International Journal of Computer Vision

Interest Point Detectors and Descriptors - What is the Best Interest Point Detector and Descriptor Combination?

--Manuscript Draft--

Manuscript Number:	
Full Title:	Interest Point Detectors and Descriptors - What is the Best Interest Point Detector and Descriptor Combination?
Article Type:	Manuscript
Keywords:	Interest point detectors, Interest point descriptors, Feature matching, Performance evaluation
Corresponding Author:	Anders Lindbjerg Dahl, Ph.d. DTU Informatics, Technical University of Denmark Lyngby, DENMARK
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	DTU Informatics, Technical University of Denmark
Corresponding Author's Secondary Institution:	
First Author:	Anders Lindbjerg Dahl, Ph.d.
First Author Secondary Information:	
Order of Authors:	Anders Lindbjerg Dahl, Ph.d.
	Henrik Aanæs, PhD
	Kim Steenstrup Pedersen, PhD
Order of Authors Secondary Information:	
Abstract:	The matching of interest points between image pairs is a fundamental part of many central computer vision applications. The performance of the plethora of interest point detector and descriptor methods are therefore of natural interest within the field. We present such a performance evaluation, considering both interest point detectors (the extraction of points) and descriptors (the basis of the similarity metric), as well as their combined effect. This is done by evaluating the exhaustive combination of a large number of state of the art detectors and descriptors. The investigation is based on our extensive DTU robot dataset, with ground truth, which allows us not only to identify the best performing methods, but also quantify the strength of these findings via their statistical significance. This data set allow us to evaluate based on varying viewing angle, feature scale and changing lighting conditions over sixty different scenes. The evaluated descriptors are all, except for correlation, various derivatives of the SIFT (Scale Invariant Image Transform) descriptor. In contrast to previous studies, the performance difference is shown to be very marginal, although the opponent SIFT descriptor perform best. The best performing detector is a novel formulation of the multi-scale Harris corner detector, which is given here. Very few cross effects between detector and descriptor performance was observed. Lastly, an analysis giving insights into the workings of the interest point matching based on this thorough evaluation is presented.

International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Interest Point Detectors and Descriptors

What is the Best Interest Point Detector and Descriptor Combination?

Anders Lindbjerg Dahl · Henrik Aanæs · Kim Steenstrup Pedersen

Received: date / Accepted: date

Abstract The matching of interest points between image pairs is a fundamental part of many central computer vision applications. The performance of the plethora of interest point detector and descriptor methods are therefore of natural interest within the field. We present such a performance evaluation, considering both interest point detectors (the extraction of points) and descriptors (the basis of the similarity metric), as well as their combined effect. This is done by evaluating the exhaustive combination of a large number of state of the art detectors and descriptors. The investigation is based on our extensive DTU robot dataset, with ground truth, which allows us not only to identify the best performing methods, but also quantify the strength of these findings via their statistical significance. This data set allow us to evaluate based on varying viewing angle, feature scale and changing lighting conditions over sixty different scenes. The evaluated descriptors are all, except for correlation, various derivatives of the SIFT (Scale Invariant Image Transform) descriptor. In contrast to previous studies, the performance difference is shown to be very marginal, although the opponent SIFT descriptor perform best. The best performing detector is a novel formulation of the multi-scale Harris corner detector, which is given here. Very few cross effects between detector and descriptor performance was observed. Lastly, an analysis giving insights into the workings of the interest point matching based on this thorough evaluation is presented.

Anders Lindbjerg Dahl, Henrik Aanæs
DTU Informatics
Technical University of Denmark
Denmark
E-mail: abd@imm.dtu.dk, haa@imm.dtu.dk

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

1 Introduction

Many successful computer vision applications (Agarwal et al, 2009; Brown and Lowe, 2005; Crandall et al, 2009; Snavely et al, 2006; Winder and Brown, 2007) that require image correspondence are based on local image features. Representing an image by a sparse set of salient features is computational attractive, which has greatly contributed to their success. Typically, features are found as a set of salient interest points combined with a local differential geometric description (Tuytelaars and Mikolajczyk, 2008). Various detectors and descriptors have been proposed, but the question of how to optimally design an interest point characterization still remains open.

In the image correspondence problem, feature quality can be assessed by matching features between two images and counting the number of correct and wrong matches. The problem is how to determine if a feature match is correct, i.e. to validate if correspondence exist. However, with knowledge about the geometry of the observed scene and of the camera it becomes easy to verify if two interest points, that are corresponding in feature space, also correspond in the real scene. To overcome this problem we propose to use the DTU Robot dataset[1] with known surface geometry presented by Aanæs et al (2012) (see Fig. 1 and Sec. 2 for a brief description). Based on this dataset we are able to systematically analyze the performance of feature meth-

[1] Data available from `http://www.imm.dtu.dk/robotdata`

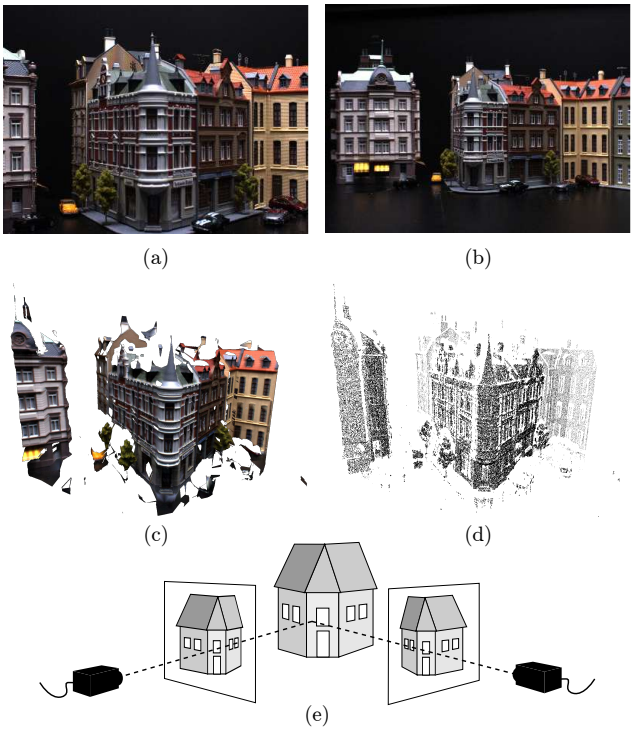


Fig. 1 Example of data and setup. (a,b) two images of the same scene – one close up (a) and one distant from the side (b). (c) The reconstructed 3D surface and (d) the surface points. (e) Corresponding interest points can be found using the geometric information of the scene with known camera positions and 3D scene surface.

ods and due to the large variation in scene types we can judge the statistical significance of our findings.

Finding correspondence between image pairs using interest points is based on the assumption that common interest points will be detected in both images. The optimal match can be obtained if corresponding interest points are localized precisely on the same scene element, and the associated region around each interest point covers the same part of the scene.

Commonly, candidate points are detected using an interest point *detector* and a description of the local image structure – the so called *descriptors* – surrounding the interest points are extracted. Following the extraction of descriptors, a comparison of these is made using a similarity metric in order to determine correspondence between interest points. The rationale is that descriptors capture the essential visual appearance of the scene region covered by the interest point. Hereby the same scene point seen from different viewpoints and/or with different lighting or color of the light will get similar descriptors. We therefore expect the performance, measured in relation to image matching, to be dependent on the combined choice of algorithms for detecting interest points and the local description of the image around the found interest points. The question is what combination of interest point detector and descriptor is optimal – this is the question we address in this paper.

1.1 Related work

An early approach for characterizing an image using interest points coupled with a local descriptor was performed by Schmid and Mohr (1997). Image correspondence was established from Harris corner interest points (Harris and Stephens, 1988) combined with differential geometric invariants in scale-space (Koenderink, 1984; Lindeberg, 1994; ter Haar Romeny, 1994; Witkin, 1983). In Schmid et al (2000) interest points from planer scenes were evaluated. Later affine invariant interest points were introduced to obtain robustness to large viewpoint changes. These methods have been surveyed by Mikolajczyk et al (2005). In this work the performance was evaluated on a small datasets consisting of ten scenes each containing six images. This dataset and the suggested evaluation criteria have since been used in numerous works.

Interest points are typically found as local differential geometric features in the image. Image corners have been a popular choice both with the Harris corner detector (Harris and Stephens, 1988) and its variations like Harris Laplace and Harris Affine (Mikolajczyk and Schmid, 2004). Other types of corner detectors include the multi-scale corners of Lindeberg (1998) based on finding points of maximal isophote curvature in scale-space. Corners have the advantage of being localization stable and perform well under varying photometric parameters. Interestingly, Loog and Lauze (2010) have shown that Harris corner points can be derived from an established measure of saliency as improbable local image structure, hence being highly informative and therefore are good candidates for interest points. Intensity blobs are another feature type that has been a popular choice for interest point detection. Blob detectors include Hessian Laplace and Hessian Affine (Mikolajczyk et al, 2005), and the multi-scale DoG (Difference of Gaussians) (Lowe, 2004). Blobs are also localization stable and has provided good performance. MSER (Maximally Stable Extremal Regions) (Matas et al, 2004) have similarities to blobs by being regions of local extreme intensity. MSER are also localization stable and additionally encode the local shape making an affine adaption easy.

Different approaches have been taken when describing the local visual appearance of interest points. A majority of approaches extract some descriptive feature, such as histograms of differential geometric image properties in each pixel (Lowe, 2004; Mikolajczyk

and Schmid, 2005; Schmid and Mohr, 1997; Tola et al, 2008), using integral images (Bay et al, 2006, 2008), or the responses of steerable filters (Freeman and Adelson, 1991), differential invariants or local jets (Balmashnova and Florack, 2008; Florack et al, 1993; Koenderink and van Doorn, 1987; Schmid et al, 2000). The SIFT (Lowe, 2004), GLOH (Mikolajczyk and Schmid, 2005), and DAISY (Tola et al, 2008, 2009; Winder et al, 2009; Winder and Brown, 2007) descriptors also includes a spatial pooling step in order to agglomerate the descriptive feature in an arrangement around the interest point. A selection of descriptors have previously been evaluated by Mikolajczyk and Schmid (2005) on the same dataset as used by Mikolajczyk et al (2005). Again the limitations of the dataset restrict the ability to generalize the results from that survey to a wider class of scene types and more natural variation in illumination.

Color invariant descriptors have been investigated by Abdel-Hakim and Farag (2006); Burghouts and Geusebroek (2009); Van De Sande et al (2010) based on the theory of color invariance (Geusebroek et al, 2001). The invariance is obtained from derivatives in the Gaussian opponent color space. Color invariant descriptors were evaluated both on objects with varying illumination and color (Geusebroek et al, 2005) and outdoor scenes (Everingham et al, 2006, 2007; Snoek et al, 2006). Color variation occurs as a result of changes in illumination both caused by shadows (changes in light source position), inter-surface reflectance, as well as changes to the color of the incident light. The conclusion of these studies are that the opponent SIFT descriptor, based on the opponent color space, has highest performance closely followed by the CSIFT descriptor, which is based on derivatives in the Gaussian opponent color space.

The ground truth in the data from Mikolajczyk et al (2005) was obtained by a manually annotated image homography. This limits the scene geometry to planar surfaces or scenes viewed from a large distance where a homography is a good approximation. Fraundorfer and Bischof (2004) addressed this limitation by generating ground truth and requiring that a matched feature should be consistent with the known camera geometry across three views. In Winder *et al.* (Brown et al, 2011; Hua et al, 2007; Winder et al, 2009; Winder and Brown, 2007) results from Photo Tourism (Snavely et al, 2006, 2008) were used as ground truth, which was based on SIFT matching.

Moreels and Perona (2007) evaluated feature descriptors, similar to Fraundorfer and Bischof (2004), based on pure geometry by requiring three view geometric consistency with the epipolar geometry. In addition, they used a depth constraint based on knowledge about the position of their experimental setup. Hereby they obtained unique correspondence between 500-1000 detected points from each object. The limitation of their experiment lies in the use of relatively simple scenes with mostly single objects resulting in little self-occlusion. However, self-occlusions are very frequent in real world scenes and many interest points are typically found near occluding boundaries. It is also interesting to investigate how this affects the performance of interest point detectors in combination with a descriptor, because the appearance of the neighborhood around interest points on occluding boundaries can change dramatically with viewpoint change.

We have previously (Aanæs et al, 2012) made a comparative study of the performance of interest point detectors on the DTU Robot dataset. This data set contains a wide variety of 60 scenes captured under controlled lighting and with a broad range of changes to view angle including large scale changes. The data set also includes 3D surface scans of each scene providing ground truth for matching points in image pairs. The study included 10 state of the art detectors and revealed that scale adapted detectors obtain the overall best performance on all scene types. We believe that the DTU Robot data set provides a more realistic scenario for evaluating interest point detectors and descriptors than the previously used datasets.

This investigation The aim of this work is to compare pairs of feature detectors and descriptors, to find the best combination. To keep the computational burden manageable, the number of candidates has to be limited, hereby reducing the combinatorial explosion. We only use candidates that have previously been reported to perform well. We choose the following interest point detectors; Harris, Harris Affine, Harris Laplace, Hessian Laplace, Hessian Affine, MSER, and Difference of Gaussian (DoG), because they are popular and reported to work well in the literature (Aanæs et al, 2012; Aanæs et al, 2010; Tuytelaars and Mikolajczyk, 2008). These are all based on the original authors implementation. In addition, we have tested scale-space corners (Lindeberg, 1998) and our implementation of the multi-scale Harris corners – a multi-scale extension of the improved Harris corner detector (Schmid et al, 2000).

SIFT (Lowe, 2004) is a general well performing descriptor and the typical choice in many applications based on interest point descriptors (Agarwal et al, 2009; Brown and Lowe, 2005; Crandall et al, 2009; Snavely et al, 2006; Winder and Brown, 2007). Many of the descriptors tested in this study are based on variations of the SIFT descriptor. The DAISY descriptor (Tola et al, 2008, 2009; Winder et al, 2009; Winder and Brown, 2007) is another state of the art descriptor that has

been shown to provide very good results. The DAISY descriptor is build as a vector of spatially histogram binned first order image derivative, and is in this way similar to SIFT. We have chosen to implement the DAISY descriptor based on the framework of Winder and Brown (2007) with parameters based on the results from Dahl et al (2011). The DAISY descriptor is purely based on gray scale images.

We use the DTU Robot dataset (Aanæs et al, 2012) in this study, which allows us to investigate the robustness of descriptors with respect to changes in light position including shadow effects and inter-surface reflectance. However, the color of incident light is fixed. Nevertheless, this allows us to investigate the robustness of each descriptor towards light variations to a higher degree than previous studies. We therefore also include the following color invariant descriptors in our investigation; CSIFT (Abdel-Hakim and Farag, 2006; Burghouts and Geusebroek, 2009) and opponent SIFT (Van De Sande et al, 2010). We also include some simple descriptors such as raw intensities and simple extensions of the SIFT descriptor to color. More details will be given in Section 3.2.

Our evaluation is based on computing all combinations of 14 descriptor variants and 11 detector variants on the DTU Robot Dataset. Each combination is evaluated using the area under the Receiver Operating Characteristics (ROC) curves, also referred to as the AUC measure. Hereby we obtain results for the combined effect of descriptors and detectors in relation to variation in camera position, light direction, and scene type. This is in line with the strategy followed by Winder et al (2009) in their evaluation of interest point descriptors.

1.2 Contribution

The contributions of this paper are:

1. Evaluation of state-of-the-art detector-descriptor combinations including color descriptors on the DTU Robot dataset.
2. Novel experimentally based conclusions and insights in the performance of detector-descriptor combinations for feature matching, including that the best performing detector performs best for all descriptors, and that there is little performance difference between the various descriptors derived from SIFT.
3. We provide implementation details of the superior performing multi-scale Harris corner detector – a detector which has previously been discarded by Mikolajczyk and Schmid (2004).

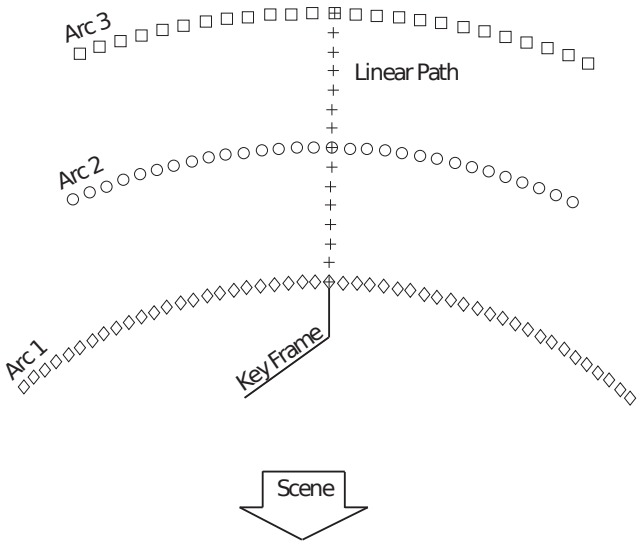


Fig. 2 The central frame in the nearest arc is the key frame, and the surface reconstruction is attempted to cover most of this frame. The three arcs are located on circular paths with radii of 0.5 m, 0.65 m and 0.8 m, which also defines the range of the linear path. Furthermore, Arc1 spans $+/-40°$, Arc2 $+/-25°$ and Arc3 $+/-20°$. Illustration from Aanæs et al (2010).

4. Demonstration that the use of a large set of systematically collected data, like the DTU Robot dataset, provide results with high statistical confidence.

This paper is an extension of our previous conference paper (Dahl et al, 2011) containing an evaluation study of combined interest point detectors and local image descriptors together with an evaluation of the popular *DAISY* descriptor (Brown et al, 2011; Tola et al, 2009). In this paper, we have additionally investigated the influence of light variation and the use of color image descriptors, and significantly extended the number of detectors and descriptors. Furthermore, we have added an extensive statistical analysis providing strong evidence for our findings.

2 Data

In this investigation we use the DTU Robot dataset (Aanæs et al, 2012)[2] outlined in Fig. 1. The dataset has been constructed to enable point correspondence verification, i.e. if we have matching points in two images, we can verify if they originate from the same scene position. Furthermore, it is possible to perform image based scene relighting of the dataset, which we have used for our light variation experiment.

The dataset has been acquired using a six axis industrial robot arm which is shown in Fig. 4. Very pre-

[2] `http://roboimagedata.imm.dtu.dk/`

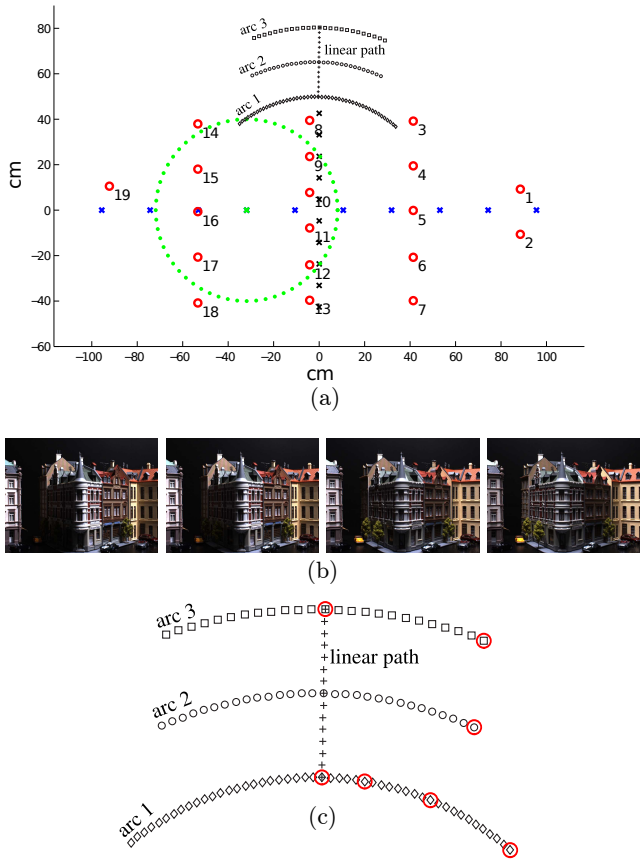


Fig. 3 Illustration of the relighting setup. (a) The layout of the light setup is illustrated with the red circles showing the positions of the white LEDs. The camera path is shown above the light layout. At each position an image is taken with one diode turned on at a time. The crosses show the relight sampling points from right to left (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. (b) Image examples of right to left relighting. (c) The spatial camera positions of light experiment.

cise camera positions have been obtained by mounting a camera on the robot arm. Based on this setup we have photographed 60 complex scenes from the 119 fixed camera positions shown in Fig. 2. The acquired images are 24 bit RGB images of size 1200×1600 pixels. The scenes have been illuminated by 19 point light sources, which we have used for scene relighting as illustrated in Fig. 3. Furthermore, the surface of each scene has been scanned using structured light. Together with the camera geometry this allows us to accurately determine the correct camera correspondences *without* matching visual features.

In real outdoor scenes, like the ones presented in (Winder et al, 2009), LIDAR (light detection and ranging) as employed in Strecha et al (2008) could be an alternative. They have scanned seven buildings for evaluating algorithms for structure from motion. This data has the advantage of being of real outdoor scenes, but the scale and light variation is limited as well as the material properties of the scenes. With the DTU Robot dataset we obtain a much larger variation in camera positioning, light variation, and scene types, which is the motivation for using it in our experiments.

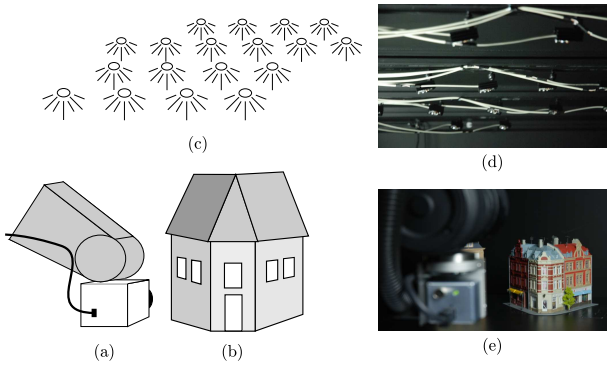


Fig. 4 Illustration of data collection setup. (a) The camera is mounted on a robot arm capturing images of the scene shown in (b). (c) LED point light sources illuminate the scene from 19 individual positions. (d,e) Photos of the real setup. Illustration from Aanæs et al (2010).

Table 1 Azimuth and elevation of the ten relighting positions from left to right and back to front respectively.

	Right to left		Back to front	
# Pos.	Azimuth	Elevation	Azimuth	Elevation
1	270.0°	55.5°	0.0°	73.0°
2	270.0°	61.9°	0.0°	76.6°
3	270.0°	69.1°	0.0°	80.3°
4	270.0°	77.1°	0.0°	84.2°
5	270.0°	85.6°	0.0°	88.0°
6	90.0°	85.6°	180.0°	88.0°
7	90.0°	77.1°	180.0°	84.2°
8	90.0°	69.1°	180.0°	80.3°
9	90.0°	61.9°	180.0°	76.6°
10	90.0°	55.5°	180.0°	73.0°

3 Evaluation

The detectors and descriptors chosen for our experiments are listed in Tab. 2 and 3. All combinations of chosen detectors and descriptors have been computed on the DTU Robot dataset, both with diffuse light setting and with light variation.

3.1 Interest point detectors

The interest point detectors included in this study are the improved Harris (Schmid et al, 2000), Harris Laplace, and Harris Affine corner detectors (Mikolajczyk and

Schmid, 2004), and the Hessian Laplace, Hessian Affine (Mikolajczyk et al, 2005), and difference of Gaussians (DoG) (Lowe, 2004) blob detectors, as well as the region based Maximally Stable Extremal Regions (MSER) (Matas et al, 2004) detector. All have previously been shown to work well with various combinations of interest point descriptors (Aanæs et al, 2010; Tuytelaars and Mikolajczyk, 2008; Mikolajczyk et al, 2005), hence one would expect similar good performance on the DTU Robot dataset. We use the implementations of these interest point detectors provided by the authors[3]. The detectors are listed in Tab. 2.

In addition, we have implemented a multi-scale corner detector by Lindeberg (1998) and a multi-scale version of the Harris corner detector with scale selection using the general scale selection principle proposed by Lindeberg (1998). We include these detectors because we noticed in the survey by Aanæs et al (2012) that some of the supposedly scale invariant detectors, such as Harris Laplace and Hessian Laplace and their affine variants, did not perform well under large scale changes. We suspect the cause is to be found in the scale selection approach utilized by these detectors and not as a result of artifacts in the data set. We investigate this by including the two new detector implementations in the study, both of which are scale invariant by construction. Below we give the implementation details of these detectors and the source code is available at the DTU Robot dataset website[4]. All the provided parameters are valid for 8 bit gray-scale images with intensities in the range 0 to 255.

Lindeberg corners The Lindeberg (1998) corner detector is based on finding points of locally maximal isophote curvature in the linear scale-space representation of the image (Koenderink, 1984; Witkin, 1983). This approach simultaneously detects the location of the corner point in space and estimates its scale (scale selection).

The scale-space of an image $I : \mathbb{R}^2 \to \mathbb{R}$ is given by the family of images parameterized by scale σ produced by Gaussian blurring, $L(\mathbf{x};\sigma) = (I * G)(\mathbf{x};\sigma)\ ,\ \sigma \geq 0$, with boundary condition $L(\mathbf{x};\sigma = 0) \equiv I(\mathbf{x})$. Here $*$ denotes convolution with respect to the space variable $\mathbf{x}$ and

$$G(\mathbf{x};\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right) \tag{1}$$

is the Gaussian function. The scale-space representation allows us to compute well-defined image derivatives by convolving the image with the corresponding Gaussian derivative,

$$L_{x^n y^m}(\mathbf{x};\sigma) = \left(I * \frac{\partial^{n+m} G}{\partial x^n \partial y^m}\right)(\mathbf{x};\sigma)\ . \tag{2}$$

Lindeberg (1998) propose a general principle for simultaneous localization and scale selection for image features. This principle is based on detection of local extrema in scale-space of scale-normalized measures of feature strength. Among several examples of use of this principle, Lindeberg defines a corner detector based on finding local maxima of the isophote curvature multiplied by the gradient magnitude raised to the power of three. Lindeberg's scale normalized corner strength can be expressed in terms of image derivatives as

$$\kappa(\mathbf{x};\sigma) = \sigma^{8(\gamma-1)}(L_y^2 L_{xx} - 2L_x L_y L_{xy} + L_x^2 L_{yy})^2\ , \tag{3}$$

where γ is a scaling parameter.

We implement the scale-space representation by performing convolution by products in the frequency domain. This is done by performing a Fourier transform of the image, multiplying with the Gaussian filter constructed in the frequency domain, followed by an inverse Fourier transform. We use an implementation of the fast Fourier transform algorithm. We have chosen not to apply a Gaussian pyramid representation, hence the image resolution stays the same for all scale levels in the scale-space. This may potentially provide better localization performance at the cost of longer computation time.

We have found that we achieve the best overall performance of this detector by choosing to sample the scale-space at 31 scale levels, such that the ith scale level is $\sigma_i = k^{(i-1)}\sigma_0$, $i = 1, \ldots, 31$, with $k = 1.1$ and $\sigma_0 = 1.5$ pixels. At each scale level we compute the feature strength (3) ending up with a discretized scale-space of κ. For the scaling parameter we choose $\gamma = 7/8$, which leads to a detection of corners at small focused scales (Lindeberg, 1998).

The detection of local maxima of the feature strength (3) is simply performed by searching through the discretized 3-dimensional scale-space representation of κ and comparing the value of κ at every scale-space pixel with its 26 neighbors both spatially and across scale. A pixel at $\mathbf{x}$ and σ corresponds to a local maxima, if its κ value is larger than or equal to all its neighbors, $\kappa(\mathbf{x};\sigma) \geq \kappa(\mathbf{x}_j;\sigma_j)$, where j indexes all the 26 neighbors. Passing this test of local maximality leads to the location being detected as a corner point. In order to remove noise in the detected points we apply a threshold of $10^{7.5}$ and discard all detected points with a κ value less than this threshold. We also remove points on the boundary of the image (points closer than 2σ

[3] `http://www.robots.ox.ac.uk/~vgg/research/affine/index.html` and `http://www.cs.ubc.ca/~lowe/keypoints/`

[4] `http://roboimagedata.imm.dtu.dk/`

to the boundary) in order to avoid spurious detections caused by boundary effects in the implementation of the scale-space representation.

In order to improve the spatial localization and refine the selected scale, Lindeberg (1998) proposes an iterative localization principle to be applied after initial detection of a local scale-space maxima. The principle is to find a gradient weighted point in the spatial neighborhood of the initially detected point $\mathbf{x}_0$ that has minimum orthogonal distance to the nearest edges forming the corner. This optimal point can be found in closed form by

$$\mathbf{x} = \hat{\mathbf{A}}^{-1}\mathbf{b}, \tag{4}$$

where

$$\hat{\mathbf{A}} = \int_{\mathbf{x}' \in \mathbb{R}^2} (\nabla L)(\nabla L)^T w_{\mathbf{x}_0} d\mathbf{x}' \tag{5}$$

$$\mathbf{b} = \int_{\mathbf{x}' \in \mathbb{R}^2} (\nabla L)(\nabla L)^T \mathbf{x}' w_{\mathbf{x}_0} d\mathbf{x}' \ , \tag{6}$$

with $\nabla L = (L_x, L_y)^T$ denoting the local image gradient. We choose the window function $w_{\mathbf{x}_0}(\mathbf{x}; \sigma_I) = G(\mathbf{x} - \mathbf{x}_0; \sigma_I)$ to be a Gaussian function with the same scale as the originally detected scale for the corner point $\mathbf{x}_0$. The method starts at the detected point $\mathbf{x}_0$ and then iteratively refines the location of this point by applying (4). At every iteration the latest position estimate $\mathbf{x}$ is substituted for $\mathbf{x}_0$ in (5) and (6).

In our implementation of this procedure we consider a point as converged, and hence we keep it as a candidate corner point, if it moves less than 1 pixel between two iterations. We drop the point, if it is not converged within 3 iterations or move further than $2\sigma_I$ between iterations. Finally, we also drop points that during the iterations move too close to the image boundary. This is done in order to avoid boundary effects from the implementation of the scale-space representation. This is done by dropping points closer than $2\sigma_I$ to the image boundary. For more details on the algorithm for localization see Lindeberg (1998). In the experiments conducted in this study, we include results with and without this localization step.

Multi-scale Harris corners The Harris corner detector (Harris and Stephens, 1988) is based on the principal curvatures of the local autocorrelation function. The curvature of the autocorrelation function can be estimated by the eigenvalues of the structure tensor

$$\mathbf{T}(\mathbf{x}_0) = \int_{\mathbf{x}' \in \mathbb{R}^2} (\nabla L)(\nabla L)^T w_{\mathbf{x}_0} \, d\mathbf{x}' \ , \tag{7}$$

where $w_{\mathbf{x}_0}(\mathbf{x})$ is a window function giving higher weight to points close to the center point $\mathbf{x}_0$ and

$$(\nabla L)(\nabla L)^T(\mathbf{x}) = \begin{bmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{bmatrix}(\mathbf{x}) \ . \tag{8}$$

In the improved Harris corner detector (Schmid et al, 2000), image derivatives are computed using the linear scale-space representation (2) at a fixed scale σ_D. We will instead utilize the scale-space of $\nabla L(\mathbf{x}; \sigma_D)$ and hence the scale-spaces $(\nabla L)(\nabla L)^T(\mathbf{x}; \sigma_D)$ and $\mathbf{T}(\mathbf{x}; \sigma_D)$. In order to compensate for change of scale in the scale-space, $\nabla L(\mathbf{x}; \sigma_D)$ is substituted in (8) with the scale normalized gradient $\nabla_{\text{norm}} L(\mathbf{x}; \sigma_D) = \sigma_D \nabla L(\mathbf{x}; \sigma_D)$ (Lindeberg, 1998).

Harris and Stephens (1988) notice that the eigenvalues of (8) are rotational invariant estimates of the curvature, and a corner feature is present if both are large. For a 2×2 matrix $\mathbf{A}$, the eigenvalues (λ_1, λ_2) are related to the trace and determinant of $\mathbf{A}$ by $\lambda_1 \lambda_2 = \det(\mathbf{A})$ and $\lambda_1 + \lambda_2 = \text{Tr}(\mathbf{A})$. Based on this, the Harris corner measure is defined as

$$R = \det(\mathbf{T}) - \alpha \text{Tr}(\mathbf{T}) \ . \tag{9}$$

In our implementation we choose the weight $\alpha = 0.06$ as is common in the literature (Schmid et al, 2000). We use the Gaussian function as window function,

$$w_{\mathbf{x}_0}(\mathbf{x}; \sigma_I) = G(\mathbf{x} - \mathbf{x}_0; \sigma_I) \ , \tag{10}$$

and fix the relationship between integration window scale σ_I and differentiation scale σ_D as $\sigma_D = 0.7 \times \sigma_I$ similar to Mikolajczyk and Schmid (2004). Hence we define the multi-scale Harris corner measure $R(\mathbf{x}; \sigma_I)$ by using the scale-space $\mathbf{T}(\mathbf{x}; \sigma_I)$ in (9). We define the multi-scale Harris corner detector by detecting local maxima of the scale normalized multi-scale Harris measure $R(\mathbf{x}; \sigma_I)$.

The scale-space representation is implemented as described under the Lindeberg corner detector with 31 scale levels, such that the ith integration scale level is $\sigma_{I,i} = k^{(i-1)}\sigma_0$, $i = 1, \ldots, 31$, with $k = 1.1$ and $\sigma_0 = 1.5$ pixels. Similar to the Lindeberg corner detector we detect local maxima of $R(\mathbf{x}; \sigma_I)$, (9), in the discrete scale-space representation using the same maximum criterion as described above. Similar to the Lindeberg corner detector, we choose a threshold of 1500 and discard all points where $R(\mathbf{x}; \sigma_I)$ is less than this threshold. We also remove points closer than $2\sigma_I$ to the image boundary, in order to avoid spurious detections.

In our implementation of the multi-scale Harris corner detector, we further add a precise localization of the corner response based on the suggested approach by Lindeberg (1998) and outlined above. In the experiments conducted in this study, we include results with and without this localization step. We use the initially detected integration scale σ_I as the scale in the Gaussian integration window in (5) and (6).

Table 2 Summary of detectors. The *nl.* indicates that detector number 9 and 11 have no localization optimization, which is included in number 8 and 10.

Detector	#Detect.	Affine Invariant
Harris	1	no
Harris Laplace	2	no
Harris Affine	3	yes
Hessian Laplace	4	no
Hessian Affine	5	yes
MSER	6	yes
DoG	7	no
Multiscale Harris	8	no
Multiscale Harris nl.	9	no
Lindeberg corner	10	no
Lindeberg corner nl.	11	no

Notice that the Harris, (9), and Lindeberg, (3), corner measures differ and will detect different local image structure as corners. The multi-scale Harris corner detector as described above also differ from the Harris Laplace detector (Mikolajczyk and Schmid, 2001, 2004). In the Harris Laplace detector, the corner measure R in (9) is applied to each scale level image to detect a collection of candidate points. Following this, each candidate point is tested for whether or not it is a maxima across scale of the Laplacian $\Delta L = L_{xx} + L_{yy}$ measured at the point. The application of the Laplacian leads to different detected points than those produced by the multi-scale Harris corner detector.

Finally, note that a rigorous optimization of the parameters of both of the detectors has not been performed. We instead choose parameters that appeared sensible and gave good results on a couple of test images. Hence one may expect that better performance of the detectors is possible to achieve.

3.2 Interest point descriptors

We have chosen to investigate a number of interest point descriptors including a number of SIFT type descriptors (Lowe, 2004), DAISY descriptors with different parameters (Winder et al, 2009), and raw image patches compared with normalized cross correlation. The descriptors are listed in Tab. 3. As mentioned earlier, these were chosen, because they reportedly are the best performing (Mikolajczyk and Schmid, 2005; Moreels and Perona, 2007; Winder et al, 2009). We have used our own implementation of the descriptors to ensure that exactly the same image patch was used for the different descriptors. The descriptors are estimated on an affine warped image patch sampled according to the parameters obtained from the interest point detection and rotated to one dominant gradient direction.

Table 3 Summary of descriptors. The *Raw Patches* are matched using normalized cross correlation (*NCC*).

Descriptor	#Desc.	Color	#Dim.
Raw patch (*NCC*)	1	no	$34 \times 34 \times 3$
SIFT gray	2	no	128
SIFT RGB bin	3	yes	128
SIFT RGB	4	yes	384
Opponent SIFT	5	yes	384
CSIFT	6	yes	384
Gaussian opponent SIFT	7	yes	384
Hist. eq. SIFT	8	no	128
Hist. eq. SIFT RGB bin	9	yes	128
Hist. eq. SIFT RGB	10	yes	384
DAISY 1-6-6 s	11	no	52
DAISY 1-6-6 l	12	no	104
DAISY 1-8-8-8 s	13	no	100
DAISY 1-8-8-8 l	14	no	200

The image patches are chosen to cover an area of $\pm$ three times the scale of the interest point detector in pixels, which we found to give good matching properties. We discard patches that exceed the image boundaries. The image patches are estimated on an affine warped image patch sampled according to the parameters obtained from the interest point detection and rotated to one dominant gradient direction. Sampling is done in images convolved with a Gaussian according to the size of the image patch to avoid sampling artifacts. To make the implementation efficient we iteratively presmoothed the image ten times using a Gaussian convolution kernel with a standard deviation of $\sigma = 1$. This gave us a scale space in which we sampled the patch at the most appropriate scale relative to the size of the sampling patch. We found this to give satisfactory results compared to smoothing the image for individual samples, but with significant speed-up. The patch size was chosen to be 34×34 pixels, which was a tradeoff between calculation time and precision. Especially the normalized cross correlation is time consuming, because no approximate nearest neighbor search methods gave satisfactory results. For the other descriptors we employed the FLANN library (Muja and Lowe, 2009) for approximate nearest neighbor search, which gave significant speed-up and very little loss in precision.

We have chosen a range of descriptor implementations based on the SIFT framework. We investigate descriptor performance for both color and gray scale images. The primary difference is the color representation of the input image and the binning of the descriptor.

Raw image patches The raw image patches have been included as a baseline match based on normalized cross correlation.

SIFT gray is the traditional descriptor proposed by Lowe (2004). The descriptor is based on binning gradients in 8 directional bins sampled in a 4×4 spatial grid. The resulting descriptor is 128 dimensional.

SIFT RGB bin is the SIFT descriptor estimated independently for the red, green, and blue band, and the three obtained descriptors are averaged, giving a 128 dimensional descriptor.

SIFT RGB is the SIFT descriptor estimated independently for the red, green, and blue band concatenated to form one normalized vector of 384 dimensions.

Opponent SIFT is the color invariant descriptor proposed by Van De Sande et al (2010), which is based on the opponent color space. The opponent color space is given by

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{11}$$

The Opponent SIFT descriptor is estimated independently for the three opponent color channels (O_1, O_2, O_3), and concatenated to form a 384 dimensional descriptor.

CSIFT is the color invariant descriptor proposed by Abdel-Hakim and Farag (2006) and extended in Burghouts and Geusebroek (2009). The descriptor is based on the Gaussian opponent color model (Geusebroek et al, 2001), which is a linear transformation of the RGB channels

$$\begin{bmatrix} E \\ E_\lambda \\ E_{\lambda\lambda} \end{bmatrix} = \begin{bmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.60 & 0.17 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{12}$$

where E, E_λ and $E_{\lambda\lambda}$ are the decorrelation of the RGB channels into intensity, blue-yellow and green-red channels. From the Gaussian opponent color model the color invariants $W_{\lambda j}$, $C_{\lambda j}$ and $C_{\lambda\lambda j}$ can be calculated

$$W_{\lambda j} = \frac{E_{\lambda j}}{E}, \tag{13}$$

$$C_{\lambda j} = \frac{E_{\lambda j}E - E_\lambda E_j}{E^2}, \tag{14}$$

$$C_{\lambda\lambda j} = \frac{E_{\lambda\lambda j}E - E_{\lambda\lambda}E_j}{E^2}, \tag{15}$$

where $j \in \{x, y\}$ is used for spatial derivatives in the x and y directions. These color invariant derivatives are used to build the SIFT descriptor, resulting in a 384 dimensional normalized descriptor.

Gaussian opponent SIFT This descriptor is built similar to the SIFT RGB descriptor just based on the linear transformation to the Gaussian opponent color model in Eq. 12. This also results in a 384 dimensional descriptor.

Histogram Equalized SIFT Tang et al (2009) suggest an intensity invariant descriptor based on an ordinal transform of the local image patch. The ordinal transform is invariant to monotonic brightness change including nonlinear changes. An ordinal transformation is equivalent to histogram equalization. To include this approach for obtaining photometric invariance, we have chosen to use the SIFT framework instead of the radial spatial binning suggested in (Tang et al, 2009). We have tested three histogram equalized SIFT descriptors including the 128 dimensional gray level version. The second choise is a 128 dimensional binned RGB version, where each color band is histogram equalized independently and subsequently averaged over the bins. Finally, a 384 dimensional histogram equalized SIFT where each color band is equalized independently and the SIFT descriptors concatenated into one normalized descriptor.

DAISY In addition to the SIFT descriptors we tested four descriptors of the DAISY type. Our implementation closely follow the work of Winder et al (2009)[5] for the Type-II descriptor. This descriptor is parameterized by the number of spatial sampling points, their spatial extension and the type and sampling of differential geometric operator applied. Each spatial sampling point covers an image area, where the pixels are weighted by a Gaussian. The standard deviation σ of the Gaussian determines the spatial extent of the sampling region. The Type-II descriptor is estimated by taking the average positive and negative spatial differential in the x and y directions respectively

$$\psi_{\mathbf{x},s} = [\mathrm{p}(I_x), \mathrm{p}(-I_x), \mathrm{p}(I_y), \mathrm{p}(-I_y)], \tag{16}$$

$$\begin{aligned} \psi_{\mathbf{x},l} = [&\mathrm{p}(I_x), \mathrm{p}(-I_x), \mathrm{p}(I_y), \mathrm{p}(-I_y), \\ &\mathrm{p}(I_x - I_y), \mathrm{p}(I_y - I_x), \mathrm{p}(I_x + I_y), \mathrm{p}(-I_x - I_y)], \end{aligned} \tag{17}$$

where $\mathrm{p}(\cdot)$ indicates the positive part

$$\mathrm{p}(x) = \begin{cases} 0 \text{ if } x \leq 0 \\ x \text{ if } x > 0 \end{cases}. \tag{18}$$

[5] http://cvlab.epfl.ch/~brown/patchdata/patchdata.html

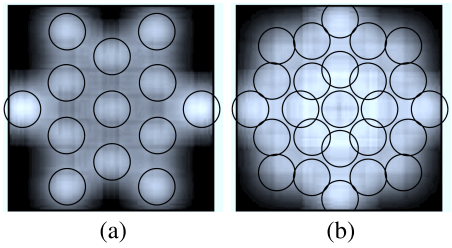


Fig. 5 Spatial layout of the DAISY descriptors. (a) Is DAISY 1-6-6 and (b) is DAISY 1-8-8-8. Rings indicate one standard deviation of the Gaussian sample weight, and the color indicate the weight of the pixels, with brighter colors indicating higher weights.

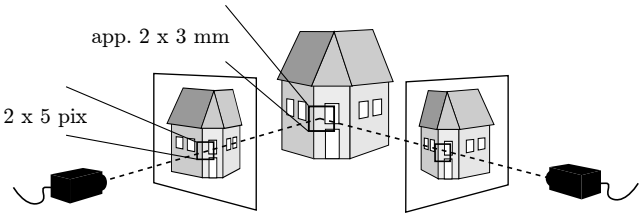


Fig. 6 Matching criterion for interest points. This figure gives a schematic illustration of a house scene depicted from two viewpoints. Corresponding descriptors are allowed a misalignment within a window with a radius of 5 pixels, which is approximately 3 mm on the scene surface. Ground truth is obtained from the surface geometry.

The $\psi_{\mathbf{x},s}$ is a part of the daisy descriptor sampled at position $\mathbf{x}$ and s and l indicates small and large dimensionality respectively. The full DAISY descriptors are sampled at a number of positions, as shown in Fig. 5, and the descriptors from the sample points are concatenated and normalized.

3.3 Evaluation criterion

The evaluation is based on determining if corresponding features are correctly matched or not. Features are matched by measuring the distance between interest point descriptors, and these matches are validated using the 3D surface scans of the scenes. We use the Euclidean distance when comparing descriptors, expect for the raw patch descriptor where we use normalized cross correlation. We employ an evaluation framework similar to the procedure described in (Dahl et al, 2011). Fig. 2 illustrates the image acquisition layout and we match all images to the key frame, which is the central image in Arc 1. Fig. 6 shows the criteria that we use for determining correct correspondence.

Given an image pair, where one image is the key frame, a detector-descriptor pair is evaluated by

1. For each feature in the key frame the distance to the best δ_b and the second best δ_s matching feature in the other image is found.
2. For each feature correspondence, we compute the ratio, $r = \frac{\delta_b}{\delta_s}$, between the match score of the best δ_b and the second best δ_s correspondence. r is close to zero if the best match score is much smaller than the second best (the match is unique), and close to one if the two best match scores are similar.
3. Using this ratio, r, as a predictor for correct matches, c.f. Lowe (2004), the ROC (Receiver Operating Characteristic) curve as a function of r is constructed based on all features in an image pair. The ROC curve is the *true positive rate* as a function of *false positive rate*. We compare the area under the ROC curve (AUC). The area is between zero and one, where one indicates perfect performance of the detector-descriptor pair and 0.5 corresponds to random matching.
4. The AUC is used as the performance measure of a detector-descriptor combination on a pair of images.

These AUCs are the basis for our statistical analysis. The AUC is chosen as a performance measure, in line with Winder et al (2009), because it elegantly removes the need to balance between many false positive or many false negatives, by integrating r out. As a result, it strongly relates to the underlying discriminative power of the method.

Carneiro and Jepson (2002) propose to use the receiver-operating characteristics (ROC) curve as a measure of performance for local image descriptors. It has later been argued by Ke and Sukthankar (2004) that ROC is not a good measure of performance for interest point matching based on arguments by Agarwal and Roth (2002) for object detection. This measure was subsequently used by Mikolajczyk and Schmid (2005) in their evaluation of interest point descriptors. The main argument against ROC is that it is difficult to estimate the total number of negatives in a dataset, which is needed in ROC. However, in our setup the uncertainty in evaluating negative and positive matches is the same and is governed by the imprecision of the 3D surface scan used as ground truth in the DTU Robot dataset. We therefore prefer to use the AUC as measure of performance, since it strikes a balance between positive and negative matches.

3.4 Statistical analysis

We have performed an analysis of variance (ANOVA) (Conradsen and Ersbøll, 2002) for the obtained AUC-scores to test if there is a significant effect of choosing

different descriptors, detectors or a specific combination of descriptors and detectors. The statistical model used is

$$\begin{aligned} \mathrm{AUC}_{obs} = \mu_{\mathrm{AUC}_{all}} &+ \alpha_{\mathrm{detector}} + \alpha_{\mathrm{descriptor}} \\ &+ \alpha_{\mathrm{light}} + \alpha_{\mathrm{camera\ position}} \\ &+ \alpha_{\mathrm{scene}} + \alpha_{\mathrm{detector}} \cdot \alpha_{\mathrm{descriptor}} + \epsilon \end{aligned} \tag{19}$$

where AUC_{obs} is the observed AUC, $\mu_{\mathrm{AUC}_{all}}$ is the overall mean AUC-score, and $\alpha_{(.)}$ is the difference from the mean for detector, descriptor, light, camera position, scene, and combined effect of detector and descriptor for each of the factors. The last term ϵ is the residual error.

The ANOVA is based on the assumption that each observation is an independent normal distributed stochastic variable with the same variance. The mean values of the factors in (19) are assumed to be additive. An F-test is employed to test if a factor has a significant effect on prediction of the model. The mean value of a given factor, e.g. the mean of the different detectors, are assumed to be the same, i.e. $\alpha_{detector\#1} = \alpha_{detector\#2} = \cdots = \alpha_{detector\#14} = 0$.

In our previous work (Dahl et al, 2011), we employed a simpler statistical model including only a few factors. In that work we were not able to obtain a statistical significant performance difference, but with the model (19) we obtain significance, even for very small performance differences.

The ANOVA tests, if a factor has significantly different levels, but this can be caused by one factor type performing very different from the other. So in addition to the ANOVA we have performed students t-tests for the most similar detectors, descriptors and combinations of the two. Students t-test is performed by

$$t = \frac{\mu_1 - \mu_2}{\sqrt{2\hat{\sigma}^2/n}} \ , \tag{20}$$

where μ_1 and μ_2 are the two means to be compared, $\hat{\sigma}$ is an estimate of the standard deviation and n is the number of observations. The ANOVA and t-test are based on an assumption of normal distributed noise. This is a safe assumption, since statistics is made on means of AUC, which due to the law of large numbers are normal distributed for all practical purposes.

4 Results

Our experiments illustrate the combined performance of interest point detectors and descriptors under diffuse and directional lighting conditions. In addition we have investigated the effect of scene type.

Table 4 Average performance of the interest point detectors in the diffuse lighting experiment μ_{AUC_d}, directional lighting experiment with light from right to left $\mu_{AUC_{rl}}$, light from back to front $\mu_{AUC_{bf}}$, and combined μ_{AUC}. Best performance is marked with bold face.

	μ_{AUC_d}	$\mu_{AUC_{rl}}$	$\mu_{AUC_{bf}}$	μ_{AUC}
Harris	0.801	0.754	0.774	0.776
Harris Laplace	0.808	0.767	0.773	0.782
Harris Affine	0.810	0.766	0.777	0.784
Hessian Laplace	0.760	0.715	0.735	0.737
Hessian Affine	0.786	0.757	0.765	0.769
MSER	0.852	0.776	0.798	0.809
DoG	0.856	0.804	0.814	0.825
Multiscale Harris	0.864	0.818	**0.837**	0.839
Multiscale Harris nl.	**0.870**	**0.819**	0.835	**0.841**
Lindeberg	0.819	0.775	0.782	0.792
Lindeberg nl.	0.813	0.772	0.779	0.788

Tab. 4 and 5 show the average performance for interest point detectors and descriptors respectively. In Tab. 6 the number of detected interest points and their standard deviation over scenes are shown. Tab. 7 shows the average performance of the scene type experiment, and in Fig. 7 and Tab. 8 - 11 the combined effect of detectors and descriptors are shown for the different experiments.

Our experiments show that the difference in detector performance is large, and the best performing detector is the multi-scale Harris corner detector. The difference in performance with (multi-scale Harris) and without localization (multi-scale Harris nl.) is marginal – in many situations the optimized localization degrades the performance, but most of these differences are not significant. DoG interest points performs second best followed by MSER, both with a performance that is statistically significant lower than the multi-scale Harris.

The descriptors deviate much less in their performance – only the raw patches have inferior performance. Opponent SIFT performs best, but its performance is only slightly better than many of the other descriptors, including Gaussian opponent SIFT, Histogram equalized SIFT RGB, and DAISY 1-8-8-8 l, which all have superior performance in some experiments. The difference between the best performing descriptors is in many cases not statistically significant.

The ANOVA analysis shows that there is a significant effect for all main factors, which includes detector, descriptor, light, camera position and scene. Also the combined factor of detector and descriptor is significant. From the ANOVA we get a statistical summary of all variables at once, e.g. we test if all interest point

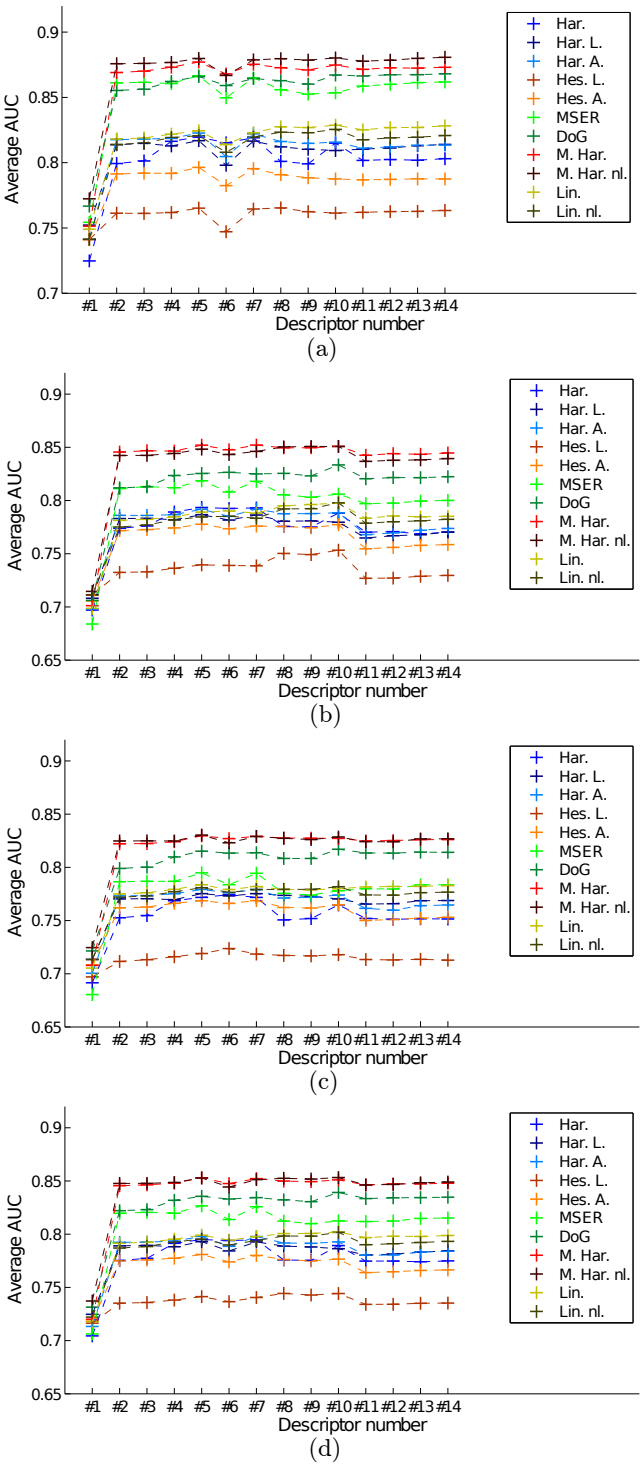


Fig. 7 Average AUC for detectors and descriptors for (a) diffuse lighting, (b) directional light from right to left, (c) directional light from back to front, and (d) overall average performance. Numbering is done according to Tab. 3. Note that there is no special ordering of the descriptors, so the dashed line connecting the points are just used to make it easier to distinguish the individual points.

detectors have the same mean value. To test for factors that have very similar performance we use a pairwise

Table 5 Average performance of the interest point descriptors in the diffuse lighting experiment μ_{AUC_d}, directional lighting experiment with light from right to left $\mu_{AUC_{rl}}$, light from back to front $\mu_{AUC_{bf}}$, and combined μ_{AUC}. Best performance is marked with bold face.

	μ_{AUC_d}	$\mu_{AUC_{rl}}$	$\mu_{AUC_{bf}}$	μ_{AUC}
Raw patch	0.749	0.706	0.703	0.719
SIFT gray	0.825	0.777	0.792	0.798
SIFT RGB bin	0.826	0.778	0.793	0.799
SIFT RGB	0.829	0.782	0.796	0.802
Opponent SIFT	**0.832**	**0.786**	0.801	**0.806**
CSIFT	0.819	0.783	0.797	0.800
Gauss. opp. SIFT	0.831	**0.786**	0.800	**0.806**
Heq. SIFT	0.828	0.779	0.799	0.802
Heq. SIFT RGB bin	0.826	0.779	0.798	0.801
Heq. SIFT RGB	0.829	0.782	**0.802**	0.804
DAISY 1-6-6 s	0.826	0.776	0.786	0.796
DAISY 1-6-6 l	0.827	0.776	0.787	0.797
DAISY 1-8-8-8 s	0.828	0.778	0.788	0.798
DAISY 1-8-8-8 l	0.829	0.778	0.789	0.799

t-test, but this only test the significance between two variables. Our experiments is however very similar for the different variables, which has allowed us to obtain a general measure for the mean values in Tab. 8 - 11 being significantly different.

The t-statistics is based on the difference in mean value of the two variables that are tested, their variance and the number of observations. We have the same number of observations in each experiment and they also have similar variance, except for the normalized cross correlation. Hereby we can estimate an approximate difference for the average AUC-measures being different. In the diffuse lighting experiment shown in Tab. 8 a difference of 0.001 is statistical significant on a 5% level whereas a difference of 0.002 is significant on a 0.05% level. In the relighting experiment from left to right shown in Tab. 9 the levels are 0.003 and 0.006 respectively, in back to front relighting experiment shown in Tab. 10 the levels are 0.002 and 0.004 respectively, and the overall performance in Tab. 11 has 0.001 and 0.002 respectively. From this it can be seen that many of the high performing descriptors are not significantly different.

The small difference in performance for the descriptors, especially when applied together with the multi-scale Harris detector, makes the choice of descriptor less important. Choosing the multi-scale Harris detector will give high performance independently of the choice of descriptor, with raw patches being the only exception. It is worth noting that there is a larger change in performance for other detectors, e.g. the relatively high performing MSER. With MSER the choice of de-

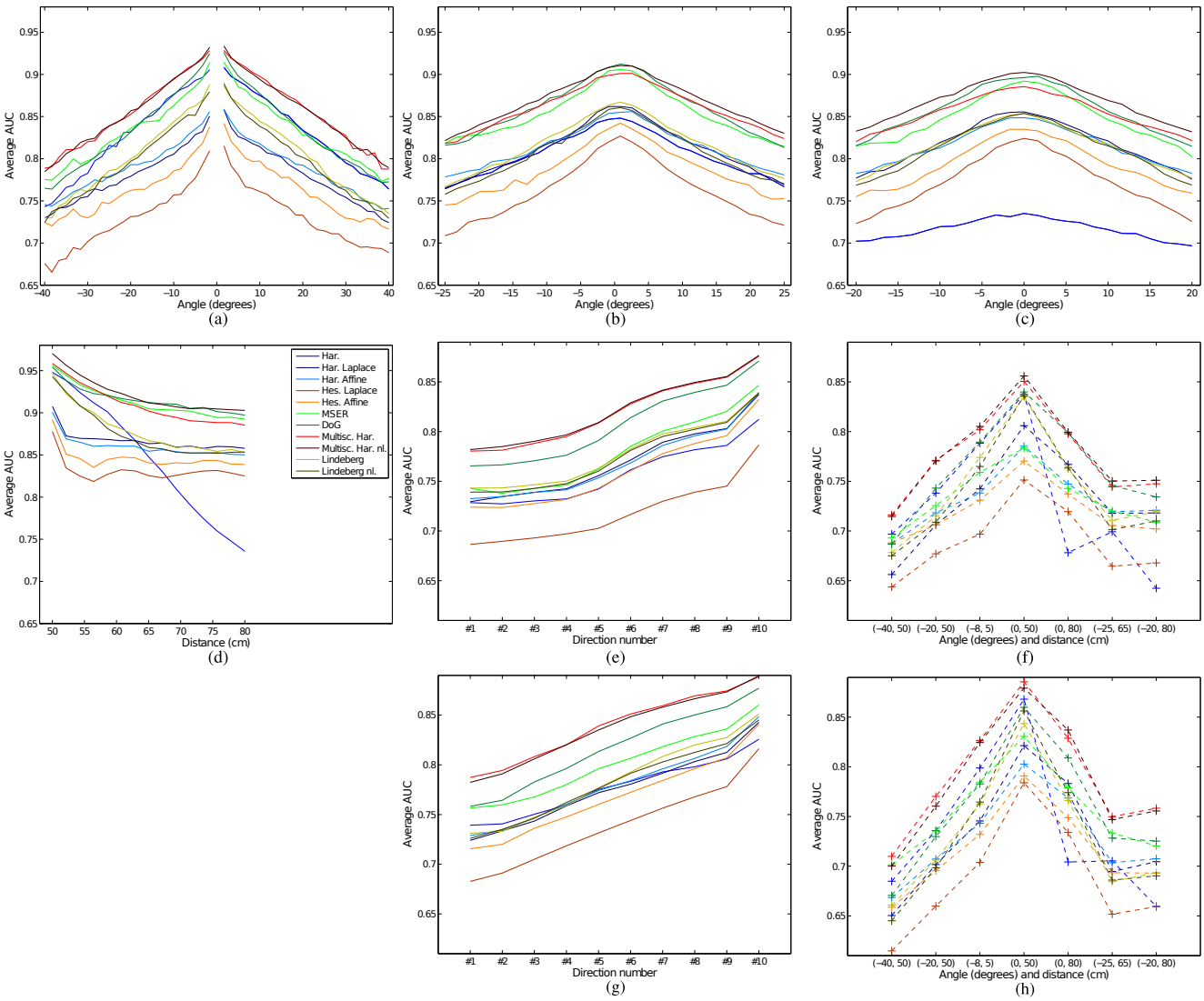


Fig. 8 Average AUC for interest point detectors for all spatial positions. The first four graphs cover all spatial positions for the diffuse light experiment. (a) Arc 1, (b) arc 2, (c) arc 3, and (d) linear path – see Fig. 2. The last four illustrate the directional lighting experiment. Average over the ten lighting directions for (e) right to left and (g) back to front, and the direction furthest away from the reference for each of the seven spatial positions in the (f) right to left lighting experiment and (h) back to front lighting experiment.

Table 6 Average number of detected interest points μ_n and the standard deviation over scenes σ_n for the diffuse lighting experiment.

	μ_n	σ_n
Harris	1834	1421
Harris Laplace	1839	1268
Harris Affine	1787	1215
Hessian Laplace	1722	1148
Hessian Affine	1373	981
MSER	819	629
DoG	1141	686
Multiscale Harris	1184	977
Multiscale Harris nl.	1605	1204
Lindeberg	1488	1147
Lindeberg nl.	2116	1541

scriptor is more important – here Opponent SIFT and Gaussian opponent SIFT are the best choices.

Camera position experiment Fig. 8 and 9 show the average performance of detectors and descriptors for all camera positions in the diffuse lighting experiment (a - d). In addition the directional lighting experiment is shown as average for the ten lighting directions (e, g) and the extreme positions (f, h) – i.e. the light direction furthest away from the reference. These positions are included to show the difference in performance when the descriptors and detectors are challenged the most.

The detector experiment (Fig. 8) shows a general high performance for the multi-scale Harris detector both with and without localization. However, DoG and

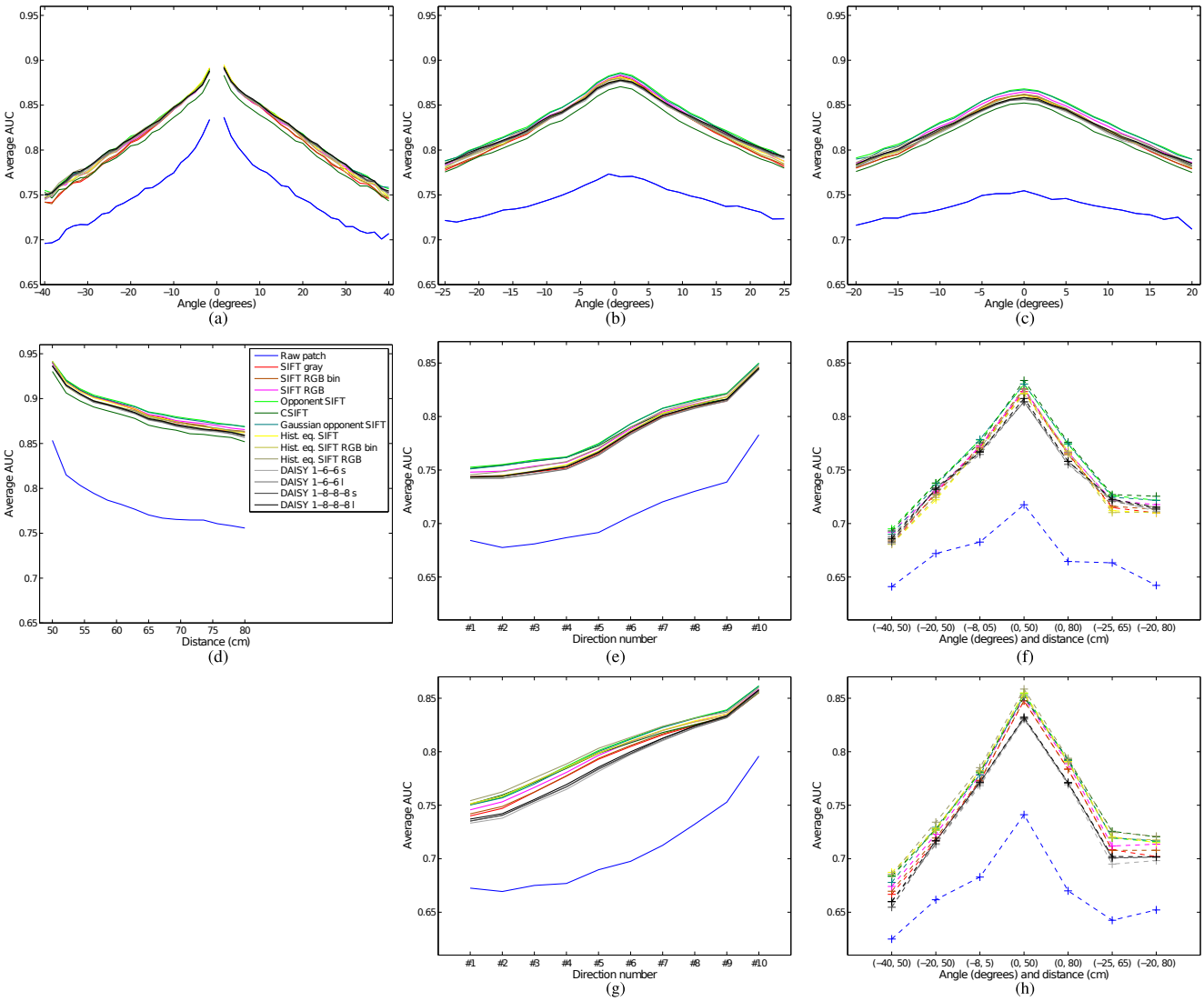


Fig. 9 Average AUC for interest point descriptors for all spatial positions. The first four graphs cover all spatial positions for the diffuse light experiment. (a) Arc 1, (b) arc 2, (c) arc 3, and (d) linear path – see Fig. 2. The last four illustrate the directional lighting experiment. Average over the ten lighting directions for (e) right to left and (g) back to front, and the direction furthest away from the reference for each of the seven spatial positions in the (f) right to left lighting experiment and (h) back to front lighting experiment.

MSER performs slightly better than multi-scale Harris with localization during change in scale (Fig. 8 (d)), but multi-scale Harris (nl.) without localization performs comparable to DoG and MSER. The Harris detectors with no scale adaption perform well when no scale change occurs, but this decreases dramatically with scale change. A general trend is that the Hessian Laplace and Hessian Affine detectors perform poorly.

Both of the multi-scale Harris detectors outperform all other detectors in the light directions experiment. Both as an overall average (Fig. 8 (e, g)) and at the extreme positions with largest changes (Fig. 8 (f, h)). DoG is generally performing better than MSER with light changes, and the Harris detectors also show good performance when no scale change takes place.

The descriptor experiment (Fig. 9) shows high performance for all descriptors compared to the normalized cross correlation of the raw image patches. But there are differences that are more expressed with changes in scale, angle, and light direction. Especially the Opponent SIFT and the Gaussian opponent SIFT descriptors perform overall good, closely followed by the Hist. eq. SIFT RGB. The CSIFT does not perform as good as the other descriptors in the diffuse lighting experiment, but in the directional lighting experiments it performs as good as the opponent descriptors (there is no statistical difference between these).

Scene type experiment Half of the 60 scenes in the DTU Robot Dataset have been categorized into five scene ty-

Fig. 10 Image examples of the five scene types. From left to right it is *Houses*, *Books*, *Fabric*, *Greens*, and *Beer cans*. Illustration from Aanæs et al (2010).

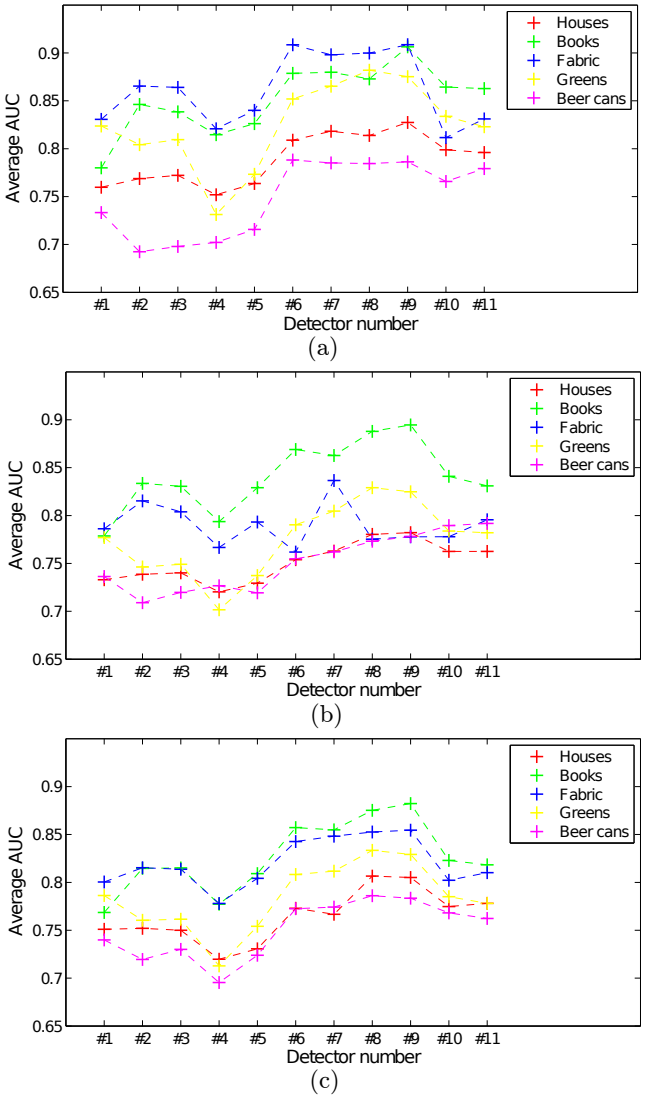


Fig. 11 Average AUC for detectors for different scene types with (a) diffuse lighting, (b) varying lighting from right to left, and (c) varying lighting from back to front. The numbering is done according to Tab. 2.

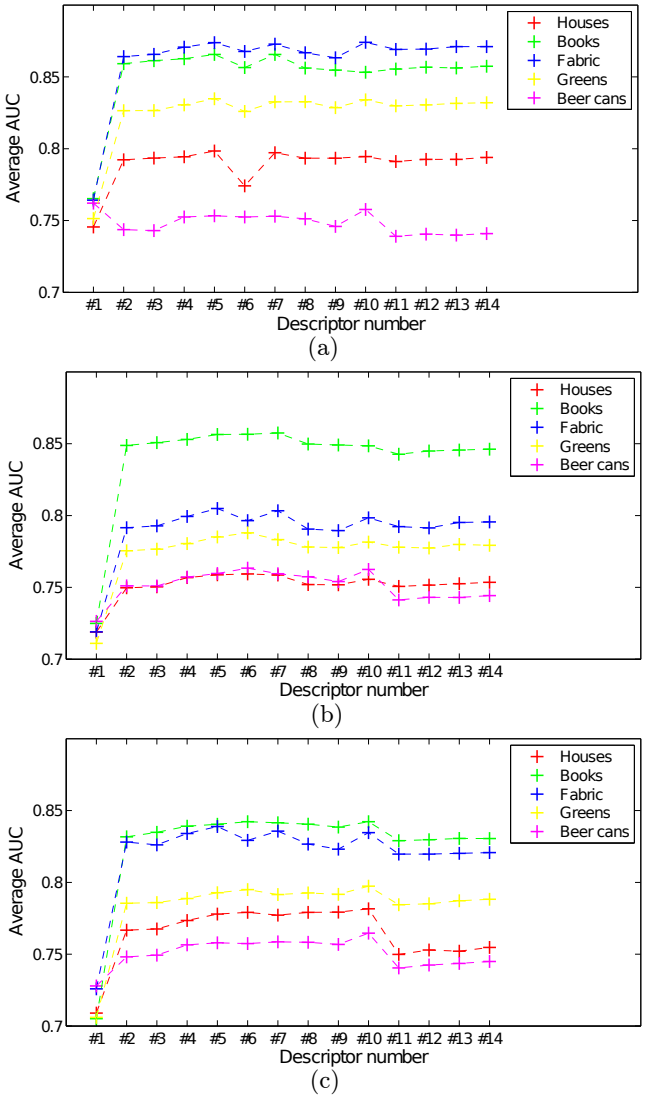


Fig. 12 Average AUC for descriptors for different scene types with (a) diffuse lighting, (b) varying lighting from right to left, and (c) varying lighting from back to front. The numbering is done according to Tab. 3.

pes, with image examples shown in Fig. 10 and average performance on these scene types is shown in Tab. 7. In this experiment we want to investigate how the combinations of detectors or descriptors perform on different scene types. The results are illustrated in Fig. 11 and 12.

For diffuse lighting Fig. 11 (a) and 12 (a), we see that *Fabric* and *Books* lead to high performance and are in general easy to handle for most combinations of detectors and descriptors. This is most likely caused by the distinct textures of the fabrics combined with their diffuse reflectance proptery as well as the flat dis-

Table 7 Average performance of all detector-descriptor combinations on the scene types illustrated in Fig. 10.

	# Scenes	μ_{AUC}
Houses	8	0.770
Books	4	**0.846**
Fabric	6	0.826
Greens	10	0.800
Beer cans	2	0.750

tinct texture of the books. We also see that *Beer cans* and *Houses* leads to poor performance for all combinations of detectors and descriptors. For the beer cans this may be caused by the specular surface reflectivity of the beer can surfaces. The *Houses* category consists of scenes of model houses, which has high surface complexity combined with repetitive patterns, which leads to poor performance across the board.

For both of the directional lighting experiments Fig. 11 (b), (c) and 12 (b), (c), we see that performance drops in general on *Fabric*, *Greens*, and *Houses*. At the same time we see that the performance on *Books* and *Beer cans* stays almost the same as in the diffuse lighting experiment. This is most likely caused by the geometric complexity of *Fabric*, *Greens*, and *Houses*, which leads to a high degree of shadows which in turns throws off the detectors and descriptors. Shadow effects is not as dominating for *Books* and *Beer cans*.

The general trend for both the detectors and the descriptors is the same as for the entire dataset. There are large differences in the performance of the interest point detectors, but only little difference in the descriptors. Again the multi-scale Harris detector performs good as well as the Opponent SIFT and Gaussian opponent SIFT.

5 Discussion

Our main motivation for this paper was to investigate the combined performance of interest point detectors and descriptors under varying camera geometry, light, and scene type, based on the DTU Robot dataset (Aanæs et al, 2012). Our results clearly show that the independently superior detectors and descriptors also have superior performance in combination. The optimal combination is the multi-scale Harris corner interest points with the color invariant opponent SIFT descriptor. The multi-scale Harris without localization (nl.) has slightly better performance than the one with localization but this difference is not significant. Also the Histogram equalized SIFT RGB descriptor has a general high performance together with the multi-scale Harris corners without localization, which is not significantly different from the opponent SIFT.

The difference in performance is much larger for the interest point detectors than for the descriptors. Despite the little difference in performance of the descriptors, most of them are statistical significant. This is caused by the size of the DTU Robot dataset, which has allowed us to obtain statistical significant results, even for small differences in performance, which generally strengthens our findings.

In our investigation we have not seen dependence between interest point detectors and descriptors. We investigated this dependence because both the interest point detectors and descriptors are based on local differential geometric properties. Hence, the local image appearance for e.g. a corner interest point will be biased towards a corner pattern and the appearance of a blob will be a blob pattern. Therefore there could have been a cross effect resulting in a combined superior performance. But we did not observe a performance difference, which indicates that the detectors can be chosen based on their performance independently of the descriptor and visa versa.

One explanation why we see independence between interest point detectors and descriptors can be that we sample an image patch at three times the scale of the interest point. Hereby a larger part of the image is included for the descriptor than in the interest point, and these additional patterns are independent of the pattern of the interest point.

Descriptors The descriptors that we have included in our evaluation are the raw image patches, nine variations of the SIFT descriptor, and four variations of the DAISY descriptor. All descriptors, except the raw image patches, are based on first order image derivatives. The variation is primarily the color representation (variations of SIFT) and spatial sampling (variations of DAISY). The performance difference between these descriptors turned out to be relatively small, which might be a reason for the very small difference in the combined effect of the detectors and descriptors, compared to their individual performance.

The very small difference between descriptors is not expected, because previous studies have shown a much clearer difference in performance of these descriptors. In (Winder et al, 2009; Brown et al, 2011), the DAISY significantly outperforms SIFT, whereas DAISY and SIFT performs very similar on our dataset. This is similar to what we observed in our previous studies (Dahl et al, 2011) on the DAISY descriptor, where we tested a range of parameter choices. The primary difference between our investigations and the results of (Winder

Table 8 Average performance of descriptors and detectors for the diffuse lighting experiment. A difference of approximately 0.001 is statistical significant on a 5% level whereas a difference of approximately 0.002 is significant on a 0.05% level.

	Har	HarL	HarA	HesL	HesA	MSER	DoG	MHar	MHarNL	Lin	LinNL
Raw patch	0.725	0.752	0.741	0.741	0.741	0.754	0.767	0.751	0.772	0.749	0.741
SIFT gray	0.799	0.814	0.818	0.761	0.791	0.861	0.855	0.869	0.876	0.818	0.814
SIFT RGB bin	0.801	0.815	0.818	0.761	0.792	0.862	0.856	0.870	0.876	0.819	0.815
SIFT RGB	0.816	0.813	0.819	0.762	0.792	0.861	0.862	0.873	0.877	0.822	0.819
Opponent SIFT	0.820	0.817	0.823	0.765	0.796	0.867	0.866	0.877	0.880	0.824	0.821
CSIFT	0.815	0.798	0.805	0.747	0.782	0.850	0.859	0.868	0.867	0.814	0.808
Gaussian opponent SIFT	0.819	0.817	0.822	0.765	0.795	0.865	0.864	0.875	0.879	0.823	0.820
Hist. eq. SIFT	0.801	0.812	0.816	0.765	0.791	0.856	0.863	0.873	0.880	0.827	0.823
Hist. eq. SIFT RGB bin	0.799	0.810	0.815	0.762	0.788	0.852	0.860	0.871	0.879	0.827	0.823
Hist. eq. SIFT RGB	0.815	0.809	0.816	0.761	0.788	0.853	0.867	0.875	0.880	0.829	0.826
DAISY 1-6-6 s	0.802	0.810	0.811	0.762	0.787	0.859	0.866	0.871	0.878	0.825	0.818
DAISY 1-6-6 l	0.802	0.811	0.812	0.763	0.787	0.860	0.867	0.873	0.878	0.827	0.819
DAISY 1-8-8-8 s	0.802	0.813	0.813	0.763	0.788	0.861	0.867	0.872	0.880	0.827	0.819
DAISY 1-8-8-8 l	0.803	0.814	0.814	0.763	0.788	0.862	0.868	0.873	**0.881**	0.828	0.821

Table 9 Average performance of descriptors and detectors for the experiment with light variation from right to left. A difference of approximately 0.003 is statistical significant on a 5% level whereas a difference of approximately 0.006 is significant on a 0.05% level.

	Har	HarL	HarA	HesL	HesA	MSER	DoG	MHar	MHarNL	Lin	LinNL
Raw patch	0.691	0.714	0.700	0.697	0.708	0.680	0.721	0.708	0.725	0.705	0.713
SIFT gray	0.753	0.770	0.773	0.712	0.762	0.786	0.799	0.822	0.825	0.775	0.772
SIFT RGB bin	0.755	0.771	0.773	0.713	0.763	0.787	0.800	0.822	0.825	0.776	0.773
SIFT RGB	0.769	0.770	0.775	0.716	0.766	0.787	0.810	0.824	0.825	0.779	0.777
Opponent SIFT	0.772	0.775	0.779	0.719	0.769	0.795	0.815	**0.830**	**0.830**	0.784	0.781
CSIFT	0.774	0.773	0.776	0.724	0.766	0.784	0.813	0.827	0.823	0.779	0.776
Gaussian opponent SIFT	0.772	0.775	0.779	0.718	0.769	0.794	0.814	0.829	0.829	0.782	0.779
Hist. eq. SIFT	0.751	0.774	0.771	0.717	0.762	0.776	0.808	0.828	0.827	0.779	0.779
Hist. eq. SIFT RGB bin	0.752	0.773	0.772	0.717	0.762	0.774	0.808	0.827	0.826	0.779	0.779
Hist. eq. SIFT RGB	0.765	0.770	0.774	0.718	0.765	0.778	0.817	0.827	0.829	0.780	0.782
DAISY 1-6-6 s	0.752	0.765	0.761	0.713	0.750	0.780	0.814	0.825	0.824	0.782	0.774
DAISY 1-6-6 l	0.751	0.766	0.760	0.713	0.751	0.780	0.813	0.825	0.824	0.782	0.774
DAISY 1-8-8-8 s	0.752	0.769	0.764	0.714	0.752	0.784	0.814	0.826	0.827	0.782	0.776
DAISY 1-8-8-8 l	0.751	0.769	0.765	0.713	0.753	0.784	0.814	0.826	0.827	0.783	0.777

Table 10 Average performance of descriptors and detectors for the experiment with light variation from back to front. A difference of approximately 0.002 is statistical significant on a 5% level whereas a difference of approximately 0.004 is significant on a 0.05% level.

	Har	HarL	HarA	HesL	HesA	MSER	DoG	MHar	MHarNL	Lin	LinNL
Raw patch	0.697	0.708	0.698	0.711	0.706	0.684	0.706	0.701	0.715	0.699	0.711
SIFT gray	0.774	0.783	0.786	0.732	0.772	0.811	0.812	0.846	0.842	0.781	0.775
SIFT RGB bin	0.776	0.783	0.786	0.733	0.773	0.813	0.813	0.847	0.842	0.782	0.777
SIFT RGB	0.789	0.782	0.787	0.736	0.774	0.812	0.823	0.847	0.844	0.785	0.782
Opponent SIFT	0.793	0.787	0.792	0.740	0.778	0.818	0.825	**0.852**	0.848	0.789	0.785
CSIFT	0.792	0.782	0.785	0.739	0.773	0.808	0.826	0.848	0.843	0.790	0.785
Gaussian opponent SIFT	0.793	0.786	0.792	0.739	0.776	0.818	0.825	**0.852**	0.846	0.788	0.783
Hist. eq. SIFT	0.776	0.781	0.788	0.750	0.776	0.805	0.825	0.849	0.850	0.795	0.792
Hist. eq. SIFT RGB bin	0.775	0.781	0.788	0.749	0.774	0.803	0.823	0.850	0.851	0.796	0.792
Hist. eq. SIFT RGB	0.788	0.780	0.788	0.753	0.777	0.806	0.833	0.851	0.851	0.798	0.798
DAISY 1-6-6 s	0.770	0.765	0.768	0.727	0.755	0.797	0.820	0.843	0.837	0.783	0.779
DAISY 1-6-6 l	0.771	0.767	0.770	0.727	0.756	0.798	0.822	0.844	0.838	0.785	0.780
DAISY 1-8-8-8 s	0.769	0.768	0.772	0.729	0.758	0.799	0.821	0.843	0.838	0.785	0.781
DAISY 1-8-8-8 l	0.770	0.770	0.774	0.730	0.759	0.800	0.822	0.845	0.839	0.785	0.782

Table 11 Overall average performance of descriptors and detectors for all experiments. A difference of approximately 0.001 is statistical significant on a 5% level whereas a difference of approximately 0.002 is significant on a 0.05% level.

	Har	HarL	HarA	HesL	HesA	MSER	DoG	MHar	MHarNL	Lin	LinNL
Raw patch	0.704	0.725	0.713	0.716	0.718	0.706	0.731	0.720	0.737	0.718	0.722
SIFT gray	0.775	0.789	0.792	0.735	0.775	0.820	0.822	0.846	0.848	0.791	0.787
SIFT RGB bin	0.777	0.790	0.793	0.736	0.776	0.821	0.823	0.846	0.848	0.792	0.788
SIFT RGB	0.791	0.788	0.793	0.738	0.777	0.820	0.832	0.848	0.849	0.795	0.793
Opponent SIFT	0.795	0.793	0.798	0.741	0.781	0.827	0.836	**0.853**	**0.853**	0.799	0.795
CSIFT	0.794	0.784	0.789	0.737	0.774	0.814	0.833	0.848	0.844	0.794	0.790
Gaussian opponent SIFT	0.795	0.793	0.798	0.741	0.780	0.826	0.834	0.852	0.851	0.798	0.794
Hist. eq. SIFT	0.776	0.789	0.792	0.744	0.776	0.812	0.832	0.850	0.852	0.801	0.798
Hist. eq. SIFT RGB bin	0.775	0.788	0.791	0.743	0.775	0.810	0.831	0.849	0.852	0.801	0.798
Hist. eq. SIFT RGB	0.789	0.786	0.793	0.744	0.777	0.812	0.839	0.851	**0.853**	0.802	0.802
DAISY 1-6-6 s	0.775	0.780	0.780	0.734	0.764	0.812	0.833	0.846	0.846	0.797	0.790
DAISY 1-6-6 l	0.775	0.781	0.780	0.734	0.765	0.812	0.834	0.847	0.847	0.798	0.791
DAISY 1-8-8-8 s	0.774	0.783	0.783	0.735	0.766	0.815	0.834	0.847	0.848	0.798	0.792
DAISY 1-8-8-8 l	0.775	0.784	0.784	0.735	0.766	0.815	0.835	0.848	0.849	0.799	0.793

et al, 2009; Brown et al, 2011) is the dataset employed, where we have a much larger number of very different scenes. Similarly the color descriptors CSIFT and opponent SIFT have in (Abdel-Hakim and Farag, 2006; Burghouts and Geusebroek, 2009; Van De Sande et al, 2010) been reported to significantly outperform the traditional SIFT. In our investigation this difference is relatively small. The reason might be that the DTU Robot dataset has changing lighting direction, which indirectly changes the scene colors.

The histogram equalized SIFT descriptors, which was inspired by Tang et al (2009), is harder to compare to this work, because their descriptor also includes a different spatial layout than the SIFT descriptor. But their ordinal spatial intensity distribution (OSID) descriptor has much better performance on the Oxford dataset from Mikolajczyk et al (2005). The ordinal labeling of the OSID descriptor is equivalent to histogram equalization. To make a descriptor that is similar to the OSID descriptor, but comparable to the other descriptors in our experiment, we chose to make SIFT descriptors from histogram equalized image patches. Hereby we obtained descriptors with the same non-linear invariance properties as the OSID descriptors. These descriptors did however not show as large a performance gain in our experiments as reported for the OSID features on the Oxford dataset.

The illumination variation in our investigations is caused by change in lighting angle and camera position. This results in some change in surface brightness and color, but especially surface texture changes due to small shadows, which changes the differential geometric properties of the images drastically. The color invariant descriptors are designed to be invariant to nonlinear color variation like color temperature, which is not directly a part of the DTU Robot dataset. Therefore, the color invariant descriptors could show larger performance gain on other datasets that reflect other types of color variation. Our investigations show that there is no loss in general performance by choosing the color invariant descriptors.

The reflectance model used in the opponent SIFT descriptors (Van De Sande et al, 2010) are particularly suited for diffuse materials like fabric. But in our experiments we do not see a relatively better performance of the opponent SIFT for scene types like Fabric. The opponent SIFT descriptor just has an overall good performance.

The descriptors vary in dimensionality, because some are based on gray scale images and others are sampled in a three-dimensional color space. Burghouts and Geusebroek (2009) chose to project the descriptors to a subspace using PCA to compare equally sized descriptors. But the choice of subspace projection will affect the discriminative properties of the descriptors and in this way their performance. We choose not to do so, which has resulted in an improved performance for the high dimensional descriptors. The relatively small performance gain for the color descriptors is therefore a combined effect of high dimensional descriptors and the color transformation – so the choice of color transformation is not very important using this dataset.

Detectors A more profound difference in performance is seen in the choice of detector, where the multi-scale Harris detector is superior closely followed by DoG and MSER detectors. This performance ranking is consistent across various descriptors. The simple fixed scale Harris corner detector also performs well for small scale changes. The benefit of the original Harris corner detector is its algorithmic simplicity.

It is worth noting that the DoG interest point detector (Lowe, 2004) is an approximation of the Hes-

sian Laplace interest point and the multi-scale Harris is somewhat similar to the Harris Laplace interest point detector. Despite these similarities, the difference in performance is profound. This implies that small algorithmic differences can have significant impact on the performance of the interest point detectors.

The general high performance of the MSER detector also complies with the investigations by Mikolajczyk et al (2005), whereas the Harris Affine detector, which was found to be almost as good as MSER, has a relative low performance in our investigation. The good performance of the DoG detector is consistent with our previous results (Dahl et al, 2011; Aanæs et al, 2012), but the superior performance of the multi-scale Harris detector was a surprise. Mikolajczyk and Schmid (2001, 2004) reported that scale selection with Harris corners was problematic, because the Harris corners rarely attains maxima over scale. This observation was based on experiments carried out on the Oxford dataset, and they suggested to use the Laplacian for scale selection, resulting in the Harris Laplace detector. In our work we found a solution for the Harris scale selection problem by allowing corners to shift across the scale space volume. This multi-scale Harris detector significantly outperforms all other detectors including the Harris Laplace detector. We also implemented Lindeberg corners, but it only showed medium performance.

For both the multi-scale Harris and the Lindeberg corners we applied the localization approach suggested by Lindeberg (1998). But we were not able to show performance improvements. This might be explained by the fact that this localization optimization only results in small position changes that are much smaller than what is acceptable by our matching criteria. Similarly we did not find an improved effect with the affine adapted interest points, which is employed in MSER, Harris Affine, and Hessian Affine detectors. This is also consistent with findings by Aanæs et al (2012). However, this type of invariance may have merit in e.g. 3D reconstruction of urban type scenes or other near-planar scenes.

The DTU Robot dataset also allows for an investigation into the effects of light change, with varying direction and diffusivity of the light. The change of light significantly degrades the performance of all interest point detectors and descriptors, which is similar to our findings for the investigated detectors in Aanæs et al (2012). The light variations in the DTU Robot dataset appears less severe than in real outdoor scenes, so detectors and descriptors could potentially be improved by investigating ways to achieve better invariance to light variation.

6 Conclusions

The best interest point detector and descriptor combination is the multi-scale Harris detector and the opponent SIFT descriptor. These are also the superior detectors and descriptors independently. We have performed a systematic investigation of the combined performance of a number of interest point detectors and descriptors in order to answer this question. We used the DTU Robot dataset, allowing a thorough statistical analysis. In addition to the finding that the detectors and descriptors performing best independently also performs best in combination, we found a large difference in performance of interest point detectors with the multi-scale Harris detector as the superior closely followed by MSER and DoG. We did not find improvements by applying affine invariance or optimized corner localization of the interest points. The difference in descriptors is much smaller but still significant with the opponent SIFT as the superior descriptor closely followed by CSIFT. The overall performance was significantly reduced by change in lighting directions and generally poor performance was obtained on specular surfaces. There is potentially a large gain to be obtained by introducing invariance to change in light directions and specular object, however these are hard future challenges.

7 Acknowledgements

We would like to thank the Oxford Vision Group[6] and David Lowe[7] for making their code available online. Furthermore, this work was in part 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 AL, Pedersen KS (2010) On recall rate of interest point detectors. In: Proceedings of 3DPVT, URL http://campwww.informatik.tu-muenchen.de/3DPVT2010/data/media/e-proceeding/session07.html#paper97

Aanæs H, Dahl AL, Pedersen KS (2012) Interesting interest points - a comparative study of interest point performance on a unique data set. International Journal of Computer Vision 97(1):18–35

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

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

Abdel-Hakim AE, Farag AA (2006) Csift: A sift descriptor with color invariant characteristics. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'06, vol 2, pp 1978–1983

Agarwal S, Roth D (2002) Learning a sparse representation for object detection. In: Proceedings of ECCV'02, Springer, LNCS, pp 113–130

Agarwal S, Snavely N, Simon I, Seitz S, Szeliski R (2009) Building rome in a day. In: Computer Vision, 2009 IEEE 12th International Conference on, Ieee, pp 72–79

Balmashnova E, Florack L (2008) Novel similarity measures for differential invariant descriptors for generic object retrieval. Journal of Mathematical Imaging and Vision 31(2-3):121–132

Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. Proceedings of ECCV'06 pp 404–417

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3):346–359

Brown M, Hua G, Winder S (2011) Discriminative Learning of Local Image Descriptors. IEEE Transaction on Pattern Analysis and Machine Intelligence 33(1):43–57

Brown M, Lowe D (2005) Unsupervised 3d object recognition and reconstruction in unordered datasets. In: 3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on, IEEE, pp 56–63

Burghouts GJ, Geusebroek JM (2009) Performance evaluation of local colour invariants. Computer Vision and Image Understanding 113(1):48–62

Carneiro G, Jepson A (2002) Phase-based local features. In: Heyden A, Sparr G, Nielsen M, Johansen P (eds) Proceedings of ECCV'02, Springer, Springer LNCS, vol LNCS 2350, pp 282–296

Conradsen K, Ersbøll BK (2002) An Introduction to Statistics. DTU Informatics

Crandall D, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: Proceedings of the 18th international conference on World wide web, ACM, pp 761–770

Dahl A, Aanæs H, Pedersen K (2011) Finding the best feature detector-descriptor combination. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011, pp 318–325

Everingham M, Zisserman A, Williams CKI, Van Gool L (2006) The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2007) The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

Florack L, ter Haar Romeny B, Koenderink J, Viergever M (1993) Cartesian differential invariants in scale-space. Journal of Mathematical Imaging and Vision 3(4):327–348

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

Freeman W, Adelson E (1991) The design and use of steerable filters. IEEE Transaction on Pattern Analysis and Machine Intelligence 13(9):891–906

Geusebroek JM, van den Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(12):1338–1350

Geusebroek JM, Burghouts GJ, Smeulders AWM (2005) The Amsterdam library of object images. International Journal of Computer Vision 61(1):103–112

ter Haar Romeny B (1994) Geometry-driven diffusion in computer vision, vol 320. Kluwer Academic Publishers

Harris C, Stephens M (1988) A combined corner and edge detector. In: 4th Alvey Vision Conf., pp 147–151

Hua G, Brown M, Winder S (2007) Discriminant embedding for local image descriptors. Proceedings of ICCV'07 pp 1–8

Ke Y, Sukthankar R (2004) Pca-sift: A more distinctive representation for local image descriptors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'04, Los Alamitos, CA, USA, vol 2, pp 506–513

Koenderink JJ, van Doorn AJ (1987) Representation of local geometry in the visual system. Biological Cybernetics 55:367–375

Koenderink JJ (1984) The structure of images. Biological Cybernetics 50:363–370

Lindeberg T (1994) Scale-space theory in computer vision. Springer

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

Loog M, Lauze F (2010) The improbability of harris interest points. IEEE Transactions on Pattern Analysis and Machine Intelligence 32:1141–1147, DOI 10.1109/TPAMI.2010.53, URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5432198

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 (2001) Indexing based on scale invariant interest points. In: Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, pp 525–531, URL `http://perception.inrialpes.fr/Publications/2001/MS01a`

Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. IJCV 60(1):63–86

Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Transaction 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

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

Muja M, Lowe DG (2009) Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application VISSAPP'09), pp 331–340

Schmid C, Mohr R (1997) Local grayvalue invariants for image retrieval. IEEE Transaction 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

Snavely N, Seitz S, Szeliski R (2006) Photo tourism: exploring photo collections in 3D. In: ACM SIGGRAPH 2006 Papers, ACM, pp 835–846

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

Snoek C, Worring M, Van Gemert J, Geusebroek J, Smeulders A (2006) The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, pp 421–430

Strecha C, von Hansen W, Van Gool L, Fua P, Thoennessen U (2008) On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'08, pp 1–8

Tang F, Lim SH, Chang NL, Tao H (2009) A novel feature descriptor invariant to complex brightness changes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'09., IEEE, pp 2631–2638

Tola E, Lepetit V, Fua P (2008) A Fast Local Descriptor for Dense Matching. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'08

Tola E, Lepetit V, Fua P (2009) DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE Transaction on Pattern Analysis and Machine Intelligence 32(5):815–830

Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: a survey. Found Trends Comput Graph Vis 3(3):177–280

Van De Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1582–1596

Winder SAJ, Brown M (2007) Learning local image descriptors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'07, pp 1–8

Winder S, Hua G, Brown M (2009) Picking the best daisy. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'09, pp 178–185

Witkin AP (1983) Scale space filtering. In: Proc. of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, Germany, vol 2, pp 1019–1023

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

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.

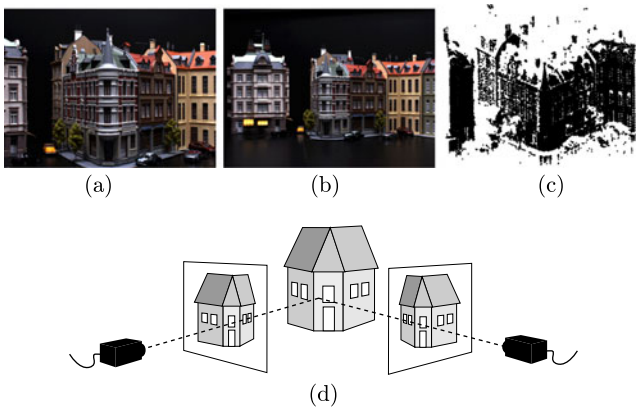


Fig. 1 Example of data and setup. Two images of the same scene with (**a**) one close up, (**b**) one distant from the side, and (**c**) the reconstructed 3D points. Illustration (**d**) of corresponding images with known geometric information including camera positions and 3D scene surface

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.

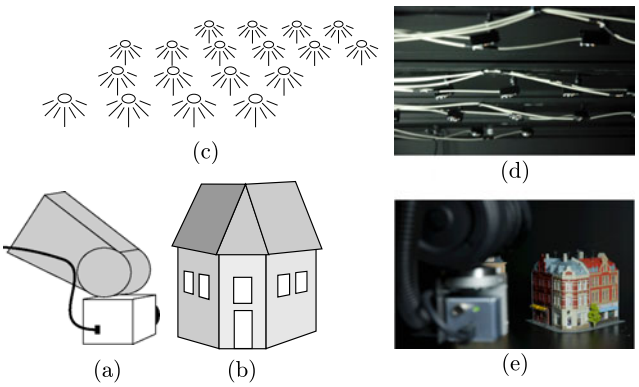


Fig. 2 Illustration of data collection setup. The camera (**a**) is mounted on a robot arm (**b**) capturing images of the scene. (**c**) LED point light sources illuminate the scene from 19 individual positions. (**d**), (**e**) photos of the experimental setup

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

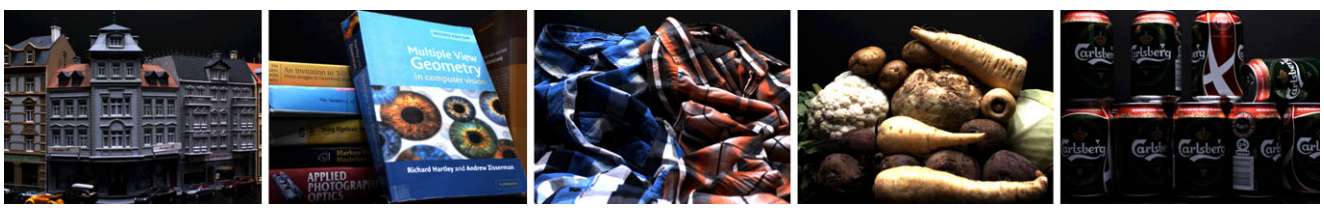


Fig. 3 Example images from our data set. The images show a diffuse relighting obtained by a linear combination of the 18 directional illuminated images. From *left* the scenes are examples of houses, books, fabric, greens, and beer cans, which have been used in our feature matching experiment with light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

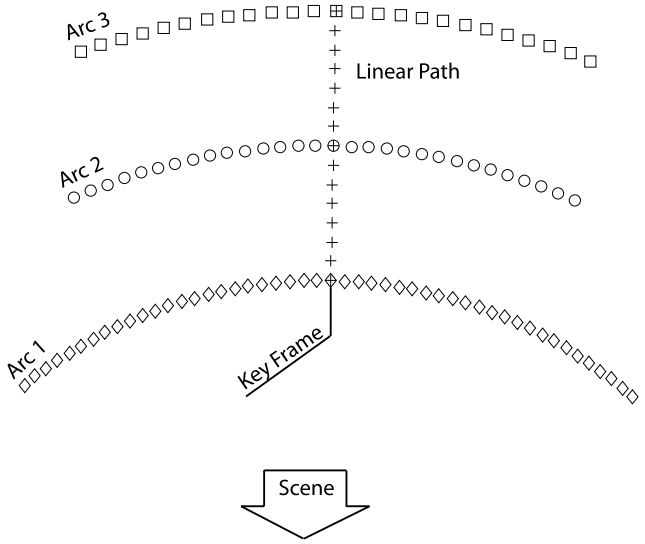


Fig. 4 Camera positions. The camera is placed in 119 positions in three horizontal arcs and a linear path away from the scene. The central frame in the nearest arc is the key frame, and the surface reconstruction is attempted to cover most of this frame. The three arcs are located on circular paths with 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°

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.

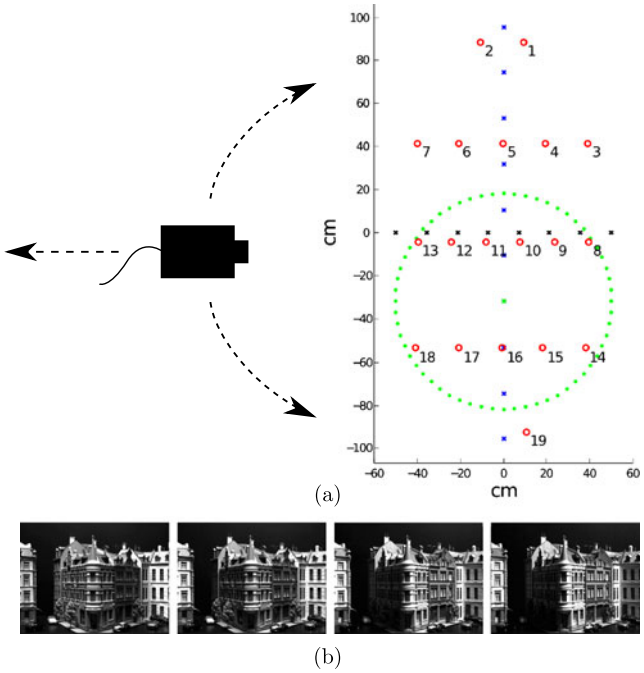


Fig. 5 (**a**) light stage setup seen from above and (**b**) example images with light from left to right. The layout of the light setup is illustrated with the *red circles* showing the positions of the white LEDs. The *axis* shown in (**a**) are in cm to illustrate the actual size of the setup, and 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

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

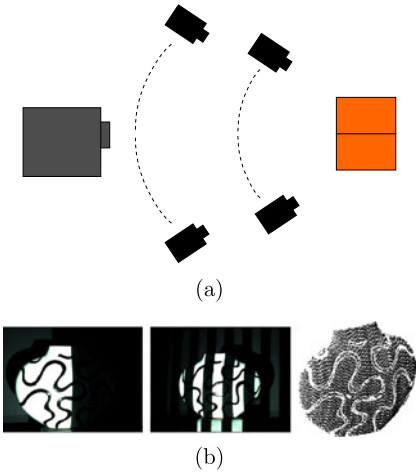


Fig. 6 Surface reconstruction is done with the setup shown in (**a**). We use a projector (*left*) to project a stripe pattern onto the scene (*right*), and we acquire images from four positions (*middle*). In (**b**) two stripe image examples are shown together with the reconstructed 3D point set (*right*)

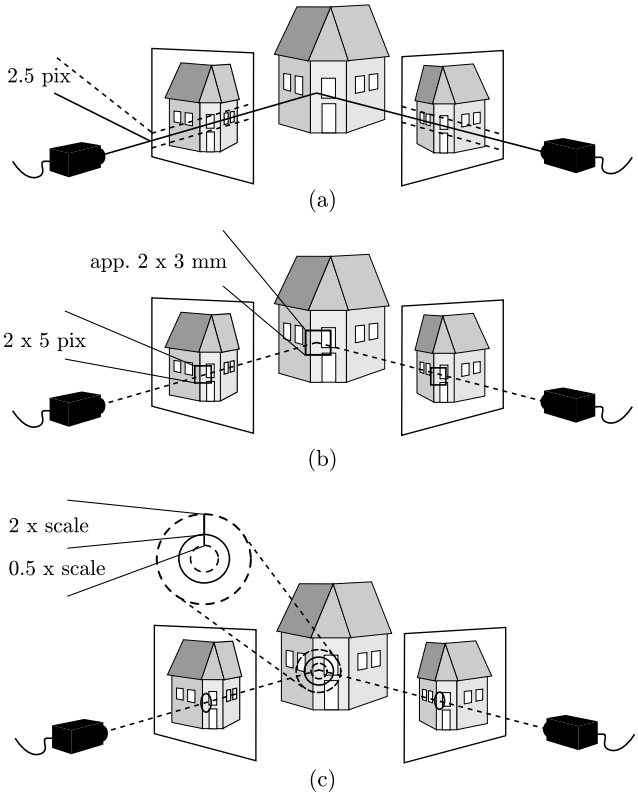


Fig. 7 Matching criteria for interest points. This figure gives a schematic illustration of a scene of a house and two images of the scene from two viewpoints. (**a**) The consistency with epipolar geometry, where corresponding descriptors should be within 2.5 pixels from the epipolar line. (**b**) Window 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

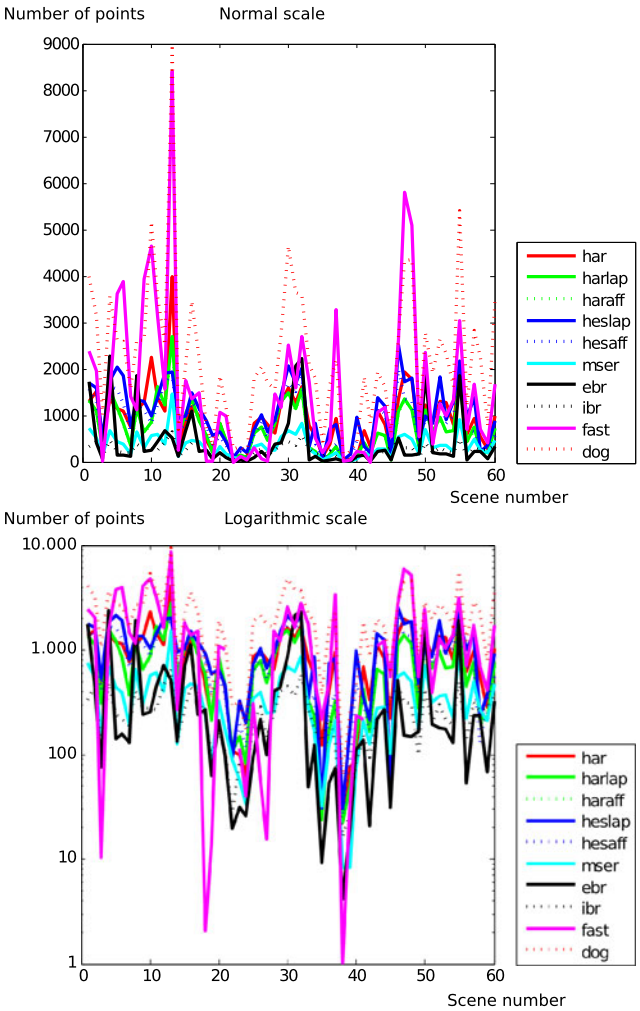


Fig. 8 Number of points in each scene for the different detector types. The horizontal axis shows the scene number and the vertical axis shows the number of points on (**a**) normal scale and (**b**) logarithmic scale. The high count outliers are especially clear in (**a**) and the low count outliers 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

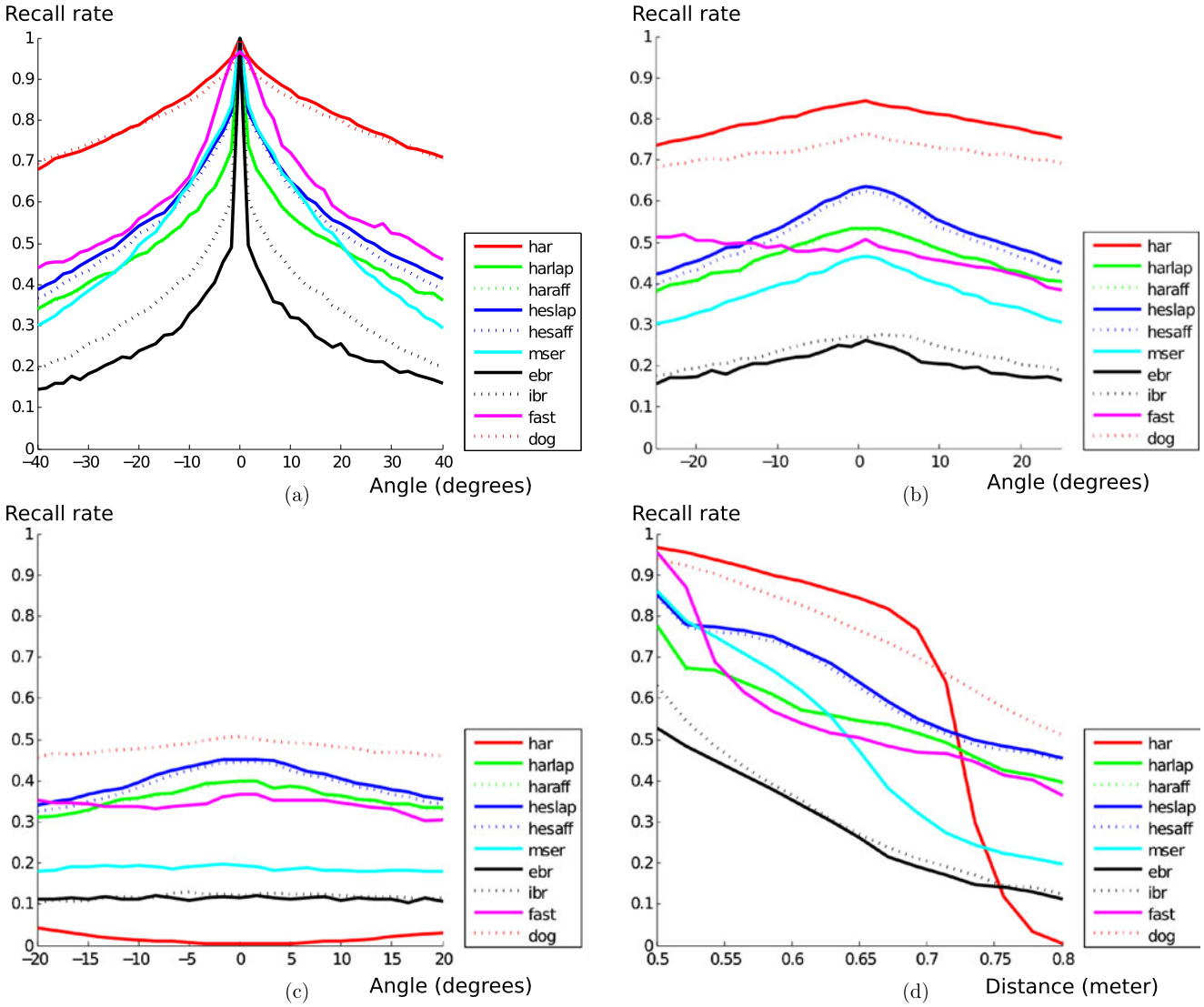


Fig. 9 Mean recall rate. The graphs show the recall rate relative to the paths shown in Fig. 4 with (**a**) Arc 1, (**b**) Arc 2, (**c**) Arc 3, and (**d**) Line Path. The *horizontal axis* is the angle relative to the scene in (**a**)–(**c**) and distance to the scene in (**d**). The *vertical axis* 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

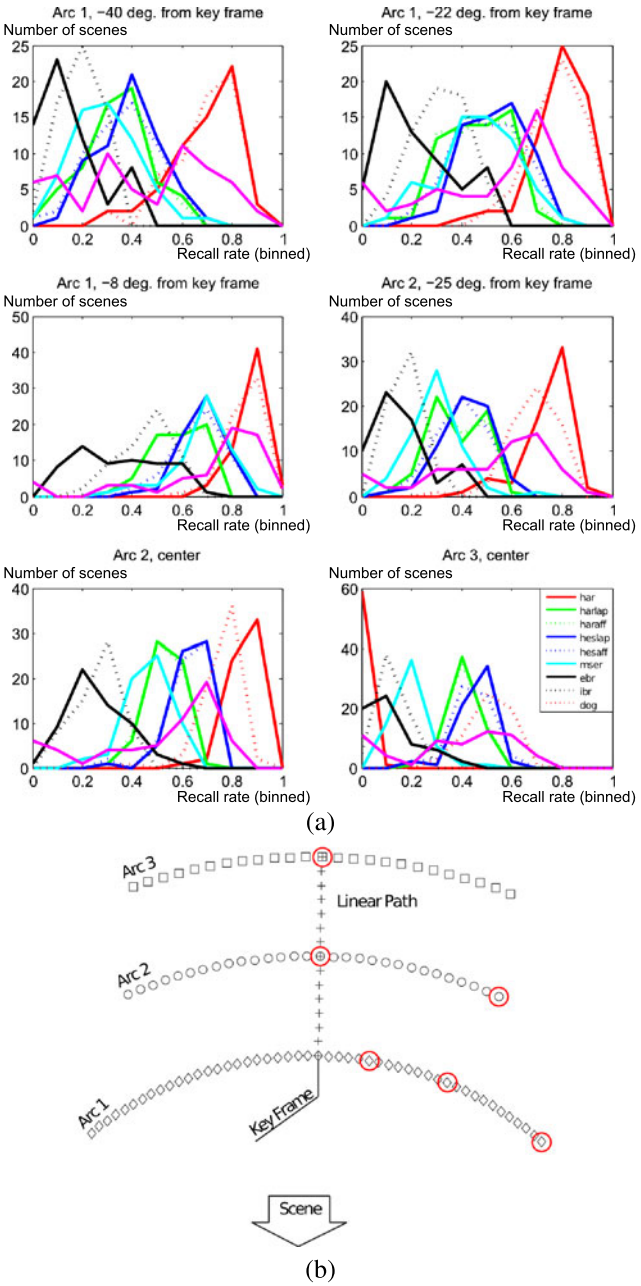


Fig. 10 (**a**) Probability distribution functions for selected image positions and (**b**) their positions on the path. The *colors* show the detector types—see Fig. 9. The *horizontal axis* shows binned recall rates and the *vertical axis* show number of scenes. This figure provides more detail in 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;

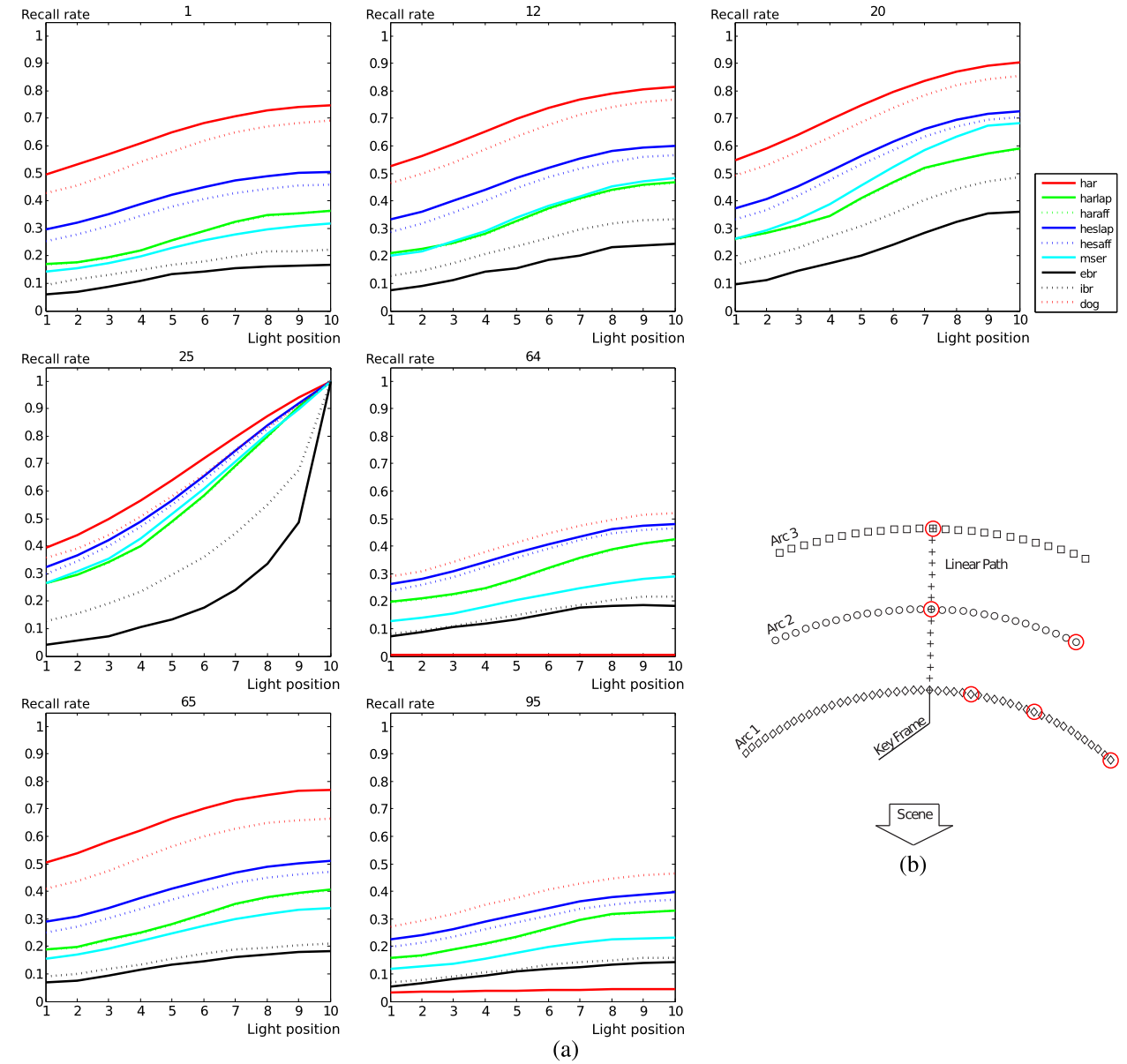
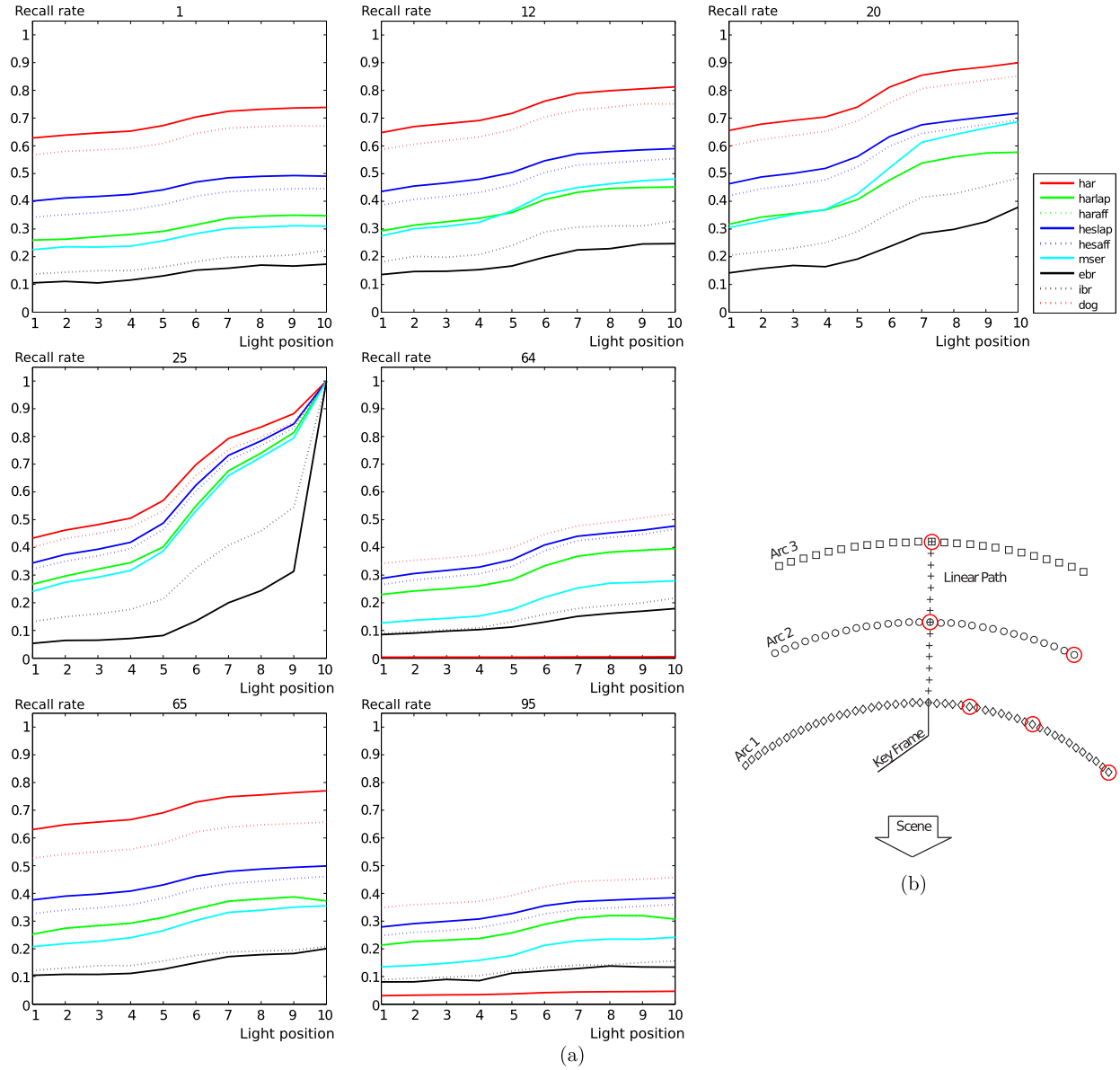
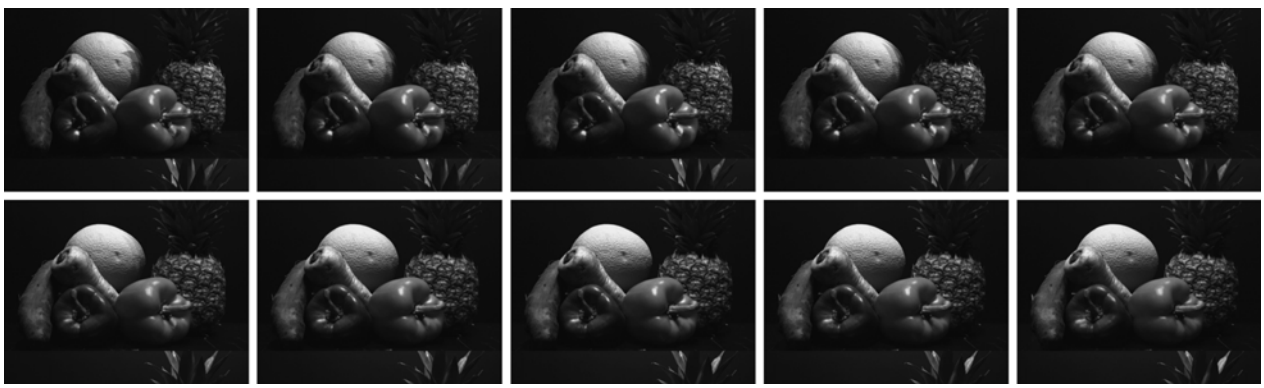


Fig. 11 Mean recall rate relative to change in light direction from back to front for seven camera positions averaged over all 60 scenes. The graphs (**a**) show the performance of the different detector types, with the average recall rate at the *vertical axis* and the light direction at the *horizontal 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,

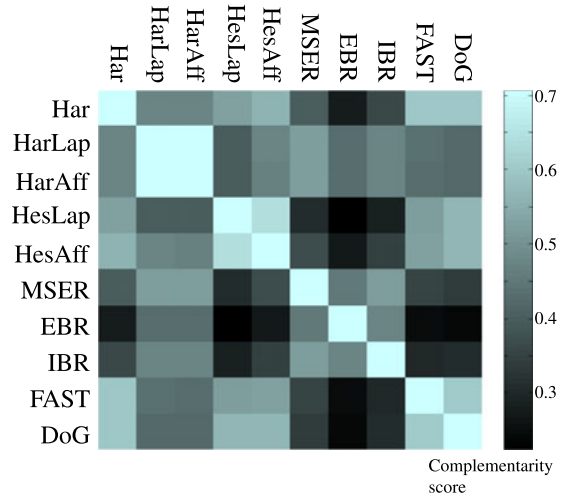


Fig. 13 The complementarity of the interest point detectors. The score is computed as described by (2) and averaged over all 60 scenes. Dark values imply that the two sets of interest points cover different parts of the scene and this way complement each other

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).

Letter to reviewer

This manuscript is an extension of our previous paper *Finding the Best Feature Detector-Descriptor Combination* (Dahl et al, 2011), published at International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011. The dataset used is presented in our IJCV paper (Aanæs et al, 2012). Both papers are included as supplementary material. As stated in the submitted manuscript, the conference publication is extended in this work by

- The number of interest point descriptors and detectors have been increased, e.g. including color invariant descriptors.
- We have proposed a new and superior performing version of the multi-scale Harris corner detector.
- We have included light variation in our tests.
- The analysis and statistical evaluation has been considerably improved.

We have chosen to submit a manuscript with a somewhat above average length. This is due to the fact that the tables and figures needed to present our considerable experimental results in sufficient detail takes up a lot of space. We have also prioritized to present our version of the multi-scale Harris corner detector in some detail, since our findings are that the details of this detector is crucially important, as such these details need to be described for our results to be reproducible.

References

Aanæs H, Dahl AL, Pedersen KS (2012) Interesting interest points - a comparative study of interest point performance on a unique data set. International Journal of Computer Vision 97(1):18–35

Dahl A, Aanæs H, Pedersen K (2011) Finding the best feature detector-descriptor combination. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011, pp 318–325

Finding the Best Feature Detector-Descriptor Combination

Anders Lindbjerg Dahl, Henrik Aanæs
DTU Informatics
Technical University of Denmark
Lyngby, Denmark
abd@imm.dtu.dk, haa@imm.dtu.dk

Kim Steenstrup Pedersen
Department of Computer Science, DIKU
University of Copenhagen
Copenhagen, Denmark
kimstp@diku.dk

***Abstract*—Addressing the image correspondence problem by feature matching is a central part of computer vision and 3D inference from images. Consequently, there is a substantial amount of work on evaluating feature detection and feature description methodology. However, the performance of the feature matching is an interplay of both detector and descriptor methodology. Our main contribution is to evaluate the performance of some of the most popular descriptor and detector combinations on the DTU Robot dataset, which is a very large dataset with massive amounts of systematic data aimed at two view matching. The size of the dataset implies that we can also reasonably make deductions about the statistical significance of our results. We conclude, that the MSER and Difference of Gaussian (DoG) detectors with a SIFT or DAISY descriptor are the top performers. This performance is, however, not statistically significantly better than some other methods. As a byproduct of this investigation, we have also tested various DAISY type descriptors, and found that the difference among their performance is statistically insignificant using this dataset. Furthermore, we have not been able to produce results collaborating that using affine invariant feature detectors carries a statistical significant advantage on general scene types.**

***Keywords*-Interest point detector, Interest point descriptor, Feature evaluation, Combined descriptor detector evaluation**

I. Introduction

The computational efficiency of a sparse image representation consisting of salient interest points, also referred to as features, is a major motivation for feature based methods for solving the image correspondence problem. Various detectors and descriptors have been proposed, but the question of how to optimally design an interest point characterization still remains open. The success of feature-based methods depends on the quality of the local characterization. In general it is not an easy task to judge the performance of such methods, because it is hard to validate if correspondence exist. However given knowledge about the geometry of the observed scene, it becomes easy to verify if two interest points corresponding in feature space also corresponds in the real scene. We therefore propose to use the DTU Robot dataset with known surface geometry presented in [1], [2] (see Sec. II for a brief description and Fig. 1). Based on this dataset we are able to systematically analyze the design of feature methods and due to the large variation in scene types we can judge the statistical significance of our findings.

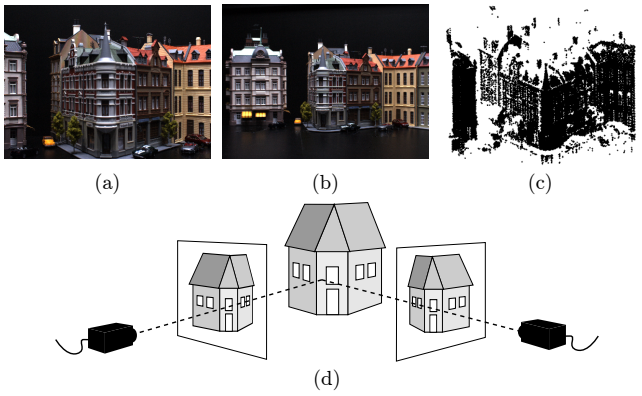


Figure 1. Example of data and setup. Two images of the same scene with one close up (a) and one distant from the side (b), and the reconstructed 3D points (c). Corresponding interest points can be found using the geometric information of the scene with known camera positions and 3D scene surface as schematically illustrated in (d). Illustration from [1].

Finding correspondence between image pairs using interest points is based on the assumption that common interest points will be detected in both images. For this to be useful, 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. Commonly, candidate points are detected using an interest point *detector* and a description of the local image structure – the so-called *descriptors* – surrounding the interest points are extracted. Following the extraction of descriptors, a comparison of these is made using a relevant similarity metric in order to determine correspondence between interest points. The rationale is that descriptors capture the essential visual appearance of the scene region covered by the interest point, and as a consequence the same scene point seen from different viewpoints and/or with different lighting should have similar descriptors. Therefore descriptors should preferably be invariant, or approximately, with respect to changes in viewpoint and lighting.

Early work on correspondence from local image features was based on rotation and scale invariant features [3], [4],

and interest points from planer scenes was evaluated in [5]. Later the interest points have been adapted to affine transformation, to obtain robust characterization to larger viewpoint changes. These methods have been surveyed in [6], but the performance has been evaluated on quite limited datasets consisting of ten scenes each containing six images. The suggested evaluation criteria have since been used in numerous works together with this small dataset.

Different approaches have been taken when describing the local visual appearance of interest points. A majority of approaches extract some descriptive feature, such as histograms of differential geometric image properties in each pixel [3], [7], [4], [8], using integral images [9], [10], or the responses of steerable filters [11], differential invariants or local jets [12], [13], [14], [5]. The SIFT [3], GLOH [7], and DAISY [8], [15], [16], [17] descriptors also includes a spatial pooling step in order to agglomerate the descriptive feature in an arrangement around the interest point. A selection of descriptors have previously been evaluated in [7] on the same dataset as used in [6]. Again the limitations of the dataset restricts the ability to generalize the results from this survey to a wider class of scene types and more natural variation in illumination.

The ground truth in the data from [6] was obtained by an image homography. This limits the scene geometry to planar surfaces or scenes viewed from a large distance where a homography is a good approximation. Fraundorfer and Bishof [18] addressed this limitation by generating ground truth and requiring that a matched feature should be consistent with the camera geometry across three views. In Winder *et al.* [17], [19], [16], [20] results from Photo Tourism [21] were used as ground truth.

Moreels and Perona [22] evaluated feature descriptors similar to [18] based on pure geometry by requiring three view geometric consistency with the epipolar geometry. In addition they used a depth constraint based on knowledge about the position of their experimental setup. Hereby they obtained unique correspondence between 500-1000 detected points from each object. The limitation of their experiment is the use of relatively simple scenes with mostly single objects resulting in little self-occlusion. However, self-occlusions are very frequent in real world scenes and many interest points are typically found near occluding boundaries, limiting the applicability of their conclusions.

The aim of this work is to compare pairs of feature detectors and descriptors, to find the best combination. To keep the computational burden manageable the number of candidates have to be limited, and we thus only use candidates which have previously been reported to perform well. As for the detectors we choose Harris, Harris Affine, Harris Laplace, Hessian Laplace, Hessian Affine, MSER, and Difference of Gaussian (DoG), because they are popular and reported to work well in the literature [1], [23].

As for the feature descriptors, the state of the art is currently the SIFT [3] and DAISY descriptors [8], [15], [16], [17] which we choose to use and implement using the framework of Winder and Brown [17]. We also include conventional (normalized) cross correlation as a baseline, since much work has been done using this descriptor. The DAISY descriptors however cover a wide range of descriptors; as such we choose to divide our analysis into two, where we first identify the best DAISY descriptors on a subset of the detectors. This is the subject of Sec. III, where 21 different variants of the DAISY descriptor are evaluated. Each combination is evaluated using ROC-curves (Receiver Operating Characteristics). Two representative descriptors are carried on to the last part of the analysis, reported in Sec. IV, where a matrix of the seven detectors and four descriptors are evaluated. A discussion of our results and recommendations is found in Sec. V.

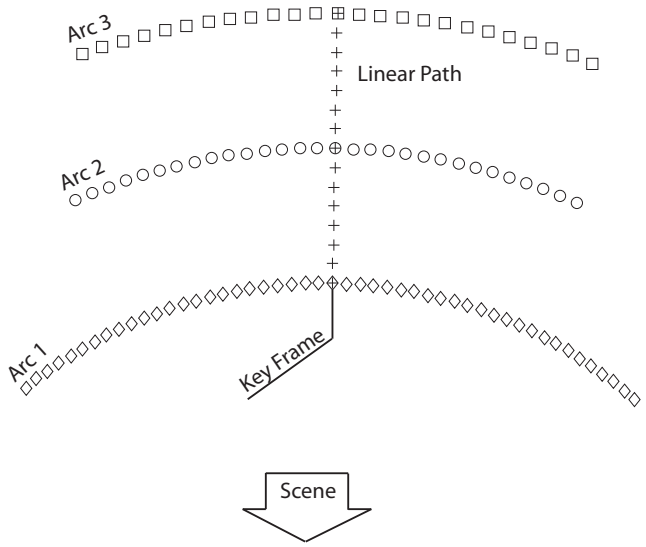


Figure 2. The central frame in the nearest arc is the key frame, and the surface reconstruction is attempted to cover most of this frame. The three arcs are located on circular paths with radii of 0.5 m, 0.65 m and 0.8 m, which also defines the range of the linear path. Furthermore, Arc1 spans $+/-40^\circ$, Arc2 $+/-25^\circ$ and Arc3 $+/-20^\circ$. Illustration from [1].

II. Data and Evaluation

In this investigation we use the DTU Robot dataset [1], [2][1] illustrated in Fig. 1. This dataset is constructed under controlled settings using an industrial robot. The set consists of 60 complex scenes, and Fig. 2 shows how each scene is viewed from 119 positions with known camera geometry. The dataset also incorporates light variation, but in this work we only focus on diffuse lighting. In addition the 60 scenes have been surface scanned using structured light. Together with the camera geometry this allows us to accurately determine the correct camera correspondences

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

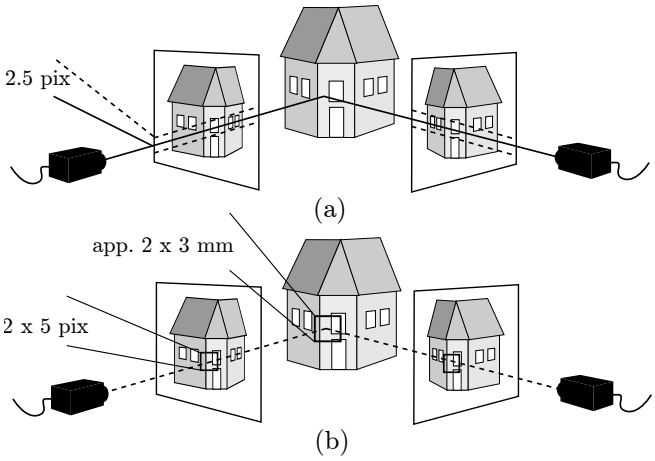


Figure 3. Matching criteria for interest points. This figure gives a schematic illustration of a scene of a house and two images of the scene from two viewpoints. (a) The consistency with epipolar geometry, where corresponding descriptors should be within 2.5 pixels from the epipolar line. (b) Window 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. Illustration from [1].

without matching visual features. In real outdoor scenes, as presented in [16], there is no alternative to have ground truth based on feature matching, but this could likely bias the result.

A. Evaluation criteria

The evaluation framework used is similar to the one reported in [1], which only includes an evaluation of the matching performance of different detector methods on the DTU Robot dataset. We want to determine if a pair of corresponding features are correct or not, where correspondence is found by the Euclidean distance between feature descriptors. Fig. 2 illustrates how the features are matched between one key frame and all other images. Fig. 3 shows the two criteria that we use for determining correct correspondence. Correct matches have to be within 2.5 pixels of the epipolar line *and* the corresponding 3D point must be within a 5 pixel error margin corresponding to approximately 3 mm.

Given an image pair, where one image is the key frame, a detector-descriptor pair is evaluated by

1) For each feature in the key frame find the distance to the best δ_b and the *second* best δ_s matching feature in the other image.
2) For each feature correspondence compute the ratio, $r = \frac{\delta_b}{\delta_s}$, between the match score of the second best and the best correspondence. It is also determined if the best match is correct or not.
3) Using this ratio, r, as a predictor for correct matches, c.f. [3], the ROC (Receiver Operating Characteristic) curve, as a function of r, is constructed based on all features in an image pair. We compare the area under the ROC curve (AUC). The area is between zero and one, where one indicates perfect performance of the detector-descriptor pair.
4) The AUC is used as the performance measure of a detector-descriptor combination on a pair of images.

These AUCs are the basis for our statistical analysis. The AUC is chosen as a performance measure, in line with [16], because it elegantly removes the need to balance between many false positive or many false negatives. As a result it strongly relates to the underlying discriminative power of the method.

We compare different detector-descriptor methods by computing the mean performance, i.e. the mean AUC over the 60 sets for each position, c.f. Fig. 6, 7 and Tab. II. Based on the central limit theorem, we assume these means to be normal distributed. We compare the means using students t-test

$$\frac{\mu_1 - \mu_2}{\hat{\sigma}} \quad , \tag{1}$$

where μ_1 and μ_2 are the two means to be compared and $\hat{\sigma}$ is an estimate of the standard deviation. When computing an estimate of the variances, $\hat{\sigma}^2$, we perform an analysis of variance, assuming that for a given method and a given problem, performance is given by two factors

$$\text{Performance} = \text{Problem Difficulty} + \text{Method} + \text{Noise} \ .$$

Since we are interested in comparing the methods the variance due to the Problem Difficulty is factored out, which reduces the overall variance, $\hat{\sigma}^2$, making it easier for a difference in means to be significant.

B. Implementation

All feature detectors are computed by implementations provided by the authors of [3], [24], [25][2], whereas we implemented our own interest point descriptors. They are estimated on an affine warped image patch sampled according to the parameters obtained from the interest point detection and rotated to one dominant gradient direction. The image patch is sampled with a radius of three times the scale of the feature point and we discard points that exceed the image borders. We found this to be a good tradeoff between performance and number of discarded sample points. In the experiments described in Sec. III we use a patch size of 66×66 pixels whereas the patches in the experiments in Sec. IV are 30×30. This is especially a consequence of the pixel similarity estimates where we have feature vectors of 900 dimensions. Using the 66×66 pixel patches this would be 4356 dimensions, which approximately slows the calculation down with a factor four. We only observed a minor loss in precision, which is shown in Tab. I "spatial layout – 1-8-8"

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

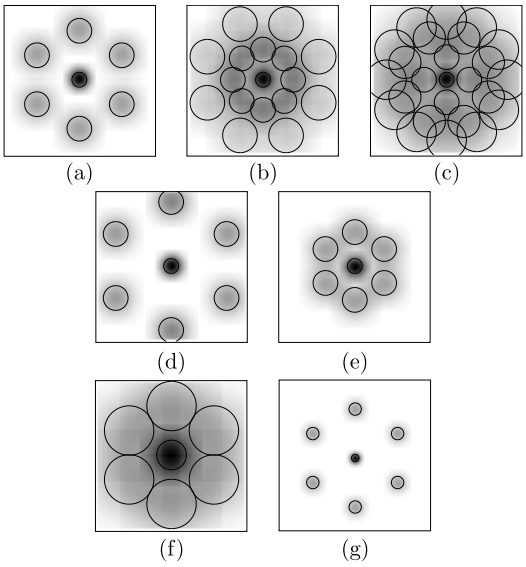


Figure 4. Layout of the descriptors for spatial summation. The circles mark the size of the sample points and the dark color shows the Gaussian weighing. *First row* one ring with six samples – (1-6) (a), two rings with eight samples in each – (1-8-8) (b), three rings with four, eight and twelve samples – (1-4-8-12) (c). *Second row* one ring with six samples – large footprint – (1-6 lf) (d), small footprint – (1-6 sf) (e). One ring with six samples – large sample area – (1-6 lg) (f), small sample area – (1-6 sg) (g).

should be compared to "HesAff" and "HarAff" – "DAISY-I" and "DAISY-II" in Tab. II. It shows a performance loss of 0.013 caused by reduction in patch size.

Our implementation of the DAISY descriptor closely follows the description of Winder *et al.* [16][3]. To ensure that the only difference between the DAISY and SIFT descriptors were the sampling, we chose to implement our own SIFT descriptor. To validate the performance we did a small experiment to compare to the original implementation of Lowe [3][4], and we obtained similar performance with patches of 66×66 pixels and about 5% fewer matching descriptors with the 30×30 patches.

III. Comparing DAISY descriptors

Brown *et al.*[20] presents a framework for optimizing feature descriptors. They have chosen the DAISY-type descriptor presented in Winder and Brown [17], because it is easily reconfigurable. The optimization is based on three outdoor scenes where ground truth is obtained from the bundler software [21], which is based on the SIFT framework [3]. In this experiment we have performed a similar investigation to Brown *et al.*, but based on the extended DTU Robot dataset, where ground truth geometry is based on precise calibration and structured light scanning. In order to keep the computational burden manageable, we only did this experiment on the Harris affine and Harris Laplace features.

[3]http://cvlab.epfl.ch/~brown/patchdata/patchdata.html

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

Comparison	Type	Performance
Descriptor type	Type 1	0.781
	Type 2	0.785
	Type 3	0.791
Spatial layout	1-6	0.786
	1-8-8	0.804
	1-4-8-12	0.802
	1-6 lf	0.784
	1-6 sf	0.778
	1-6 lg	0.784
	1-6 sg	0.763
Descriptor dimensionality	Small	0.783
	Large	0.788
Scene types	Houses	0.751
	Books	0.791
	Fabric	0.831
	Greens	0.799
	Beer cans	0.696
Affine vs. Laplace	Laplacian	0.783
	Affine	0.788

Table I
Mean AUC for different groupings of the descriptor types. The table shows mean value of all positions. In Fig. 4 the spatial layout is shown.

The descriptors proposed by Brown *et al.*[20] are varied in the spatial layout and differential-geometric response. The spatial layout that we have tested are illustrated in Fig. 4. We have varied the number of sample points, the size of sample points and their relative distance. We employ three differential-geometric responses – the directional binned gradients in four and eight directions (Type 1), average positive and negative gradients (Type 2) and steerable filters (Type 3). The experimental result is summarized in Tab. I. This approach closely follows Winder *et al.*[16].

The results show that the effect of changing the differential-geometric response is limited, so it is clearly an advantage to select either Type 1 or 2, because the computational cost of these descriptors is much lower. There is a small advantage in selecting a spatial layout where two rings are sampled, but three ring sampling does not give an improvement. Fig. 5 shows that this advantage is seen for all positions. The dimensionality difference arise from number of sample directions in Type 1, combinations of positive and negative gradients in Type 2, and number of directions of the steerable filters in Type 3. But there is almost no difference in selecting the large dimensionality over the small. There is a clear difference in Scene type, where the AUC is significantly higher for less specular objects like fabric than for specular objects like beer cans. The difference between affine and non-affine feature detectors is surprisingly small, which might be a result of complexity of the evaluated scenes with many occluding boundaries. The findings regarding affine detectors are confirmed by the experiments presented in Sec. IV.

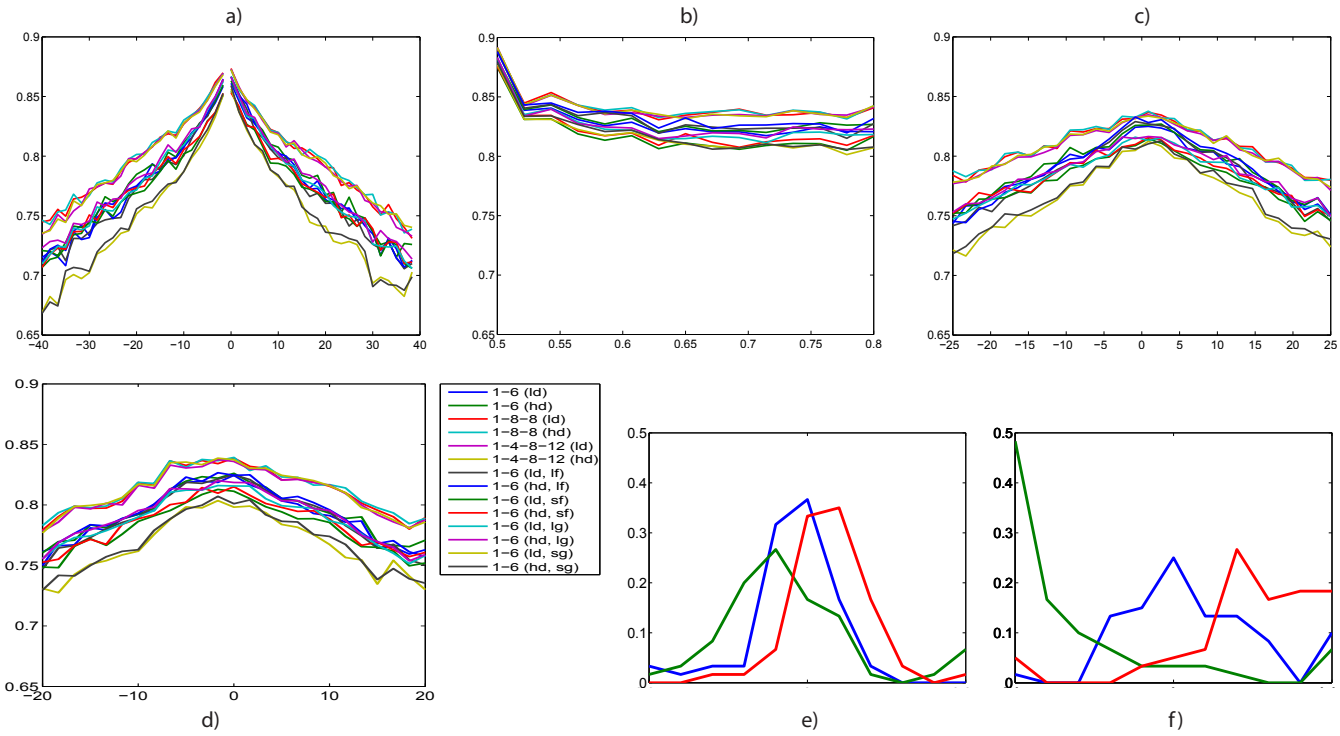


Figure 5. Performance evaluation of the DAISY descriptor. Average AUCs for Type 2 descriptors are shown in (a - d). The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). The labels relate to the descriptor design shown in Fig. 4. In (e - f) probability density functions for different descriptor designs are shown for a 30° angle where (e) is affine interest points and (f) is non-affine. This shows that with a sparse sampling the performance goes down for the non-affine, but the affine invariance can be compensated by a dense sampling.

	Corr	SIFT	DAISY-I	DAISY-II	Avg.
Har	0.615	0.767	0.729	0.741	0.713
HarAff	0.629	0.818	0.791	0.798	0.759
HarLap	0.635	0.814	0.784	0.790	0.756
HesAff	0.636	0.795	0.773	0.779	0.746
HesLap	0.630	0.757	0.740	0.742	0.717
MSER	0.648	**0.846**	0.826	0.832	0.788
DOG	0.646	**0.849**	0.837	**0.843**	0.794
Avg.	0.634	0.807	0.783	0.789	0.753

Table II
MEAN AUC OVER ALL POSITIONS FOR THE FEATURE DETECTOR AND DESCRIPTOR COMBINATIONS. TOP 3 PERFORMERS HIGHLIGHTED WITH BOLD-FACE (**HAR** IS HARRIS CORNERS, **HARAFF** IS HARRIS AFFINE, **HARLAP** IS HARRIS LAPLACE, **HESAFF** IS HESSIAN AFFINE AND **HESLAP** IS HESSIAN LAPLACE FEATURE DETECTORS RESPECTIVELY).

From this study, we choose the two-ring DAISY descriptor with small (DAISY-I) and large dimensionality (DAISY-II) for further analysis. This is done together with SIFT and a vector of simple pixel intensities (normalized cross correlation). These four descriptors are analyzed in combination with seven feature detectors.

IV. COMPARING DETECTOR-DESCRIPTOR COMBINATIONS

In this section we present the evaluation of detector-descriptor combinations with the aim of finding the best performers. We compare a combination of the four feature descriptors (SIFT, DAISY-I, DAISY-II and cross correlation) with seven feature detectors. These detectors are Harris corner detector [26], Harris Laplace, Harris affine, Hessian Laplace, Hessian affine [6], MSER [24], and Difference of Gaussians (DoG) [3]. The combined result is summarized in Tab. II and Fig. 6

To evaluate the significance of the performance difference we have estimated the average standard deviation $\hat{\sigma}$ of (1). Overall we obtain $\hat{\sigma} = 0.08$, but if we exclude cross correlation, which has a higher variance than all others, then we obtain $\hat{\sigma} = 0.05$. To give an idea of significance based on Student's t-test from (1) we consider a difference larger than 0.05 as significantly different on a 84% confidence level and 0.1 as significant on a 98% level.

The performance is computed for all 28 combinations on all 119 camera positions, where the distribution of the performance was evaluated over all 60 scenes. Our central evaluation criterion is the mean over these 60 scenes for a given position and detector-descriptor combination. Due to

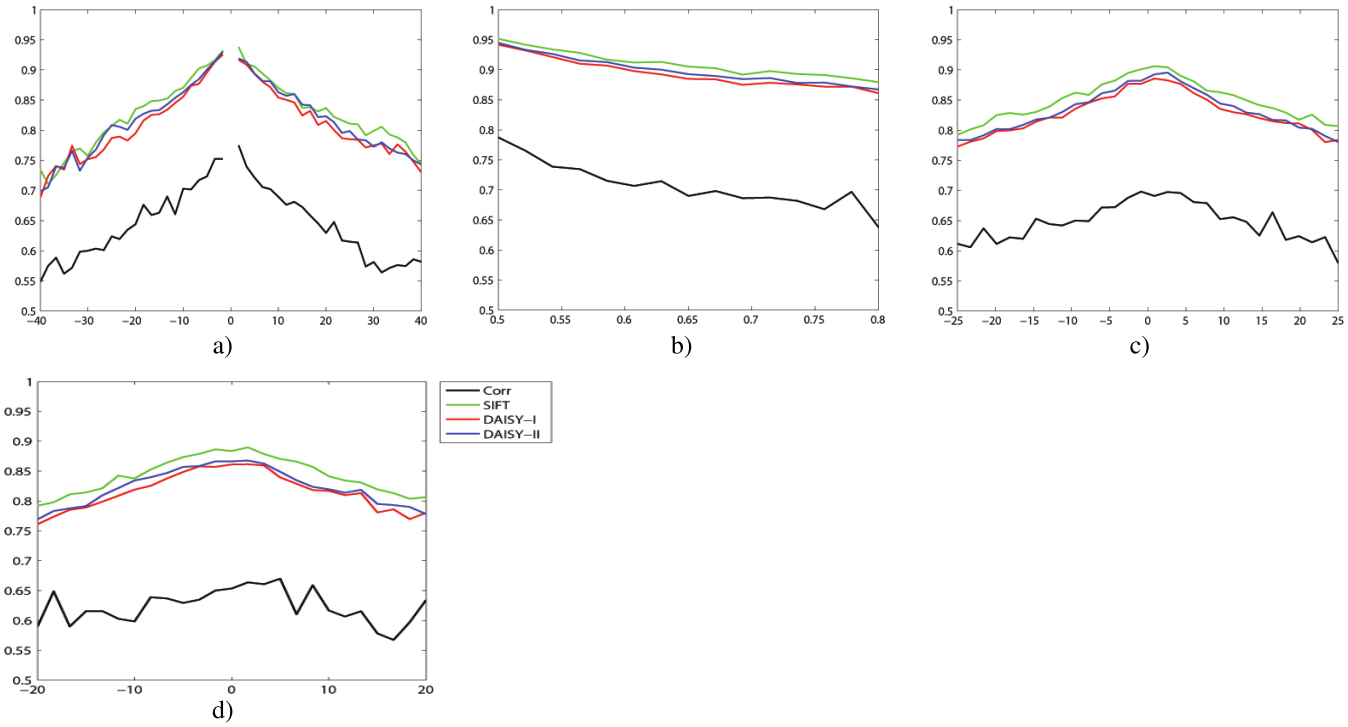


Figure 6. Mean AUC for the MSER detector displayed for all four descriptors and for all positions. The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). It is seen that the SIFT and the two DAISY descriptors have very similar performance, compared to a $\hat{\sigma} = 0.05$, but outperform the correlation.

space limitations we are only able to present a summarized evaluation as shown in Fig. 6 and Tab. II outlining our conclusions.

Fig. 6 shows a combination with the same detector but different descriptors. Cross correlation is clearly outperformed by the other descriptors. SIFT and DAISY has almost identical performance, and Tab. II shows that their average difference is less than 0.015, which is statistically insignificant.

In Fig. 7 the SIFT descriptor is shown in combination with the seven detectors. We chose to show SIFT, but very similar results were obtained for the DAISY descriptors. Here there is a difference in performance where MSER and DOG detectors perform about one standard deviation better than the Harris affine and Harris Laplace detectors, and about two standard deviations better than the Harris based detectors, which is statistically significant. Harris corner detector with no scale adaption performs well when the scale change is not to large.

So, our experiments suggest that the best choice is a DOG or MSER detector with a SIFT or DAISY descriptor, or a perhaps a Harris corner detector if the scale change is low. The dataset used also has different categories of scene types like 'fabric', 'books', 'model houses', etc. and running the experiments on a specific scene type did not change the overall picture. Compared to the results in [1], where the recall rate of detectors was evaluated on the same dataset, it is interesting to see that the best performers in a full feature tracking frame work are not identical to the ones with the best recall rates. Again this implies that the discriminative power of the extracted features vary for different feature detectors. This last point is especially noteworthy for the MSER detectors. Both the descriptor experiment presented in Sec. III and this combined experiment show that an affine detector has an advantage, but this advantage is small compared to variance making it statistically insignificant, see Fig. 8.

V. Discussion

Based on the experiments reported in this paper the general conclusion is that the best detector-descriptor combination is either the DOG or MSER detectors and SIFT or DAISY descriptors. If the scale change is low a Harris corner detector would be superior and also faster and simpler to run and implement. The experiments also show, that many other performance differences exist, which confirm other studies, but these differences are not statistically significant. This demonstrates a need for considering statistical significance when performing these type of comparisons, necessitating the use of large datasets to make meaningful estimates of significance and variance, such as the dataset used here [1].

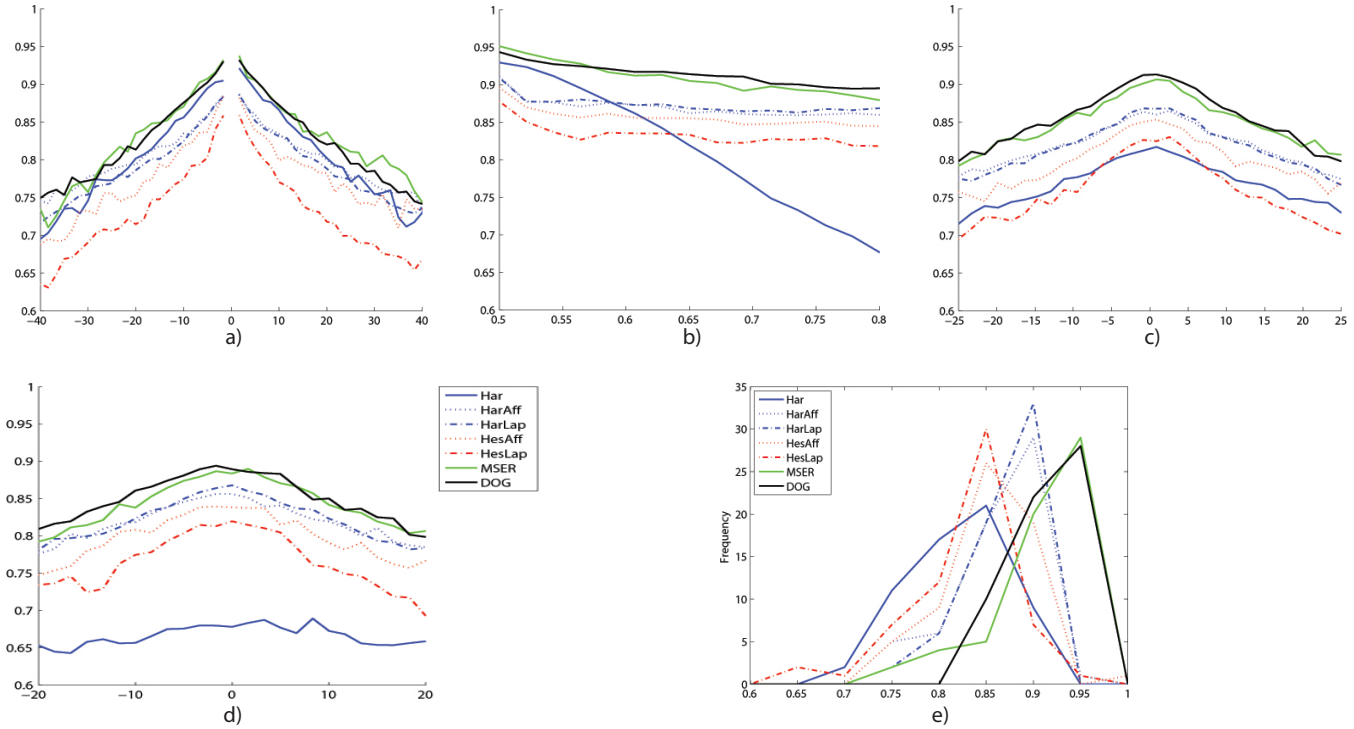


Figure 7. Mean AUC for the SIFT descriptor displayed for all seven detectors and all positions. The vertical axis in the graphs show the AUC, and the horizontal is the angle (a,c,d) and distance (b) relative to the key-frame. Each graph corresponds to the sample path shown in Fig. 2, with Arc 1 (a), Linear Path (b), Arc 2 (c) and Arc 3 (d). Here it is seen that the MSER and DOG detectors are the top performers, outperforming the Harris based detectors on a statistically borderline level, and significantly outperforming the hessian based descriptors. The performance of the 'pure' Harris corner detector is very scale dependent. Similar results are obtained for the two DAISY descriptors, as indicated in Fig. 6. The validity of our findings is further cooperated by considered the probability distribution functions for each position, in (e) the pdf is shown for 0.86° of Arc 2.

Furthermore, it is interesting to note that the DOG detectors perform much better than the Hessian type detectors, although they are very similar, i.e. the DOG is basically a well-engineered approximation of the Laplace filter, which is equal to the trace of the Hessian. This indicates that perhaps a better-engineered version of the Harris Laplace corner detector, inspired by the DOG detector, could be made. This is especially interesting in the light that the Harris corners performed better than the Hessians.

A last point of curiosity is that we have not been able to produce results collaborating that using affine invariant feature detectors carries a statistical significant advantage on general scene types. However this type of invariance may have merit in e.g. 3D reconstruction of urban type scenes or other near-planar scenes.

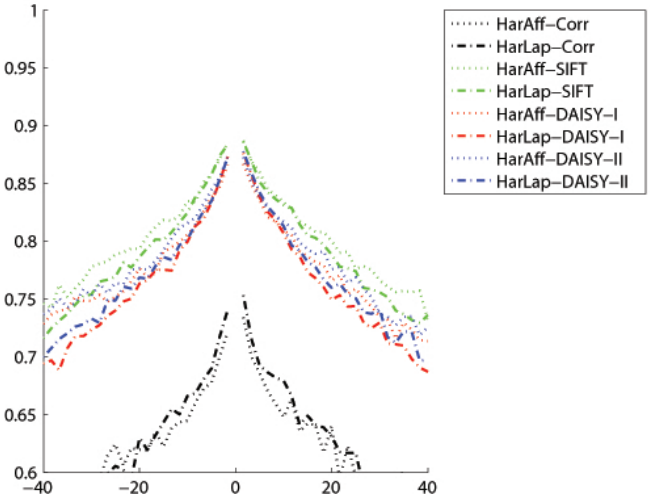


Figure 8. Affine vs. non-affine (Lap). Affine performs slightly better with large angles, but the improvement is not significant.

References

[1] H. Aanæs, A. L. Dahl, and K. S. Pedersen, "On recall rate of interest point detectors," in *3DPVT*, 2010. [Online]. Available: http://campwww.informatik.tu-muenchen.de/3DPVT2010/data/media/e-proceeding/session07.html#paper97

[2] H. Aanæs, A. L. Dahl, and V. Perfernov, "Technical report on two view ground truth image data," DTU Informatics, Technical University of Denmark, Tech. Rep., 2009.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[4] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE T-PAMI*, vol. 19, no. 5, pp. 530–535,

1997.

[5] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *IJCV*, vol. 37, no. 4, pp. 151–172, 2000.

[6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1-2, pp. 43–72, 2005.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE T-PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[8] E. Tola, V. Lepetit, and P. Fua, "A Fast Local Descriptor for Dense Matching," in *CVPR*, 2008.

[9] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *ECCV*, pp. 404–417, 2006.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] W. Freeman and E. Adelson, "The design and use of steerable filters," *IEEE T-PAMI*, vol. 13, no. 9, pp. 891–906, 1991.

[12] E. Balmashnova and L. Florack, "Novel similarity measures for differential invariant descriptors for generic object retrieval," *JMIV*, vol. 31, no. 2-3, pp. 121–132, 2008.

[13] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever, "Cartesian differential invariants in scale-space," *JMIV*, vol. 3, no. 4, pp. 327–348, 1993.

[14] J. J. Koenderink and A. J. van Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, pp. 367–375, 1987.

[15] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE T-PAMI*, vol. 32, no. 5, pp. 815–830, May 2009.

[16] S. Winder, G. Hua, and M. Brown, "Picking the best daisy," in *CVPR*, 2009.

[17] S. A. J. Winder and M. Brown, "Learning local image descriptors," in *CVPR*, 2007, pp. 1–8.

[18] F. Fraundorfer and H. Bischof, "Evaluation of local detectors on non-planar scenes," in *Proc. 28th workshop of AAPR*, 2004, pp. 125–132.

[19] G. Hua, M. Brown, and S. Winder, "Discriminant embedding for local image descriptors," *ICCV*, pp. 1–8, 2007.

[20] M. Brown, G. Hua, and S. Winder, "Discriminative Learning of Local Image Descriptors," *IEEE T-PAMI*, 2010.

[21] N. Snavely, S. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *IJCV*, vol. 80, no. 2, pp. 189–210, 2008.

[22] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," *IJCV*, vol. 73, no. 3, pp. 263–284, 2007.

[23] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[24] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[25] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

[26] C. Harris and M. Stephens, "A combined corner and edge detector," in *4th Alvey Vision Conf.*, 1988, pp. 147–151.