Int J Comput Vis
DOI 10.1007/s11263-011-0473-8

Interesting Interest Points

A Comparative Study of Interest Point Performance on a Unique Data Set

Henrik Aanæs · Anders Lindbjerg Dahl · Kim Steenstrup Pedersen

Received: 11 September 2010 / Accepted: 2 June 2011
© Springer Science+Business Media, LLC 2011

Abstract Not all interest points are equally interesting. The most valuable interest points lead to optimal performance of the computer vision method in which they are employed. But a measure of this kind will be dependent on the chosen vision application. We propose a more general performance measure based on spatial invariance of interest points under changing acquisition parameters by measuring the spatial recall rate. The scope of this paper is to investigate the performance of a number of existing well-established interest point detection methods. Automatic performance evaluation of interest points is hard because the true correspondence is generally unknown. We overcome this by providing an extensive data set with known spatial correspondence. The data is acquired with a camera mounted on a 6-axis industrial robot providing very accurate camera positioning. Furthermore the scene is scanned with a structured light scanner resulting in precise 3D surface information. In total 60 scenes are depicted ranging from model houses, building material, fruit and vegetables, fabric, printed media and more. Each scene is depicted from 119 camera positions and 19 individual LED illuminations are used for each position. The LED illumination provides the option for artificially relighting the scene from a range of light directions. This data set has given us the ability to systematically evaluate the performance of a number of interest point detectors. The highlights of the conclusions are that the fixed scale Harris corner detector performs overall best followed by the Hessian based detectors and the difference of Gaussian (DoG). The methods based on scale space features have an overall better performance than other methods especially when varying the distance to the scene, where especially FAST corner detector, Edge Based Regions (EBR) and Intensity Based Regions (IBR) have a poor performance. The performance of Maximally Stable Extremal Regions (MSER) is moderate. We observe a relatively large decline in performance with both changes in viewpoint and light direction. Some of our observations support previous findings while others contradict these findings.

Keywords Benchmark data set · Interest point detectors · Performance evaluation · Object recognition · Scene matching

H. Aanæs · A.L. Dahl (✉)
DTU Informatics, Technical University of Denmark, Lyngby, Denmark
e-mail: abd@imm.dtu.dk

H. Aanæs
e-mail: haa@imm.dtu.dk

K. Steenstrup Pedersen
E-Science Center, Image Group, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
e-mail: kimstp@diku.dk

Published online: 22 June 2011

1 Introduction

The ability to evaluate image similarity is found at the core of a wide range of computer vision problems, where local interest points provide a computational attractive representation for similarity measures. This has made methods for detecting interest points popular in many applications. The ability to match descriptors obtained from local interest points is based on the assumption that it is possible to find common interest points. For this to be useful for geometric reconstruction and similar applications, corresponding interest points have to be localized precisely on the same scene element, and the associated region around each interest point should cover the same part of the scene.

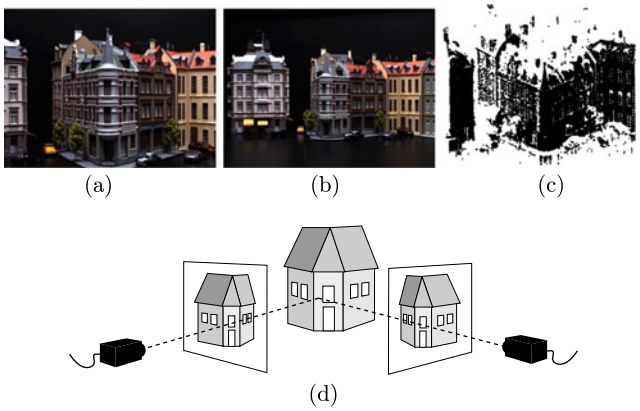


Fig. 1 Example of data and setup. Two images of the same scene with (**a**) one close up, (**b**) one distant from the side, and (**c**) the reconstructed 3D points. Illustration (**d**) of corresponding images with known geometric information including camera positions and 3D scene surface

The range of applications based on matching local image descriptors obtained from interest points includes object recognition (Lowe 2004), image retrieval (Nister and Stewenius 2006; Sivic and Zisserman 2006), and similar. For these types of applications the precision of the spatial position may appear less important. Often the relative spatial layout of interest points are used together with a tolerance for large variations in the corresponding points relative positions (Sivic et al. 2005). However in applications for 3D geometry reconstruction from interest points it is paramount to have a precise point correspondence (Snavely et al. 2008a, 2008b; Torr and Zisserman 1999; Furukawa and Ponce 2007).

It is common to distinguish between detecting interest points and computing the associated descriptor needed in order to evaluate similarity. This could indicate that the two steps are independent, see e.g. Mikolajczyk and Schmid (2005), Mikolajczyk et al. (2005). The question is, however, if this assumption of independence is reasonable. Interest points and the associated regions are found from salient image features, and the same image features will be part of the actual characterization. As a result the two parts are not completely independent, and the choice of interest point detector being a function of local image structure will influence the description of the region around the interest point. This will limit the subspace spanned by the descriptors and in this way reduce the specificity of the descriptor. We however, choose to focus on the detection step in order to avoid a complicated system where it is difficult to separate the effects of the different parts. An alternative to feature based interest points would be to pick the interest points at random, but it will be unlikely to obtain precise spatial correspondence between a sparse set of randomly picked points. The ability to detect corresponding interest points, in a precise and repeatable manner, is a desirable property for obtaining geometric scene structure. In this paper we will investigate exactly that property.

In general it is however hard to verify if correspondence exists between interest points, because it requires ground truth of the geometry of the observed scene.

Early work on correspondence from interest points and descriptors was based on rotation and scale invariant characterization (Lowe 2004; Schmid and Mohr 1997). Schmid et al. (2000) evaluated interest point detectors applied only to planar scenes. Later the interest points have been adapted to be invariant to affine transformation—an approximation to perspective distortion—thereby in principle making the characterization robust to large changes in viewpoint. These methods have been compared in Mikolajczyk et al. (2005), but the performance has been evaluated on quite limited data sets, consisting of eight scenes each containing six images. Furthermore, changes in viewing conditions are coupled with the scenes in that only two of the scenes are used for each viewing condition. However, the suggested evaluation criteria have since been used in numerous works together with this small data set.

The ground truth in the data from Mikolajczyk et al. (2005) was obtained by semi-manually fitting an image homography. As a consequence this limits the scene geometry to planar surfaces or images from a large distance where a homography is a good approximation. To address this issue Fraundorfer and Bischof (2004, 2005) generated ground truth by requiring that matched points should be consistent with the camera geometry across three views. In their study they investigate the same detectors as Mikolajczyk et al. (2005), but includes also the difference of Gaussian (DoG), Harris and Hessian detectors. Winder et al. (Hua et al. 2007; Winder et al. 2009; Winder and Brown 2007; Brown et al. 2011) studies the design of descriptors using results from Photo Tourism (Snavely et al. 2008a) as ground truth. Winder et al. only considers the DoG detector as implemented in the SIFT descriptor and a multi-scale Harris corner detector. Both the approaches of Fraundorfer and Bischof and Winder et al. use point matching to create ground truth which can be used to evaluate the matching of interest points. This can be problematic; if errors occur in the ground truth there can be a bias towards wrong correspondences in the proposed matching. As a result these wrong correspondences will not be detected.

Moreels and Perona (2007) evaluated interest point detectors and descriptors in a similar manner to the work of Fraundorfer and Bischof (2004, 2005) based on pure geometry by requiring three view geometric consistency with the epipolar geometry. They used an additional depth constraint based on knowledge about the position of their experimental setup. Hereby they obtained unique correspondence between 500–1000 interest points from each object. The studied detectors has an overlap with the previously

mentioned studies, but also including the Forstner detector (Forstner 1986) and the Kadir-Brady detector (Kadir et al. 2004). Their experiments also include limited changes in illumination in the form of 3 different lighting conditions. The focus of this study is different from ours in that Moreels and Perona (2007) consider the problem of object recognition, whereas we consider the problem of 3D reconstruction. In object recognition precise localization is not as important as in 3D reconstruction. Furthermore, a full recognition framework is needed in order to perform their evaluation, making it more difficult to separate the effects of different parts of the system, e.g. separate the effect of a particular choice for interest point detector from the choice of descriptor. The limitation of their experiment lies in relatively simple scenes with mostly single objects resulting in little self-occlusion. However, self-occlusion occurs very frequently in real world scenes and typically many interest points are found near occlusion boundaries.

We have compiled a large data set that provides a unique basis for this study. It consists of 60 scenes of varying object types, materials, and complexity of surface structures resulting in a total of 136,660 images. Figure 1 shows an example from our data set. The experimental setup consists of a camera mounted on an industrial 6-axis robot-arm, providing accurate and repeatable positioning. The scene is illuminated by 19 LED light sources. We capture an image with a single light source turned on, which allows us to do synthetic scene relighting in a controlled manner with a wide range of illumination scenarios simulating both indoor and outdoor environments. This is particularly relevant for studying performance of applications such as object recognition and image retrieval as well as computer vision applications in outdoor environments and under temporally changing lighting conditions. In addition, the scenes have been surface scanned using structured light, and, together with the camera positions, these scans supply ground truth for correspondence evaluation. As a result we can easily find corresponding interest points on the scene surface.

We evaluate ten established interest point detectors on this data set and provide new insight into the stability of these detectors with respect to large viewpoint and scale change as well as changes to the illumination conditions. The chosen detectors are Harris, Harris-Laplace, and Hessian-Laplace detectors and their two affine extensions—Harris-Affine and Hessian-Affine (Mikolajczyk and Schmid 2004; Mikolajczyk et al. 2005), Maximally Stable Extremal Regions (MSER) (Matas et al. 2004), Intensity Based Regions (IBR) and Edge Based Regions (EBR) (Tuytelaars and Van Gool 2004), the Fast corner detector (FAST) (Trajković and Hedley 1998), and the difference of Gaussian detector (DoG) (Crowley and Parker 1984; Lindeberg 1993; Lowe 1999, 2004). We recognize that this collection of detectors might not represent the complete state of the art and certainly does not cover all categories of approaches. However, they are all well-established methods commonly used in the computer vision literature and corresponds well with methods chosen in previous comparative studies (Schmid et al. 2000; Mikolajczyk et al. 2005; Mikolajczyk and Schmid 2005).

All methods investigated in this study are based on some form of extrema or zero crossing search in functionals of filter responses, and as such fall into what we could call the filter based category of detectors. In the interest of keeping the study focused and provide results comparable with previous comparative studies, we have opted not to include statistical or learning based approaches such as likelihood based approaches (Konishi et al. 2003a, 2003b; Laptev and Lindeberg 2003; Ren and Malik 2002; Ren et al. 2008), feature learning (Lillholm and Griffin 2008; Griffin et al. 2009), or outlier detection approaches (Lillholm and Pedersen 2004). Neither do we include methods based on more elaborate differential geometric definitions such as top points (Johansen et al. 1986, 2000; Nielsen and Lillholm 2001; Demirci et al. 2009).

1.1 Overview of Studied Detection Methods

The Harris corner detector was originally developed by Harris and Stephens (1988), but we use the scale-adapted Harris detector presented by Mikolajczyk and Schmid (2004). The Harris corner detector finds extrema in a corner measure based on the second moment matrix computed at fixed differentiation and integration scales, and tends to detect corner-like image structures. The Harris-Laplace (Mikolajczyk and Schmid 2004) detector is an extension of the scale-adapted Harris detector including scale selection based on extrema search in the Laplacian of Gaussian filter, an approach originally introduced by Lindeberg (1998b). The Hessian detector (Mikolajczyk et al. 2005) is based on extrema search in feature measures constructed from the Hessian matrix and its Laplacian extension includes the same scale selection approach of the Harris-Laplace detector. The Hessian detector tend to find blobs and ridges and was originally proposed by Lindeberg (1998b, 1998a). The affine extensions of both the Harris and Hessian detectors are based on the affine detection algorithm developed by Mikolajczyk and Schmid (2004), which estimate the affine shape of the interest point region using the second moment matrix. The DoG detector (Lowe 1999) is to some extent similar in spirit to the Hessian detector, because it approximates the Laplacian of Gaussian filter, which can be computed as the trace of the Hessian matrix. DoG tends to find interest points at isotropic blob structures.

In the EBR detector proposed by Tuytelaars and Van Gool (2004), both Harris corners and Canny edges (Canny

1986) are detected at multiple scales. From the Harris corner an affine region is extracted by tracing edges emanating from the corner point based on extrema search in a one-parameter family of functions of intensity moments.

The FAST corner detector proposed by Trajković and Hedley (1998) finds interest points by evaluating which of three types of image primitives the local image structure belongs to. This evaluation is based on intensity differences at crossing points between circles and lines emanating from the proposal point. The algorithm only use a limited set of scales, here represented by the radii of the circles surrounding the proposal point. This in effect should make this detector less invariant to scale changes.

MSER (Matas et al. 2004) and IBR (Tuytelaars and Van Gool 2004) are similar in spirit in that they produce regions around extremal intensities and both methods are affine invariant. IBR starts from points of local intensity extrema and detects region boundaries by tracing lines out from these points and finding extrema of a function of intensity differences along the lines. MSER detects region boundaries based on intensity thresholding.

We use the reference implementations provided by Lowe (2004), Mikolajczyk and Schmid (2005), Mikolajczyk et al. (2005), and will therefore not give further details of these methods but instead refer the reader to the papers describing the methods.

2 Contributions

The contributions of this paper are:

1. A comprehensive data set for precisely evaluating invariance properties of computer vision methods, especially with focus on geometry and recognition. The data set is freely available at our web site.[1]
2. A method for evaluating interest point detectors together with an evaluation of the ten most popular interest point detectors.
3. We evaluate the effect of view point and scale change as well as change in illumination, including both diffuse and directed lighting. Our study of the effect of illumination changes on interest point detectors are more comprehensive than previous studies (Moreels and Perona 2007).
4. Our major conclusions are:
   (a) that scale space based interest point detectors show the best performance—the exception being the fixed scale Harris corner detector which perform well, except not surprisingly in cases of large scale variations.
   (b) Large changes in view point angle and directional lighting has a devastating effect on the performance of the investigated methods. Especially, it seems that invariance to changing illumination conditions is an unsolved problem.
   (c) Contrary to previous claims in the literature (Mikolajczyk and Schmid 2004; Matas et al. 2004; Mikolajczyk et al. 2005), affine invariance has only little influence on the performance of interest point detectors, but it should be noted that such a contribution may occur when also taking interest point descriptors into account as part of the matching procedure.
   (d) Some of our results contradicts previous reported findings (Fraundorfer and Bischof 2004; Mikolajczyk et al. 2005) for some of the studied methods (see Sect. 6 for details). The main reason being that our data set is more realistically challenging than the previously used data sets.

This paper is an extension of our previous work published in Aanæs et al. (2010) including the details of the performed study as well as on the data set. Specifically, we have added the difference of Gaussian (DoG) detector to the study of interest point detectors under variation of view angle and distance to scene. Furthermore, we have added an analysis of the variation of the reported recall rate. Besides this we have also added an extensive study of the performance of the detectors with respect to varying illumination conditions. As a consequence we are able to answer several of the open questions posed in our previous conference paper.

The long-term goal of this study is to highlight successful approaches for interest point detection as well as identify potential avenues for future research in this area.

3 Data

The setup for data acquisition is illustrated in Fig. 2, and a detailed description of the data is available in Aanæs et al. (2009). The entire setup is enclosed in a black box and the scenes can be up to about half a meter, but the closest images depict about 25×35 cm. Scenes have been selected to show a large variation in scene type and they contain elements that are challenging for computer vision methods, like occlusions and various surface reflectance properties. There are 60 scenes with varying type of material and reflectance properties, including model houses, fabric, fruits and vegetables, printed media, wood branches, building material, and art objects. Image examples are shown in Fig. 3.

Color images of $1,200 \times 1,600$ pixels have been acquired, but for computational reasons we use 600×800 down-sampled versions in grayscale. The conversion to gray scale was done by $I_g = 0.299 I_R + 0.587 I_G + 0.114 I_B$, where I_g is the gray scale intensity and $I_{\{R,G,B\}}$ is the red,

[1] http://roboimagedata.imm.dtu.dk.

green and blue intensity, respectively. We have preprocessed the images to account for lens distortion by a warp based on bilinear interpolation. We also removed dark current noise by acquiring a dark frame with the same camera settings and subtracting it from the other frames.

Camera Positions For each scene we have acquired images from a precisely predefined camera path as illustrated in Fig. 4. This is possible because we employ a camera mounted industrial robot. The path is chosen relative to a central image position, which we refer to as the *key frame*. Our experiments are conducted with the key frame as a reference, so we compare all interest points found in other images to the key frame. An aim has also been to obtain the best 3D reconstruction of the scene when viewed from the key frame.

We have chosen a horizontal trajectory, so all positions are in the same plane, and for the house scenes this simulates a street view. This is chosen to avoid the robot shadowing the LEDs that are mounted in the roof. This setup provides a very accurate positioning of the camera with a standard deviation of approximately 0.1 mm. This corresponds to a standard deviation of 0.2–0.3 pixels when the point is back projected onto the images.

We have chosen 119 positions to have a dense sampling, which gives the opportunity for accurate evaluation of invariance properties of a method in relation to camera positions. In our first interest point evaluation experiment we have used all 119 positions, but such a dense sampling is in some cases not necessary. We have consequently chosen a subset of the positions for the light variation experiment.

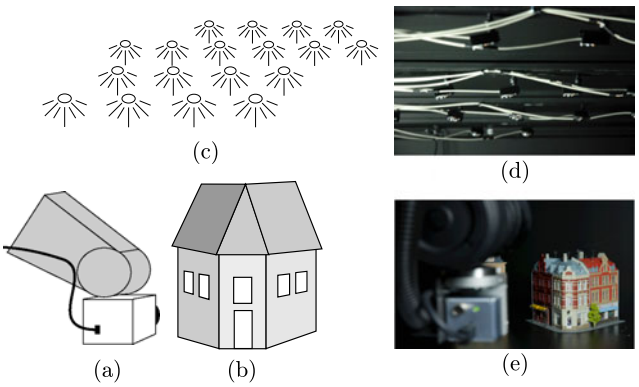


Fig. 2 Illustration of data collection setup. The camera (**a**) is mounted on a robot arm (**b**) capturing images of the scene. (**c**) LED point light sources illuminate the scene from 19 individual positions. (**d**), (**e**) photos of the experimental setup

Lighting The ability to evaluate detector methods robustness to light changes has been a central element in the design of our data acquisition setup. We have therefore chosen to use 19 individually controlled light emitting diodes (LEDs), which can be combined to provide a highly controlled and flexible light setting using image based relighting methods (see e.g. Einarsson et al. 2006; Haeberli 1992). Details and illustration of the setup is found in Fig. 5 and Table 1. The scene relighting is done by a linear combination of the directional illuminated images. We illuminate the scene according to a point, which gives us the light direction, and we use a Gaussian to weight the individual images. Choosing a large Gaussian will give a highly diffuse relighting, whereas a small Gaussian gives a directional relighting. An image $I_{\mathbf{x}}$ at position $\mathbf{x}$ is estimated by the linear combination $I_{\mathbf{x}} = \sum_{i=1}^{n} w_i I_i$, where the weight w_i is found by the Gaussian $w_i = c \exp(-\frac{(\mathbf{x}_i - \mathbf{x})^2}{2\sigma^2})$ and the scalar c is chosen such that $\sum_{i=1}^{n} w_i = 1$. σ is the parameter controlling the size of the Gaussian. In our directional relight experiment we choose $\sigma = 20$ and we used 19 LEDs ($n = 19$), and in our diffuse light experiment we choose to average all LEDs. It is important to note that the purpose of the relighting setup is to have controlled and repeatable relighting of the scenes. We did not strive at modeling a light source at approximate infinite distance, like the sun. Neither did we account for the distance of the diodes to the scene where the diodes just above the scene contribute with more light than the diodes at the sides. But the repeatability of the setup provides us with the same illumination for all scenes and simultaneously it provides a realistic light variation. The relighting has been done both from right to left and from back to front to illustrate the sensitivity of the investigated interest point detectors to changing lighting.

Surface Reconstruction We use structured light to obtain 3D surface geometry of the scenes. Figure 6 shows the setup

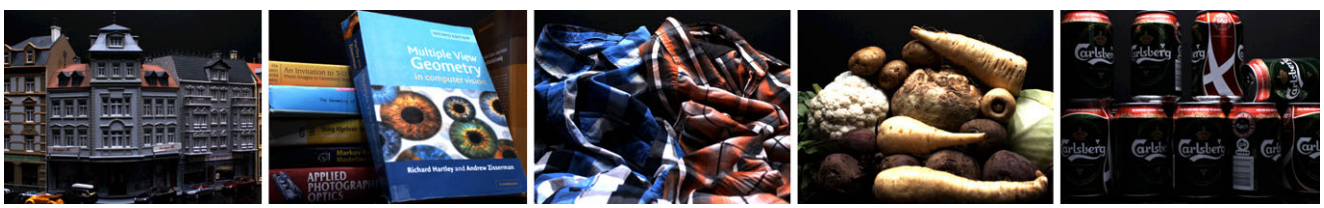


Fig. 3 Example images from our data set. The images show a diffuse relighting obtained by a linear combination of the 18 directional illuminated images. From *left* the scenes are examples of houses, books, fabric, greens, and beer cans, which have been used in our feature matching experiment with light variation

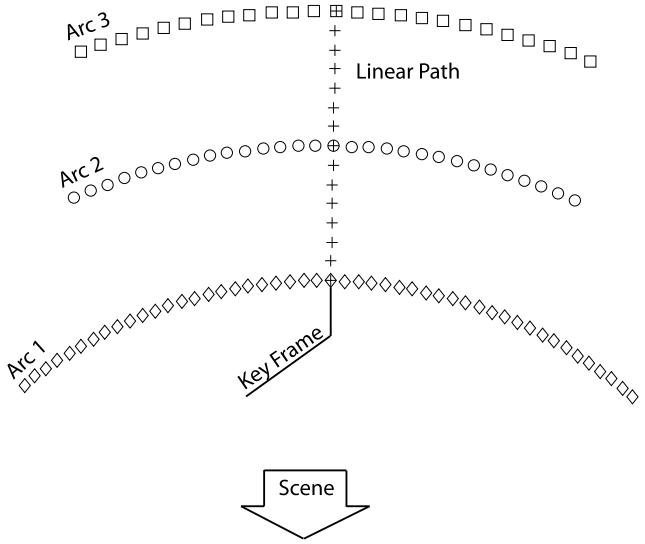


Fig. 4 Camera positions. The camera is placed in 119 positions in three horizontal arcs and a linear path away from the scene. The central frame in the nearest arc is the key frame, and the surface reconstruction is attempted to cover most of this frame. The three arcs are located on circular paths with radii of 0.5 m, 0.65 m and 0.8 m, which also defines the range of the linear path. Furthermore, Arc1 spans ±40°, Arc2 ±25° and Arc3 ±20°

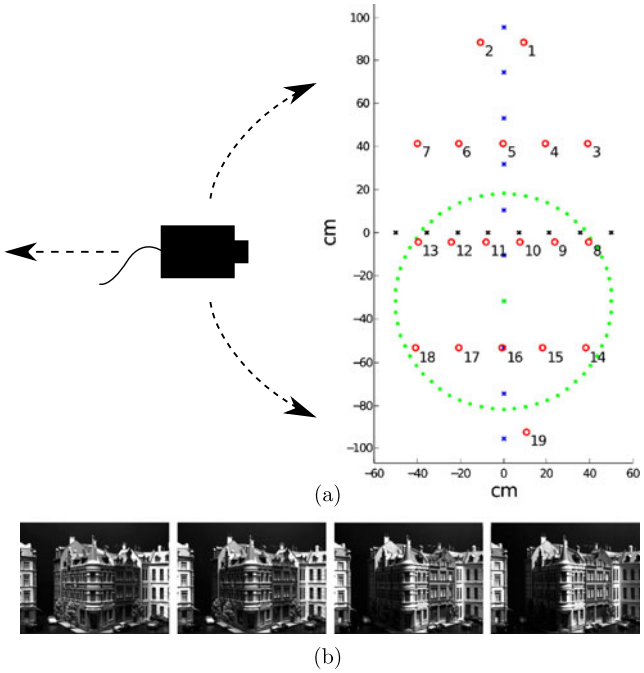


Fig. 5 (**a**) light stage setup seen from above and (**b**) example images with light from left to right. The layout of the light setup is illustrated with the *red circles* showing the positions of the white LEDs. The *axis* shown in (**a**) are in cm to illustrate the actual size of the setup, and azimuth and elevation angles can be seen in Table 1. The camera is placed to the *left* and an image is taken with one diode illuminated at a time. The crosses indicate relight sampling points from *left* to *right* (*blue*) and *back* to *front* (*black*). The images are weighted according to a Gaussian as shown with the green dots around the *green cross*. A large Gaussian will give more diffuse lighting whereas a small will give directional

for 3D surface reconstruction and an example of the point set data we obtain. The surface reconstruction is based on a stereo setup, and we use a binary stripe pattern to find correspondence between images. This method is recommended as one of the most reliable methods in both Scharstein and Szeliski (2003) and Salvi et al. (2004). We reconstruct the scene with a stereo pair from two distances to the scene to optimally cover the scene seen from the key frame. The obtained surface point sets contained outliers that were almost entirely single points with a large distance to all other points. They were easily removed by eliminating points with less than 3 other points within a distance of 1 mm. We obtain a varying number of surface points ranging from around 100,000 to 500,000 points depending on the size of the scene. The cleaned point sets are used directly in our matching procedure, so we avoid generating a triangular mesh, which could cause a bias in our performance estimates.

We verified the precision of the structured light reconstruction using a white spherical object—a bowling ball painted with white diffusive paint, and we measure the distance from the center of sphere to the surface. This gave an estimate of the surface reconstruction in the normal direction of the sphere. The advantage of a sphere is that it reveals error in all directions. We repeated the reconstruction of the sphere 10 times and we moved the projector between each scan. This gave a standard deviation of the radius estimate of 0.15 mm corresponding to a standard deviation of less than 0.6 pixels.

Table 1 Azimuth (ϕ) and elevation (θ) angles in degrees for all LEDs. The center of the coordinate system is the surface of the table where the scenes are placed

LED #	θ	ϕ	LED #	θ	ϕ
1	264°	57°	11	28°	86°
2	277°	57°	12	10°	80°
3	227°	68°	13	6°	74°
4	245°	72°	14	125°	65°
5	270°	73°	15	109°	68°
6	297°	72°	16	89°	69°
7	314°	68°	17	69°	68°
8	174°	74°	18	53°	64°
9	170°	80°	19	97°	56°
10	152°	86°			

4 Method

Our goal is to analyze invariance properties of interest points found in corresponding images. The design of our data set

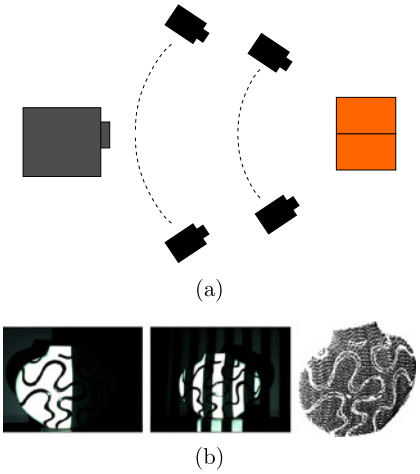


Fig. 6 Surface reconstruction is done with the setup shown in (**a**). We use a projector (*left*) to project a stripe pattern onto the scene (*right*), and we acquire images from four positions (*middle*). In (**b**) two stripe image examples are shown together with the reconstructed 3D point set (*right*)

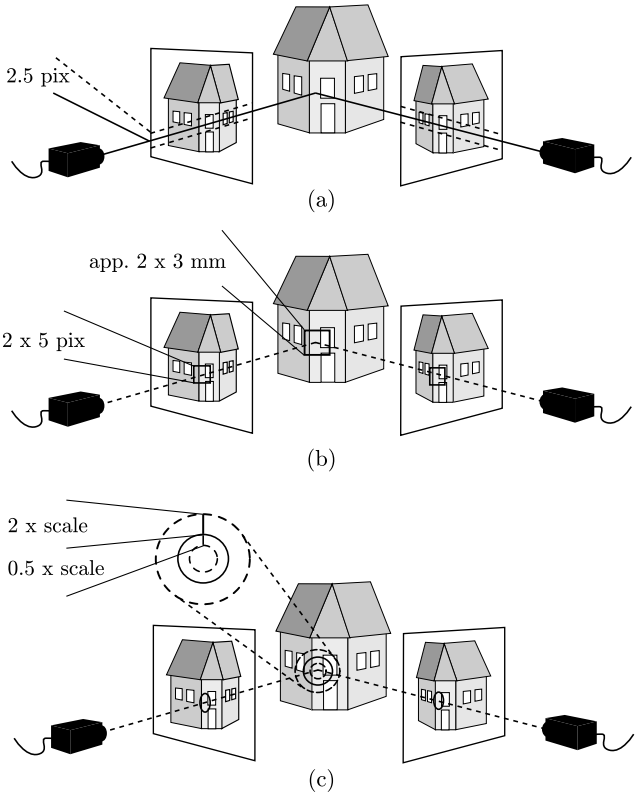


Fig. 7 Matching criteria for interest points. This figure gives a schematic illustration of a scene of a house and two images of the scene from two viewpoints. (**a**) The consistency with epipolar geometry, where corresponding descriptors should be within 2.5 pixels from the epipolar line. (**b**) Window of interest with a radius of 5 pixels and corresponding descriptors should be within this window, which is approximately 3 mm on the scene surface. Ground truth is obtained from the surface geometry. (**c**) The scale consistency, where corresponding descriptors are within a scale range factor of 2 from each other

enables us to answer questions like how do interest point detectors perform under change in view point? How many of the interest points are actually relevant? Are the detected interest points precisely located? Answers to these and related questions will provide an improved basis for choosing the appropriate methods for extracting interest points during computer vision system design. We will now provide the details of our analysis.

4.1 Evaluation Criteria

Evaluation of the performance of interest point detectors cannot be based on the associated descriptor, because the descriptors might not be unique. As a result it is impossible to tell if a given correspondence between similar looking image regions is correct or a mismatch. Therefore the evaluation has to be done independently of the interest point detection. Evidence for interest point correspondence is therefore obtained by fulfilling three criteria. We utilize the geometry of both the 3D scene surface and the camera positions to obtain this independent evaluation basis. Our evaluation criteria, with regard to pixel distances and scale, are based on a trade-off between as few double matches as possible and not eliminating points because of small variations in position of the interest points.

For each point in the key frame there has to be at least one interest point in the corresponding image fulfilling all three criteria, for the point to count as having a potential match. If more than one point fulfill all criteria it still counts as one potential match.

Epipolar Geometry Consistency with epipolar geometry is the first evaluation criterion. The camera positions of all images in our data set are known with high precision, which provides a basis for the relationship between points in one image and associated epipolar lines in another. This is used for removing false matches for a given interest point. We eliminate points that are further away than 2.5 pixels orthogonal to the epipolar line, as illustrated in Fig. 7(a).

The distance used for evaluating the epipolar constraint was computed as the back projection error of the estimated 3D point, corresponding to the match pair and their associated cameras, based on the Marquardt algorithm. It is noted that even though this 3D point estimate might be noisy, due to a very poor depth baseline ratio, this noise is due to unobservability and would as such *not* have an effect on the back projection error. This uncertainty of the 3D point estimate for short baselines is also the reason for excluding the estimate from the evaluation, e.g. as a distance to the structured light scan. Note also, that the back projection error is equal to the distance to the epipolar line, because for two cameras

this is the only source of back projection errors after it has been minimized.

Surface Geometry 3D surface reconstruction is used in the second evaluation criterion. Two points are considered a positive match if their 3D position is close to the scene surface obtained from the structured light reconstruction. This is fulfilled if there is a point from the surface reconstruction within a window of 10 pixels around a point, which corresponds to a box of approximately 6 mm on the scene surface. The surface reconstruction is not complete, so points in regions without surface reconstruction are discarded. However, only few points were removed due to this criterion. The surface geometry constraint is illustrated in Fig. 7(b).

Absolute Scale A region around each interest point provides the basis for an image descriptor. The interest points are detected in a multi-scale approach and the size of this region is dependent on what scale the interest point is detected. This image region corresponds to an area on the scene surface and corresponding descriptors should cover the same scene part. This area correspondence provides the basis for the third evaluation criterion, which is illustrated in Fig. 7(c), and the area of this region has to be within an area factor range of 0.5–2 of each other.

Parameter Choice The motivation behind the parameters used in our evaluation criteria is as follows: The image distance used for *epipolar geometry* is based on allowing for some inaccuracy in the interest point localization caused by image noise. 2.5 pixels is a standard setting for epipolar geometry threshold in a tracking algorithm corresponding to a variance of pixel position of a little more than 1.5 pixels (Hartley and Zisserman 2003). The distance used for the *surface geometry* also accounts for the effects of image noise, and the interpolation error between structured light points and the noise on the structured light points themselves. The latter was quantified by scanning objects of known geometry as described in Sect. 3. The threshold used for *absolute scale* was based on an expectation of the scale difference where a descriptor would obtain a similar characterization. To empirically validate that the tradeoff between false negatives and false positive was good and without apparent biases between detector types, we visually inspected multiple samples of interest point correspondences. In addition we counted the number of interest points a detector was matched to, where multiple matches indicated the false positive rate. Relaxing the thresholds too much would give many multiple matches, and harsh thresholds would give very few matches, indicating a high false negative rate. We found few double matches using the chosen parameter settings.

5 Experiments

We evaluate the performance of the ten interest point detectors using the recall rate, similar to the one used in Mikolajczyk and Schmid (2005), which is the ratio

$$\text{Recall} = \frac{\text{Potential Matches}}{\text{Total Interest Points}}.$$

The potential matches are points from the key frame fulfilling all three correspondence criteria. The total number of interest points is the number of interest points found in the key frame, see Fig. 4.

We have chosen the recall rate as a performance measure, because it measures the proportion of the interest points in the key frame that has a corresponding interest point in the compared frame. This measure is to some extent independent of the number of interest points detected in the key frame, because it measures the proportion of points. A very large number of interest points might give random correspondences, but it will be unlikely that random points fulfill all three correspondence criteria. If we were to measure the actual 3D precision of the interest points, we would first have to identify corresponding interest points, e.g. by applying the proposed three criteria, and then measure the distance of the interest points. Taking the uncertainty of the surface scan and the camera calibration into account, it is questionable if this distance measure will be accurate. Furthermore, if interest points are unprecisely found, a proportion will fall outside the correspondence criteria, and consequently the recall rate will to some extent also measure the precision of the 3D points. Based on this we have found the recall rate to be a good measure of performance.

Methods for interest point detection should ideally identify the same scene regions independently of camera position and illumination. As a result we have investigated the recall rate of the interest point detectors relative to variation in camera position and lighting over the 60 scenes in our data set. Furthermore, we have varied the input parameters for the methods to test if the algorithms are sensitive to parameter variation. First we will look at the detected number of interest points with the recommended parameter settings according to Lowe (2004), Matas et al. (2004), Mikolajczyk and Schmid (2004, 2005), Mikolajczyk et al. (2005), Trajković and Hedley (1998), and Tuytelaars and Van Gool (2004).

Number of Interest Points A varying number of interest points are detected in each data set, but this is highly dependent on the detection algorithm and the depicted scene. Table 2 and Fig. 8 shows the number of interest points and the standard deviation relative to the 60 scenes, where interest points have been extracted with the recommended parameter values. Some variation in number of interest points is expected, because of scene variation, but there is a noteworthy difference between the methods.

Table 2 Average number of interest points detected and the standard deviation over the 60 scenes

Detector	# Interest points	Std. interest points
Harris	925	665
Harris Laplace	736	538
Harris Affine	718	524
Hessian Laplace	1045	635
Hessian Affine	839	560
MSER	354	261
EBR	423	614
IBR	250	139
FAST	1539	1644
DoG	2236	1574

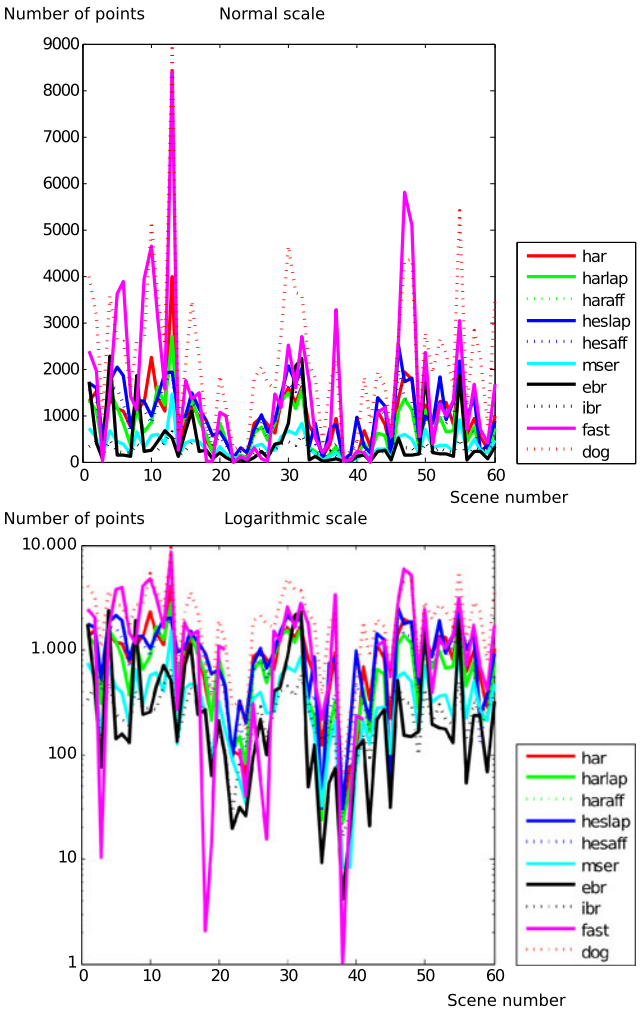


Fig. 8 Number of points in each scene for the different detector types. The horizontal axis shows the scene number and the vertical axis shows the number of points on (**a**) normal scale and (**b**) logarithmic scale. The high count outliers are especially clear in (**a**) and the low count outliers can be seen in (**b**). Note the varying number of interest points for the FAST corner detector, and for some scenes this detector has very few interest points, which is especially clear in (**b**)

Especially the FAST corner detector has some scenes with nearly 10,000 interest points and other scenes with close to 0. This is especially undesirable since it appears that scenes exist for which this algorithm will not work. At the same time other algorithms detect a reasonable amount of points for these scenes, indicating that the scenes is not degenerate, i.e. completely featureless. Also notice that for a lot of scenes FAST is an outlier to either side of the average points of all detectors. The DoG detector has a tendency to detect an above average number of interest points, and competes with FAST in detecting the most interest points on some scenes. The EBR also has a large variation, but much fewer interest points, and in general the IBR detects few interest points. Having few interest points is an undesirable property because it makes it hard to estimate the image correspondence. But also large fluctuations will result in unpredictable running time during matching, and especially a very large number of interest points can slow down the matching procedure. The Harris and Hessian corner detectors gives a reasonable number and variation of interest points, whereas MSER has relatively few points, but with a reasonable number in all scenes.

Recall and Position The recall rate of the interest point detectors as a function of the camera position is shown in Fig. 9. Interest point detectors are sensitive to the camera position, and both changing the view angle and the distance to the scene will reduce the recall. The question is what shape we can expect the curves to have.

The statistics of objects in ensembles of natural scenes exhibit statistical scale invariance (Srivastava et al. 2003). This has mainly to do with the fact that objects, or image structures, appears on all visible scales in the scale-space of the image. Empirical evidence of this is for instance seen in that the empirical distribution of area of homogeneous image segments follows a power law (Alvarez et al. 1999). A recent study (Gustavsson 2009) also shows that averaged over ensembles of scenes, this area distribution appears to be invariant to change of distance to the scene. Related to this observation, images of natural scenes also include large featureless areas such as e.g. sky areas in the horizon—this property is referred to as the "blue-sky effect" (Mumford and Gidas 2001). Even though our data set consists of indoor still-life scenes, we expect the scenes to exhibit scale invariance and as a consequence we expect the above mentioned power law behavior to be present in our scenes. Our scenes also include the "blue-sky effect" mainly because of the large black background area apparent in most scenes. Therefore, as a consequence of scale invariance we may deduce that as the camera moves away from the scene, small details, including potential interest points at low scales, will disappear (become smaller than pixel scale) in large numbers and merge into large scale structures. Furthermore, only

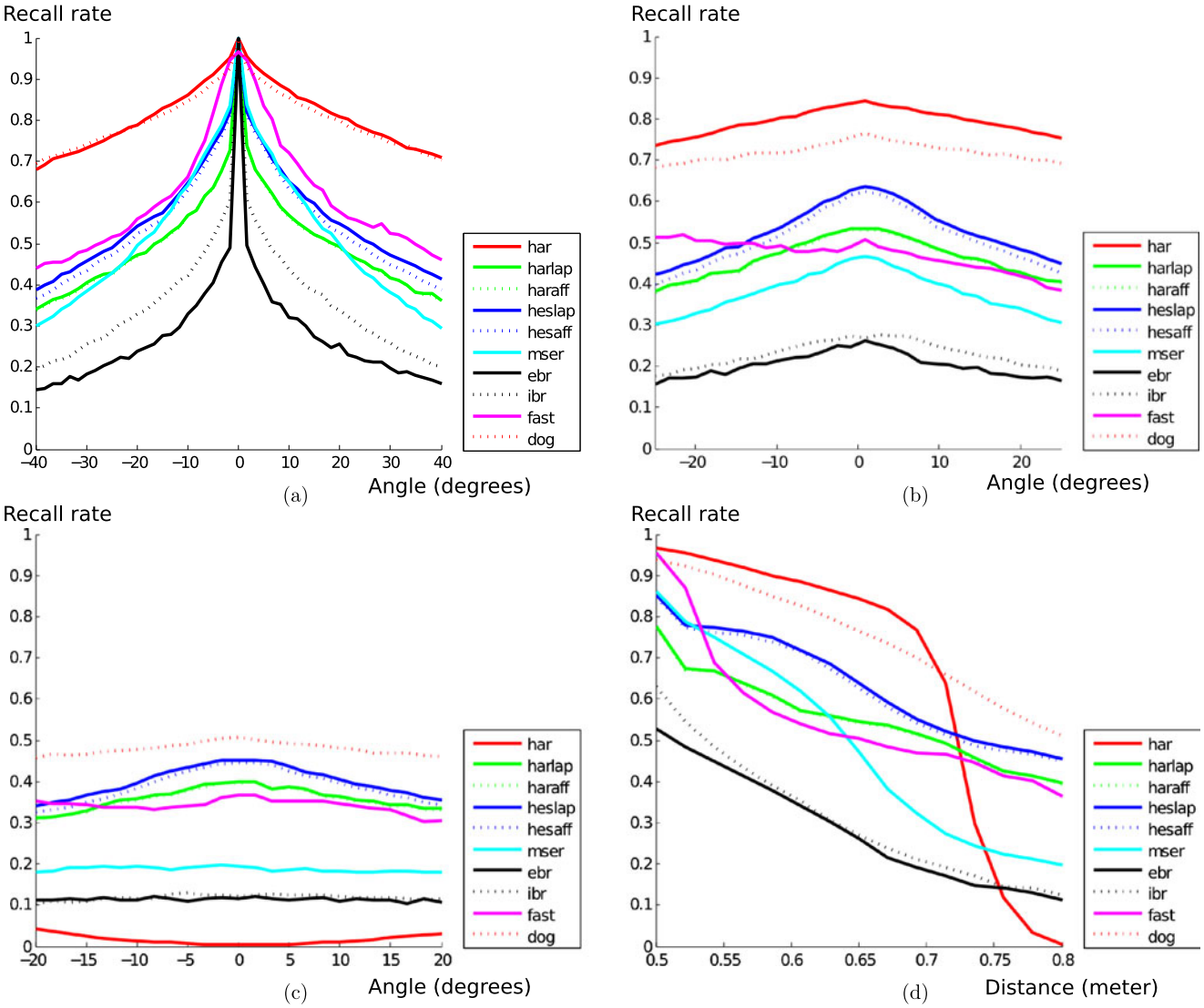


Fig. 9 Mean recall rate. The graphs show the recall rate relative to the paths shown in Fig. 4 with (**a**) Arc 1, (**b**) Arc 2, (**c**) Arc 3, and (**d**) Line Path. The *horizontal axis* is the angle relative to the scene in (**a**)–(**c**) and distance to the scene in (**d**). The *vertical axis* is the recall rate. Note that the recall rates for the FAST corner detector do not account for scale change (see text for details)

few new large-scale structures will appear leading to new potential large scale interest points due to the "blue-sky effect". Since the distribution of structure follows a power law, the consequence is that the number of matched interest points is expected to decrease as the viewing distance increases. This will in turn lead to a decrease of the recall rate. Hence for well-behaving interest point detectors we expect the number of interest points to follow a decreasing power law as a function of viewing distance, which also results in a decreasing recall rate. Furthermore, we have no reason to prefer certain view angles; hence we expect at least symmetry, if not view angle invariance, in the recall rate for well-behaved detectors when varying the view angle.

The shape of the curves in Fig. 9 are mainly as expected. The top performers are the Harris corner detector and the difference of Gaussian (DoG). Here the Harris corners perform slightly better than the DoG detector for moderate scale changes, but it has a sudden drop in recall rate at a distance of 0.7 m (Fig. 9(d)). This performance drop is caused by our scale matching criteria, which accepts a scale change of a factor two. Since the Harris corners do not incorporate scale, its performance will drop when the scale change exceeds this limit, and this is seen very clearly in our experiments.

The Hessian detectors perform overall well, but also the Harris Affine and Harris Laplace detectors have good performance. The recall rate is also high for the FAST corner detector, but this detector does not account for scale variation, so we cannot apply the third matching criterion. This favors the performance of the FAST detector, but we chose

to include it in our investigation to illustrate the large variability in performance. For small viewpoint changes FAST performs marginally better than other detectors, only beaten by DoG and Harris. However, the FAST detector exhibits asymmetries with respect to orientation (especially clear in Fig. 9(b)). An explanation for this asymmetry might be the large variation in the number of interest points detected in the various scenes. As mentioned in Sect. 1, FAST is by design not scale invariant, which accounts for the drop in performance seen in Fig. 9(d).

Notice that the ranking of the methods are preserved in the four graphs of Fig. 9, except for Harris, MSER, and FAST. Especially in Fig. 9(d), it is seen that these three methods deteriorates faster than the other methods as the distance to the scene increases.

Figure 9 shows the mean recall rate with an average taken over all 60 scenes. In addition to this we analyzed the variability of the performance by looking at the performance distribution or probability density functions (PDFs) for all 119 positions. A representative sample is shown in Fig. 10. From the overlap of these PDFs we can concluded that the DoG and Harris detectors are significantly better then the rest, which was also the observation from Fig. 9. We can furthermore see that the FAST detector at times has similar performance as the Harris and DoG detectors—especially for small viewpoint changes, but the performance is highly varying.

Changing Light In the light variation experiment our aim has been to reflect realistic light changes both in the direction of the light source and the diffuseness. In natural scenes light varies from being diffuse on an overcast day to highly directional in sunshine. To simulate this we vary the direction as shown in Figs. 11 and 12, and we have experimented with two levels of diffuseness—one with low and one with high degree of directional light. Both experiments show the same trend, but more pronounced for the high variation, so we have chosen to show results from that. Varying the light direction changes the scene surface appearances, which is seen in Figs. 11(c) and 12(c). It should be noted that we left out the FAST corner detector in this experiment, because of the missing scale information and its, in general, unreliable performance.

Ideally the interest point detection is invariant to change in light direction, but our experiments show, that this is far from the case. Our experiments is performed relative to the key frame (image number 25) illuminated from front, see the last image in Figs. 11(c) and 12(c). The light change is moderate, compared to what can be seen in natural scenes, but the reduction in performance of interest point detectors is significant. This performance reduction is similar to the effect of changing camera position, which comes as a surprise, since these variations are common in many natural

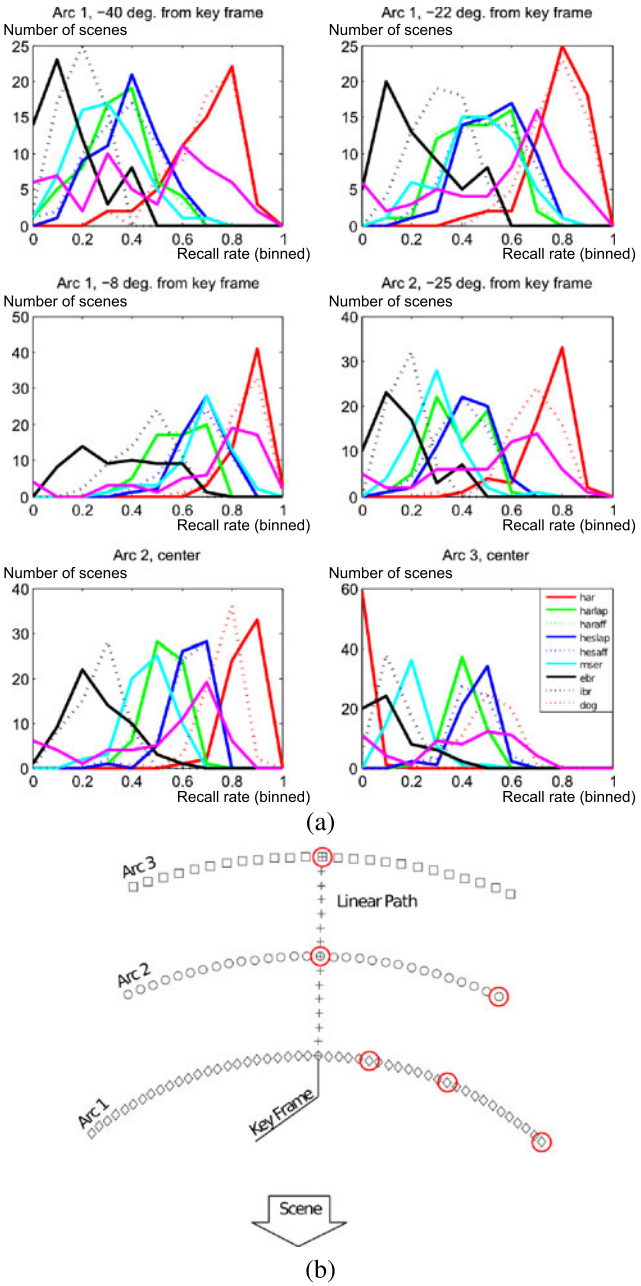


Fig. 10 (**a**) Probability distribution functions for selected image positions and (**b**) their positions on the path. The *colors* show the detector types—see Fig. 9. The *horizontal axis* shows binned recall rates and the *vertical axis* show number of scenes. This figure provides more detail in the performance of the image descriptors. Especially note how broad the distribution of the FAST corner detector that spans the range from very good to very poor performance

images. The curves have the same trend and their order are the same as in the experiment with diffuse light, see Fig. 9. This indicates that the different detectors relative sensitivity to light change is similar.

Lighting variation occurs in many applications based on interest point detection including examples like object recognition and image retrieval (Nister and Stewenius 2006;

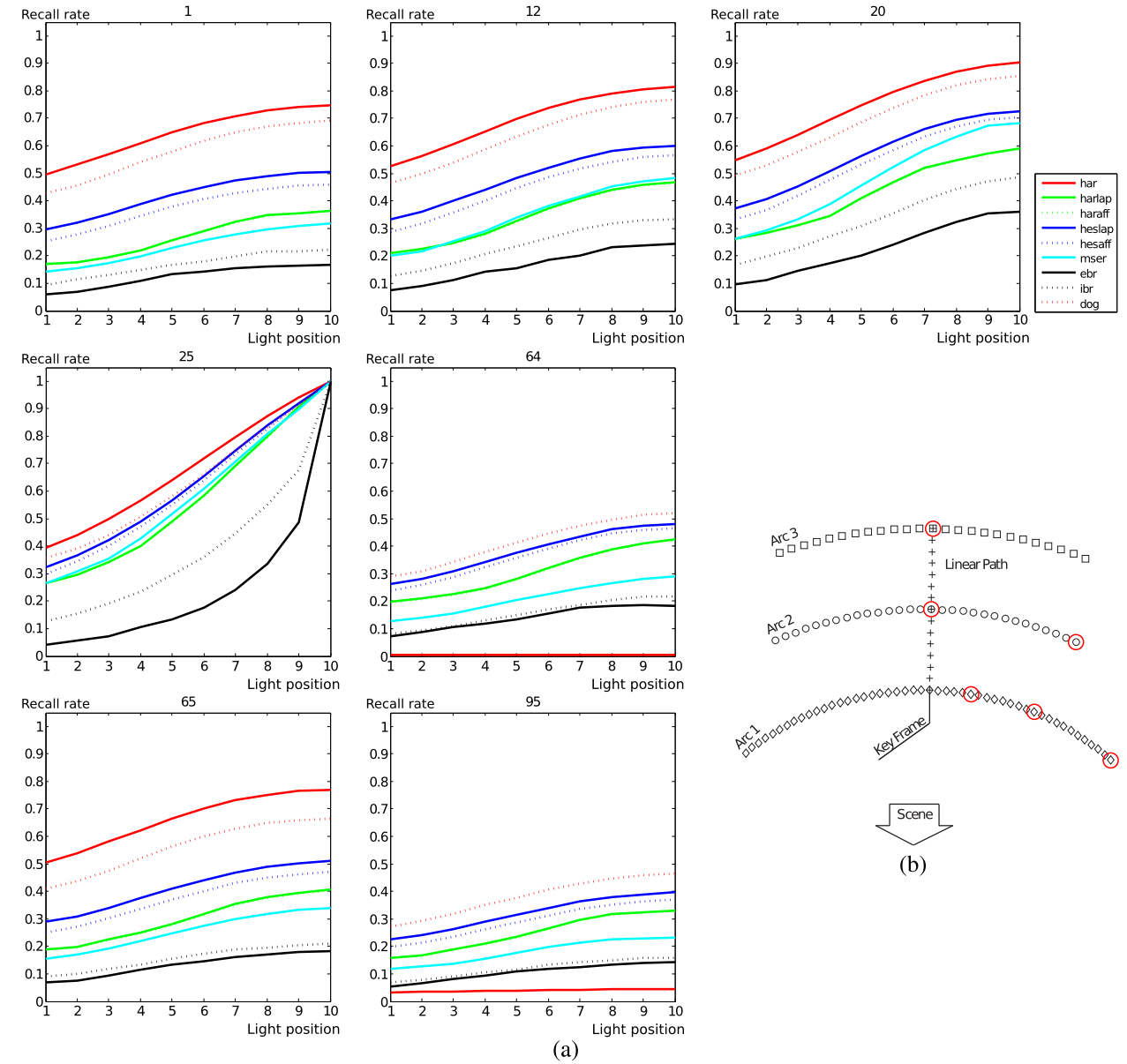
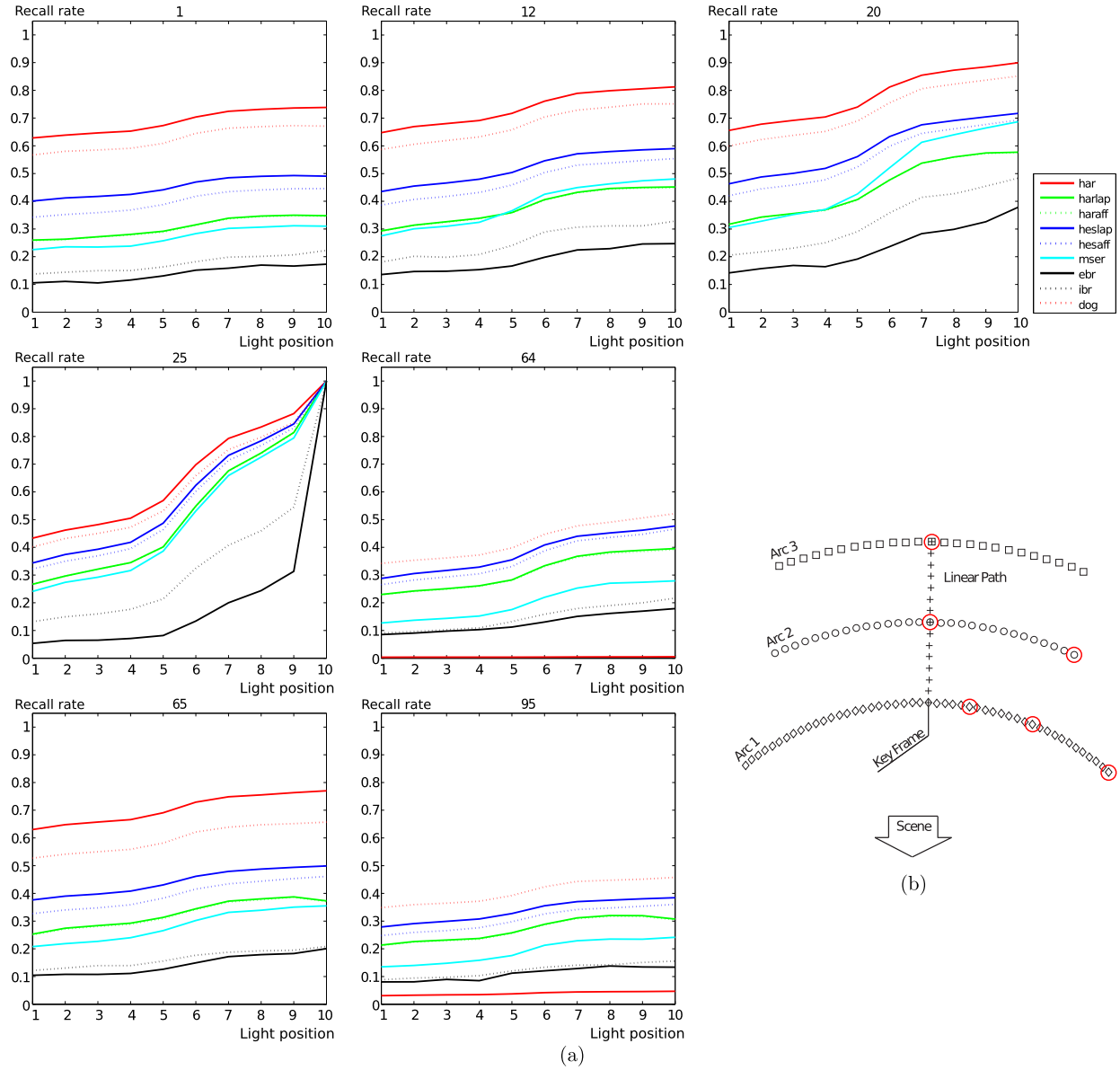
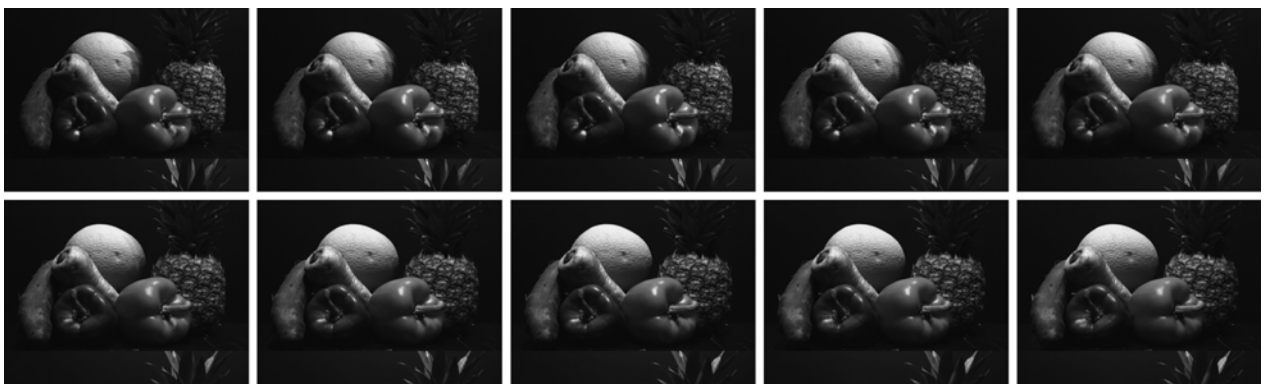


Fig. 11 Mean recall rate relative to change in light direction from back to front for seven camera positions averaged over all 60 scenes. The graphs (**a**) show the performance of the different detector types, with the average recall rate at the *vertical axis* and the light direction at the *horizontal axis*. The camera positions are shown in (**b**). An example of images from position 25 is shown in (**c**). The light changes gradually from back to front, with the *first row* being images 1–5 and *bottom row* images 6–10





(c)

Fig. 12 Mean recall rate relative to change in light direction from *right* to *left* for seven camera positions averaged over all 60 scenes. The graphs (**a**) show the performance of the different detector types, with the average recall rate at the *vertical axis* and the light direction at the *horizontal axis*. The camera positions are shown in (**b**). An example of images from position 25 is shown in (**c**). The light changes gradually from *right* to *left* with the *first row* being images 1–5 and *bottom row* images 6–10

Table 3 The average correlation of the recall rate by changing the threshold parameter from $0.107 - 9.31 \times$ the recommended parameter settings

Detector	Recall rate correlation
Harris	−0.0198
Harris Laplace	−0.0286
Harris Affine	−0.0275
Hessian Laplace	−0.2149
Hessian Affine	−0.1656

Sivic and Zisserman 2006). Our results show that the investigated interest point detectors are far from invariant to light changes, and this indicates that future research should focus on how to handle light variation to obtain more robust computer vision methods. The performance drop is relatively smaller when the scene is seen from the side. This might indicate that some feature are both robust to light and position variation, but the light variation is also smaller when the scene is viewed from the side.

Changing Model Parameters In the above experiments the recommended parameter settings were used. These correspond to standard settings of the downloaded software. To investigate the effect of these settings, we conducted the experiments with changing camera position using different cornerness and blob setting. This experiment is only conducted for the Harris and Hessian type detectors, because they are only governed by one threshold parameter. The parameter was varied on a logarithmic scale from $0.107 - 9.31 \times$ the recommended parameter settings. This is done in 21 steps by a multiplicative factor of 1.25.

From these experiments we observe that the recall rate of the Harris type detectors are unaffected by a change in the cornerness parameters and that the Hessian type detectors are only moderately affected. This happens despite these parameters drastically affect the number of interest points extracted. Our observations are quantisized in Table 3, which shows the correlation between the recall rate and the parameter setting. This implies that our results are relatively insensitive to the choice of parameter setting.

Complementarity of Interest Points Different interest points can complement each other by covering different parts of a scene. When two types of interest points complement each other it can be an advantage to apply both, which for example is used in Furukawa and Ponce (2007). We have made an investigation of how the ten interest points in this study complement each other by measuring their combined coverage of the scene. Our measure of complementarity is based on the surface reconstruction from the structured light scan and the 3D positions of the interest points. The 3D positions are found by projecting the interest point to the surface scan. We limit the complementarity measure to the key frame, see Fig. 4.

The complementarity of two sets of interest points, X and Y, is measured by computing the distance from each point in the structured light scan, S, to the nearest point in set X, set Y, and the union of X and Y. The average of these three distance distributions are then calculated by

$$D_x = \frac{1}{n}\sum_{i=1}^{n} \min_j \|X_j, S_i\|_2,$$

$$D_y = \frac{1}{n}\sum_{i=1}^{n} \min_k \|Y_k, S_i\|_2,$$

$$D_{xy} = \frac{1}{n}\sum_{i=1}^{n} \min\left(\min_j \|X_j, S_i\|_2, \min_k \|Y_k, S_i\|_2\right), \quad (1)$$

where n is the number of points in the structured light scan and S_i, X_j and Y_k denote individual points in the three point sets. We choose the following complementarity measure

$$\text{comp}(X, Y) = \frac{2\frac{D_{xy}}{\sqrt{n_x+n_y}}}{\frac{D_x}{\sqrt{n_x}} + \frac{D_y}{\sqrt{n_y}}}, \quad (2)$$

where the mean distances are divided by the square root of the number of interest points, n_x and n_y. This is done to adjust for the varying number of interest points from each detector. We chose the square root because it is proportional to the distance between nodes in a 2D grid with n_x points. Despite the fact that we are on a 2D manifold in a 3D space, we found it to be a good approximation. The average result of comparing the ten interest point detectors over the 60 scenes is shown in Fig. 13.

The motivation behind (2) is that we want a combination of interest points that represents a scene as well as possible. This is here represented as the distance from the 3D points obtained from the surface scans to the 3D positions of the interest points. If two sets of interest points complement each other well, the average distance from the structured light scan to the combined set of interest points should be reduced significantly by combining the two sets of interest points, which in essence is what (2) measures. The main result from this study is that the MSER, EBR and IBR detectors provide similar interest points but complement all other interest point detectors well, see Fig. 13.

6 Discussion

Our data set has enabled us to investigate interest point correspondence independently of descriptors for very complex, non-planar scenes. The key element that we investigate is,

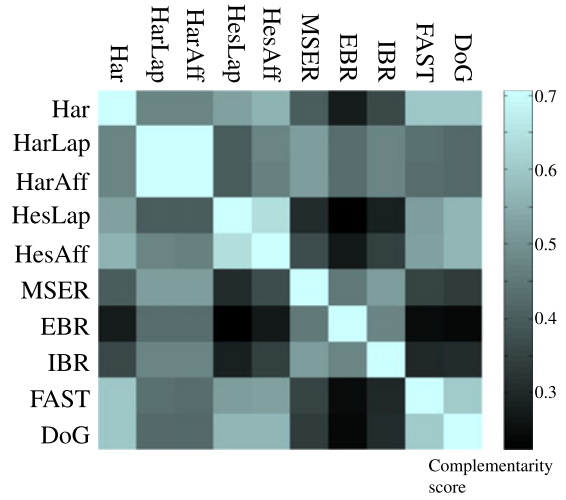


Fig. 13 The complementarity of the interest point detectors. The score is computed as described by (2) and averaged over all 60 scenes. Dark values imply that the two sets of interest points cover different parts of the scene and this way complement each other

if there for a given interest point is a potential matching interest point in a corresponding image. Our investigation is based on the same implementations as in the extensive study of interest points by Mikolajczyk et al. (2005). The novelty of our investigation is the complexity of the data set, both with regard to number of scenes, possibility of scene relighting and ground truth of geometric surface structure, which has led to nontrivial conclusions about the performance of interest point detectors.

The first investigation is concerned with the number of interest points provided by the algorithms. Most of the algorithms provide a reasonable number of interest points that varies with the depicted scene. This is expected because the number of interest points should be proportional to the number of features in the image. The FAST corner detector is highly unstable in terms of the number of interest points when we compare to the other interest point detectors. The detected number of interest points range from close to 0 to around 10,000, where most other interest point detectors are more consistent in the amount of interest points. The behavior of the FAST detector is very undesirable, because it makes this method unreliable for solving the correspondence problem. The correspondence estimate will be uncertain, especially in scenes with almost no interest points. A very large number of points can give higher certainty in solving the correspondence problem, but will slow down the matching. The EBR and IBR algorithms shows a relatively unstable behavior, but with few interest points. The best performance is achieved with the Harris and Hessian corner and blob detectors. MSER is also reasonably stable, but with few interest points. The DoG has very good performance, but with a large number of interest points.

Secondly, we have investigated the recall rate relative to the camera position, which provides very interesting results. We expect the recall rate to decrease when we change the camera position by increasing the angle or the distance to the scene. This is also what we observe for most interest point detectors. But the FAST corner detector does not show a decrease in performance with an increase in angle at the two distant arcs, see Fig. 9(b) and (c). This behavior is different than the other detectors. This is probably due to the high variation in number of interest points detected by FAST. The Harris corner detector performs very well for small-scale changes, but has a large drop in performance when scale exceeds a threshold. The reason is that this detector does not adapt to scale change and the threshold is a consequence of our scale matching criterion, as mentioned in Sect. 5. Interest points based on the FAST detector do not include scale, so the reported performance is not directly comparative to the other detectors. We have chosen to include this descriptor to illustrate its unreliable performance despite its advantage of not fulfilling the scale matching criteria. Overall the Harris corner detector and the DoG blob detector perform slightly better than the Hessian blob detector. This group of detectors is based on scale space features. Their performance is superior to MSER, but this detector does however perform reasonably well. IBR and EBR show poor performance. In general, our results do not provide a clear answer to which type of image structure (blobs or corners) is most optimal. In order to answer this question, we need to ensure that the detectors for the different type of image structure is as close in implementation details (e.g. choice of parameter settings and scale selection methods) as possible in order to provide comparable results. However, we were not able to achieve this with the current obtained implementations.

We have made an extensive scene lighting experiment in which we change the light direction. The recall rate is drastically affected by changing light, and the drop in performance is similar to the performance drop seen while changing camera position. The reflected light changes with incoming light direction resulting in a relatively large appearance change of the images. This effect is especially pronounced in specular surfaces and surfaces with local geometric variation, and less pronounced in diffuse and smooth surfaces. Looking at the images, the effect of relighting appears moderate, and much less than what is seen in an outdoor scene during the day. Therefore, the drastic reduction in performance comes as a surprise, and clearly shows that you should not expect too much of this group of methods when applied under conditions with large light variations. The ranking of the performance is similar to the experiment with diffuse conditions, showing that scale space corner and blob detectors and their approximations (Harris, Hessian and DoG) outperform the other methods. Especially EBR and IBR perform poorly.

We have investigated the recall rate in relation to changing parameters in the methods in order to see if some parameter settings are more favorable than others. Only the five Harris and Hessian interest point detectors listed in Table 3 were chosen for this investigation because they have one parameter that can be changed in a comparable way. The parameter is related to the strength of the interest points, and increasing the parameter allows lower contrast features to be included as interest points. We found the recall rate to be almost independent of the parameter settings—especially for the Harris corners, see Table 3. The Hessian blob detector showed a small decrease in recall rate when decreasing the feature strength of the interest point. The choice of parameters was also investigated in Mikolajczyk et al. (2005) where they found a stronger relation between the parameter settings and their repeatability measure, which is similar to our recall rate. Their investigation showed that choosing strong interest points would favor the repeatability, and similarly the clutter in a large number of interest points would give a high repeatability. Our investigation contradicts their observation within the broad span of parameters from 0.1–9.3 $\times$ the recommended parameter setting. The most important effect of changing parameters is the change in the number of detected interest points.

The complementarity study shows that different descriptor types cover different parts of the scene. Especially MSER, EBR and IBR detectors provide similar interest points but complement all other interest point detectors well (see Fig. 13). Since MSER outperforms the two other detectors in the other evaluations, MSER looks like the best choice of a complementary detector to the high performing Harris and Hessian detectors. It is also noticed that the Laplace and affine versions of the Harris and Hessian detectors are very alike, which is expected because the methods are almost identical. It is a little surprising that the basic Harris corner detector is not as similar to its Laplacian and affine counterparts as we would expect. An explanation might be that features detected at higher scales do not exist at the scale where the basic Harris corner detector operates. The spatial localization of the high scale interest points might also have changed due to the movement of feature points in scale space.

Overall, the simple Harris corner detector performs very well, but is not invariant to scale change. The Harris corners are closely followed by the DoG detector and outperformed by DoG when considering large-scale changes. Similar to the study in Mikolajczyk et al. (2005) we also observe an overall good performance of the variations of Harris and Hessian detectors. The difference in affine and non-affine is small, which is also expected when only looking at the interest points. The only difference is the local affine adaption, which can cause a scale variation, but the other two matching criteria are the same. We have not seen as good a performance of the MSER detector as reported in that study and by Fraundorfer and Bischof (2004). The non-planer and generally more challenging scenes in our study might cause this. MSERs performance problem on non-planar scenes, have previously been reported by Fraundorfer and Bischof (2005), but only based on one scene. Our results make it clear that this holds for complex scenes in general.

Viewed from a pure interest point detection perspective, detectors based on scale space features perform better than the other detector types, and especially better than the EBR, IBR and FAST. It is important to note, that this study only concerns interest points, which is just one element of solving the correspondence problem. The success of a system will depend on the interest point descriptor and the matching procedure as well. But the insights brought by this study show a clear performance difference and indicate what the effect of the interest point detector will be in a final system.

In some aspects, the conclusions of our study contradicts previous performance studies, e.g. for viewpoint change in Mikolajczyk et al. (2005), and underline the need for large data sets to firmly conclude on the performance, when evaluating new methods experimentally. The loss in performance is relatively large under light and viewpoint change, which should be considered when applying these methods. It is questionable how much gain there will be in suggesting new and improved interest point detectors, because image properties change when viewpoint and light change—in some scenes more than others. As a consequence perfect invariance cannot be obtained, and the accounted problems should be dealt with using other means.

7 Conclusion

The contribution of this paper is an investigation of ten established interest point detectors, which provide new insight to the stability of these detectors with respect to large changes in viewpoint, scale, and lighting. The investigation is based on a data set of 60 scenes with precise ground truth of camera position and scene surface, acquired with an industrial robot arm. Furthermore, a controlled light setting has enabled us to perform precisely controlled relighting experiments. Our conclusions are based on pure geometric constraints, and do not consider the discriminative properties of the underlying image structure. Based on this we conclude that interest points based on scale space features have the highest performance; these are the Harris corner detectors, the Hessian blobs and the difference of Gaussian (DoG), which is an approximation of the scale space Laplace operator. Especially for small-scale changes the simple Harris detector performs very well, and for scale adaptation the DoG detector is good. Maximally Stable Extremal Regions (MSER) did not show as good a performance as previously

reported, but especially the EBR and IBR are very poor in performance. Also the FAST corner was somewhat unreliable in performance.

In this study we have observed a relatively large decline in performance with change in viewpoint and lighting. This is important to account for when interest points are used for methods in natural scenes with large variation in lighting.

Acknowledgements We would like to thank the Oxford Vision Group[2] and David Lowe[3] for making their code available online. Furthermore, this work was partly financed by the Centre for Imaging Food Quality project, which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Programme Commission on Health, Food and Welfare.

References

Aanæs, H., Dahl, A. L., & Perfernov, V. (2009). *Technical report on two view ground truth image data* (Tech. rep.) DTU Informatics, Technical University of Denmark.

Aanæs, H., Dahl, A. L., & Pedersen, K. S. (2010). On recall rate of interest point detectors. In *Proceedings of 3DPVT*. http://campwww.informatik.tu-muenchen.de/3DPVT2010/data/media/e-proceeding/session07.html#paper97.

Alvarez, L., Gousseau, Y., & Morel, J. M. (1999). The size of objects in natural and artificial images. In P.W. Hawkes (Ed.), *Advances in imaging and electron physics*. New York: Academic Press.

Brown, M., Hua, G., & Winder, S. (2011). Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 43–57.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8*(6), 676–698.

Crowley, J. L., & Parker, A. C. (1984). A representation for shape based on peaks and ridges in the difference of low-pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(2), 156–170.

Demirci, A. F., Platel, B., Shokoufandeh, A., Florack, L. M. J., & Dickinson, S. J. (2009). The representation and matching of images using top points. *Journal of Mathematical Imaging and Vision*, *35*(2), 103–116.

Einarsson, P., Chabert, C., Jones, A., Ma, W., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., & Debevec, P. (2006). Relighting human locomotion with flowed reflectance fields. In *Rendering techniques* (pp. 183–194).

Forstner, W. (1986). A feature based correspondence algorithms for image matching. *International Archives of Photogrammetry and Remote Sensing*, *24*, 60–166.

Fraundorfer, F., & Bischof, H. (2004). Evaluation of local detectors on non-planar scenes. In *Proc. 28th workshop of AAPR* (pp. 125–132).

Fraundorfer, F., & Bischof, H. (2005). A novel performance evaluation method of local detectors on non-planar scenes. In *Proceedings of computer vision and pattern recognition—CVPR workshops* (pp. 33–43).

Furukawa, Y., & Ponce, J. (2007). Accurate, dense, and robust multiview stereopsis. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Griffin, L. D., Lillholm, M., Crosier, M., & van Sande, J. (2009). Basic image features (bifs) arising from approximate symmetry type. In *LNCS: Vol. 5567. Scale space and variational methods in computer vision* (pp. 343–355).

Gustavsson, D. (2009). *On texture and geometry in image analysis*. Ph.D. thesis, Department of Computer Science, University of Copenhagen, Denmark.

Haeberli, P. (1992). Synthetic lighting for photography. Grafica obscura. http://www.graficaobscura.com/synth/index.html.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *4th Alvey vision conf.* (pp. 147–151).

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hua, G., Brown, M., & Winder, S. (2007). Discriminant embedding for local image descriptors. In *ICCV* (pp. 1–8).

Johansen, P., Skelboe, S., Grue, K., & Andersen, J. (1986). Representing signals by their toppoints in scale space. In *Proceedings of the international conference on image analysis and pattern recognition* (pp. 215–217). New York: IEEE Computer Society Press.

Johansen, P., Nielsen, M., & Olsen, O. F. (2000). Branch points in one-dimensional Gaussian scale space. *Journal of Mathematical Imaging and Vision*, *13*(3), 193–203.

Kadir, T., Zisserman, A., & Brady, M. (2004). An affine invariant salient region detector. In *Proceedings of European conference on computer vision (ECCV)* (pp. 228–241).

Konishi, S., Yuille, A., & Coughlan, J. (2003a). A statistical approach to multi-scale edge detection. *Image and Vision Computing 21*(1):37–48.

Konishi, S., Yuille, A. L., Coughlan, J. M., & Zhu, S. C. (2003b). Statistical edge detection: Learning and evaluating edge cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(1), 57–74.

Laptev, I., & Lindeberg, T. (2003). A distance measure and a feature likelihood map concept for scale-invariant model matching. *International Journal of Computer Vision*, *52*(2/3), 97–120.

Lillholm, M., & Griffin, L. (2008). Novel image feature alphabets for object recognition. In *Proceedings of ICPR'08*.

Lillholm, M., & Pedersen, K. S. (2004). Jet based feature classification. In *Proceedings of international conference on pattern recognition*.

Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention. *International Journal of Computer Vision*, *11*, 283–318.

Lindeberg, T. (1998a). Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, *30*(2), 117–154.

Lindeberg, T. (1998b). Feature detection with automatic scale selection. *International Journal of Computer Vision*, *30*(2), 79–116.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proc. of 7th ICCV* (pp. 1150–1157).

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, *22*(10), 761–767.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630.

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

[2] http://www.robots.ox.ac.uk/~vgg/research/affine/index.html.

[3] http://www.cs.ubc.ca/~lowe/keypoints/.

Moreels, P., & Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, *73*(3), 263–284.

Mumford, D., & Gidas, B. (2001). Stochastic models for generic images. *Quarterly of Applied Mathematics*, *59*(1), 85–111.

Nielsen, M., & Lillholm, M. (2001). What do features tell about images. In M. Kerckhove (Ed.), *LNCS: Vol. 2106. Proc. of Scale-Space'01* (pp. 39–50). Vancouver: Springer.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *CVPR* (Vol. 5).

Ren, X., & Malik, J. (2002). A probabilistic multi-scale model for contour completion based on image statistics. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *LNCS: Vol. 2350–2353. Proc. of ECCV'02* (pp. 312–327). Copenhagen: Springer. Vol. I.

Ren, X., Fowlkes, C. C., Malik, J. (2008). Learning probabilistic models for contour completion in natural images. *International Journal of Computer Vision*, *77*(1–3), 47–63.

Salvi, J., Pages, J., & Batlle, J. (2004). Pattern codification strategies in structured light systems. *Pattern Recognition*, *37*(4), 827–849.

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *Proceedings of CVPR* (Vol. 1, pp. 195–202).

Schmid, C., & Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(5), 530–535.

Schmid, C., Mohr, R., & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, *37*(4), 151–172.

Sivic, J., & Zisserman, A. (2006). Video Google: Efficient visual search of videos. *Lecture Notes in Computer Science*, *4170*, 127.

Sivic, J., Russell, B., Efros, A., Zisserman, A., & Freeman, W. (2005). Discovering objects and their location in images. In *ICCV, 2005. Tenth IEEE international conference on computer vision* (pp. 370–377).

Snavely, N., Seitz, S., & Szeliski, R. (2008a). Modeling the world from Internet photo collections. *International Journal of Computer Vision*, *80*(2), 189–210.

Snavely, N., Seitz, S. M., & Szeliski, R. (2008b). Modeling the world from Internet photo collections. *International Journal of Computer Vision 80*(2), 189–210. http://phototour.cs.washington.edu/.

Srivastava, A., Lee, A. B., Simoncelli, E. P., & Zhu, S. C. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, *18*(1), 17–33.

Torr, P., & Zisserman, A. (1999). Feature based methods for structure and motion estimation. In *Lecture notes in computer science* (pp. 278–294).

Trajković, M., & Hedley, M. (1998). Fast corner detection. *Image and Vision Computing*, *16*(2), 75–87.

Tuytelaars, T., & Van Gool, L. (2004). Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, *59*(1), 61–85.

Winder, S. A. J., & Brown, M. (2007). Learning local image descriptors. In *Proceedings of CVPR* (pp. 1–8).

Winder, S., Hua, G., & Brown, M. (2009). Picking the best daisy. In *Proceedings of CVPR* (pp. 178–185).