.J., Cipolla, R., 2013. A performance evaluation of volumetric 3d interest point detectors. Int. J. Comput. Vision 102 (1–3), 180–197.
Zaharescu, A., Boyer, E., Horaud, R., 2012. Keypoints and local descriptors of scalar functions on 2D manifolds. Int. J. Comput. Vision 100 (1), 78–98.
Zhong, Y., 2009. Intrinsic shape signatures: a shape descriptor for 3d object recognition. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, pp. 689–696.

Mark Brown received a first class honours degree in Mathematics from the University of Bath, UK, in 2012. He is currently a PhD student at the University of Surrey, researching techniques for multi-modal data registration for digital film production. His research interests include feature detection and geometry estimation.

Dr. David Windridge is Senior Lecturer in Computer Science at Middlesex University and leads the University's Data Science activities. His research interests centre on machine learning, cognitive systems and computer vision. He has authored and played a leading role on a number of large-scale machine learning projects in academic and industrial research settings, and has also won various interdisciplinary research grants. He is a Visiting Professor at Trento University, Italy, and a Visiting Senior Lecturer at the University of Surrey (where he was previously a Senior Research Fellow). He has authored around 100 academic publications with $\sim$ 500 citations.

Jean-Yves Guillemaut received the MEng(hons) degree from the Ecole Centrale de Nantes, France, in 2001 and the PhD degree from the University of Surrey, UK, in 2005. He is a Lecturer in 3D Computer Vision at the University of Surrey. His research interests include 3D scene modelling, multi-modal data registration, image/video segmentation and matting, 3D video and animation. His current research activities focus on developing novel video-based modelling techniques suitable for the reconstruction of outdoor scenes or scenes with complex surface reflectance properties. He has co-authored over 60 peer-reviewed publications.