Reliable Point Correspondences in Scenes Dominated by Highly Reflective and Largely Homogeneous Surfaces

Srimal Jayawardena[1], Stephen Gould[2], Hongdong Li[2], Marcus Hutter[2], and Richard Hartley[2]

[1] Autonomous Systems Laboratory, CSIRO, Australia.
srimal.jayawardena@csiro.au
[2] Research School of Computer Science, The ANU, Australia.

Abstract. Common Structure from Motion (SfM) tasks require reliable point correspondences in images taken from different views to subsequently estimate model parameters which describe the 3D scene geometry. For example when estimating the fundamental matrix from point correspondences using RANSAC. The amount of noise in the point correspondences drastically affect the estimation algorithm and the number of iterations needed for convergence grows exponentially with the level of noise. In scenes dominated by highly reflective and largely homogeneous surfaces such as vehicle panels and buildings with a lot of glass, existing approaches give a very high proportion of spurious point correspondences. As a result the number of iterations required for subsequent model estimation algorithms become intractable. We propose a novel method that uses descriptors evaluated along points in image edges to obtain a sufficiently high proportion of correct point correspondences. We show experimentally that our method gives better results in recovering the epipolar geometry in scenes dominated by highly reflective and homogeneous surfaces compared to common baseline methods on stereo images taken from considerably wide baselines.

1 Introduction

Structure from Motion (SfM) tasks that recover geometric scene information from a set of images obtained from different views typically require reliable point correspondences across the images (or tracks in the case of videos) as a prerequisite. Such SfM tasks range from complete 3D scene reconstruction to stereo matching performed on uncalibrated images. Typically keypoints in images are detected and matched in order to obtain point correspondences. Much research has been done in this area and popular applications which use feature correspondences include aligning tourist photos from the Internet [1].

Motivation. Scenes dominated by highly reflective and largely homogeneous surfaces such as the body of a car [3, 4, 5], buildings with a lot of glass panes (*e.g.* failure case of [6] in Figure 7 (g) and (i)), medical images [7, 8, 9] *etc.* tend to generate unreliable point correspondences. The number of iterations required

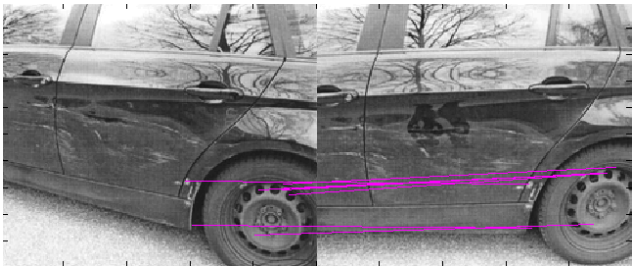


Fig. 1. Best point correspondences obtained from naive SIFT [2] matching do not give a sufficient spatial spread to recover the epipolar geometry. The reliable matches are concentrated around relatively non-reflective areas. Best viewed in color. Images may be cropped for clarity.

for convergence of subsequent model fitting algorithms such as estimating the fundamental matrix using RANSAC increase exponentially and the task becomes intractable as the level of noise in the point correspondence grows. Typically noise ratios of more than 50% tend to be impractical [10]. The high amount of noise in point correspondences obtained from scenes dominated by highly reflective and largely homogeneous surfaces can be due to the following reasons.

1. Reflections in common reflective surfaces do not represent a physical artifact in 3D space. Therefore in general they do not conform to the EPG (Epipolar Geometry) of the 3D scene. A special case is for rectilinear camera motion [11] where the epipolar deviations of specularities on surfaces that are convex and not highly undulating are usually quite small. Additionally ideal planar reflective surfaces are a limiting case where there is no epipolar deviation.
2. Parts of the same reflection may not appear the same in all images taken from different views. They are often distorted, broken up or missing in the other images. Therefore keypoints on reflections in the image tend to introduce spurious matches.
3. Large homogeneous surfaces such as the panel of a car are in general texture impoverished. Therefore descriptors evaluated on such surfaces are not sufficiently discriminate. On the other hand, textured non-reflective areas of the scene which need not necessarily be spatially well distributed across the images may generate more descriptive keypoints and therefore stronger point matches. An example is shown in Fig. 1 where SIFT keypoints were matched using SIFT descriptors with SIFT matching (nearest neighbor / ratio test [2]). The strongest matches are localized to a corner of the image containing the wheel of the car which is comparatively less reflective and better textured. Since the matches are not spatially well distributed such matches can produce degenerate configurations in subsequent SfM tasks such as estimating the fundamental matrix.

Contributions. We propose a method to obtain reliable point correspondences in scenes dominated by highly reflective and largely homogeneous sur-

faces. The noise level in correspondences obtained from our method are sufficiently low to perform subsequent SfM tasks such as recovering the epipolar geometry. Our method is able to obtain a sufficient amount of representative matches (inliers) which can be used to recover the epipolar geometry of the scene from images where baseline methods fail (Sec. 4). Unlike existing methods [12, 13] our method does not place any restrictions on the camera (*e.g.* affine camera, small motions). Moreover, it works on scenes with highly specular and reflective surfaces of vehicles, glass paneled buildings *etc.*, which create a lot of inter-object reflections. Instead of detecting keypoints, we propose to consider all points along image edges. Most of such edge points are usually disregarded in conventional keypoint detection and matching methods which are known to give good results in non-reflective and well textured scenes. We match all edge points employing a dense descriptor, DAISY [14], which can be computed quickly at all pixels in the image and therefore at all edge points. Additionally, a spatial constraint is enforced by dividing the image into a grid of buckets and selecting only the best k putative matches from each bucket, to ensure better spatial distribution of the point matches. Although it is possible to use SIFT to densely compute descriptors for all pixels, dense SIFT (DSIFT [15]) needs to be computed at a predetermined scale since there is no keypoint detection step to determine the scale. Using a hand tuned fixed scale will calculate feature descriptors that do not properly describe point features that are of a different scale, resulting in low matching scores for potential inlier point matches. Alternatively, computing dense SIFT descriptors at a range of scales for all edge points and performing subsequent matching would make the computation complexity prohibitively high and we have not attempted this in our investigation. On the other hand the DAISY descriptor naturally incorporates gradient histograms computed at a range of scales for each pixel at locations radially distributed around the interest point. DAISY has been shown to be more computationally efficient [14] than SIFT. Hence we used DAISY descriptors for our method.

2 Related work

Most SfM and multi-view stereo algorithms which work on images of several views of a 3D scene require finding some form of correspondences between the views. Much work has been done on detecting and identifying correspondences between multiple views of a 3D scene. However, much of this work is targeted towards images of non-reflective and well textured objects and scenes.

For example, work which has received much attention include *Photo Tourism* [1], which employs the SIFT [2] key point detection and matching algorithm to find point correspondences. This work was originally intended for tourist images commonly found on the Internet which include outdoor landscapes and historic buildings. As such it does not work well with images of highly reflective objects containing largely homogeneous regions.

It is worth noting at this point that feature detection and description are two separate tasks although some algorithms such as SIFT [2], SURF [16] and

BRISK [17] tend to do both. Other methods such as the key point and edge detector by Harris *et al.* [18] focus only on the detection aspect. Some common feature descriptors include the use of a histogram of oriented gradients (HoG) and Phog/Phow descriptors [19] which are commonly used in image classification and recognition.

Detecting regions covariant with a certain class of transformations can be useful in finding correspondences between views and [20] compares some common affine region detectors including MSER, IBR, EBR, Hessian-Affine and Harris-Affine. Wide-baseline correspondences have been found by [21] using MSER, [22] using edge descriptors and more recently by [14] using DAISY descriptors. However, these methods by themselves are not well suited for images of highly reflective objects with largely homogeneous regions such as cars, reflective buildings *etc.*.

Recent work by Lin *et al.* [23] finds correspondences and camera pose using motion coherence on scenes which were previously regarded as feature impoverished SfM scenes; containing largely edge cues but few corners. However, their method seems to be intended primarily for scenes consisting of long edges and few corners such as images of buildings and cupboards. Edge based features have also been used by [24] for shape recognition. However, their work seems to be focused on simple shapes such as bicycles and tennis rackets, where edges tend to give strong cues in otherwise poorly textured scenes. Our car images on the other hand, do not guarantee reliable edges that can be matched across images as edges since the edges are often fragmented and noisy.

Although we do not directly match edges, our proposed methodology matches points along image edges. Shape contexts [25] use points along object edges for matching shapes and object recognition but not for obtaining point correspondences, which is the focus of our work. Since we match all image edge points, we need a dense descriptor which can be quickly evaluated over all edge points. Although dense implementations of SIFT [2] and SURF [16] exist, we prefer to use the DAISY [14] descriptor which is faster and also better suited for wide-baseline images. Faster rotation invariant GPU implementations of the DAISY also exist [26], although we have not used it in our work.

Reflections are not necessarily harmful for the recovery of the epipolar geometry (EPG) between two images. Work done by Saminathan *et al.* [11] shows that the epipolar deviations of specularities on convex surfaces which are not highly undulating are usually quite small.

Prior work by [12] estimates the EPG using apparent contours for the limited case of affine and circular motions. [13] use straight line edges for EPG and point matching. Our method does not place such restrictions on the camera motion or type, nor on the type of edges.

3 Problem formulation and proposed solution

Our goal is to obtain point correspondences from two images of an object with highly reflective and largely homogeneous regions. The obtained correspondences

should be good enough for SfM tasks such as recovering the epipolar geometry of the scene or estimating a homography transform for near planar objects in the scene. Our proposed method for obtaining reliable point correspondences is as follows.

3.1 Putative point matches

Given two images I and I' with point sets P and P', we wish to find the correct mapping $m(\boldsymbol{p}) = \boldsymbol{p'}$ for points $\boldsymbol{p} = (u, v) \in P = \{\boldsymbol{p_1}, \boldsymbol{p_2}, ..., \boldsymbol{p_{n_1}}\}$ and $\boldsymbol{p'} = (u', v') \in P' = \{\boldsymbol{p_1}', \boldsymbol{p_2}', ..., \boldsymbol{p_{n_2}}'\}$. Suppose we have a feature descriptor $\boldsymbol{\phi}(\boldsymbol{p})$ evaluated on point $\boldsymbol{p}$ and a suitable distance measure $d(.)$ to compare two descriptors. An optimal assignment for $\boldsymbol{p} \in P$ would be

$$m(\boldsymbol{p}) = \underset{\boldsymbol{p'} \in P'}{\operatorname{argmin}}\, d\left(\boldsymbol{\phi}(\boldsymbol{p}), \boldsymbol{\phi}(\boldsymbol{p'})\right) \tag{1}$$

Selecting candidate points. It is common practice [10] to use salient feature points (key-points) in the image as candidate points $\boldsymbol{p}, \boldsymbol{p'}$ to perform matching. Commonly used methods to obtain key-points are as follows.

Harris corner points [18] are obtained in a gray-scale image I by considering the sum of squared differences (SSD) of a 2D patch at location (u, v) and shifting it by (x, y). Let I_x and I_y be the partial derivatives of I such that

$$I(u + x, v + y) \approx I(u, v) + I_x(u, v)x + I_y(u, v)y \tag{2}$$

The weighted SSD between these two patches is given by

$$S(x, y) = \sum_u \sum_v w(u, v)\ \left(I(u + x, v + y) - I(u, v)\right)^2 \tag{3}$$

A corner (or an interest point) is characterized by a large variation of S in all directions of the vector (x, y). The Harris matrix is defined as

$$A = \sum_u \sum_v w(u, v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{bmatrix} \tag{4}$$

where angle brackets denote averaging (*i.e.* summation over (u, v)). The Harris matrix A should have two "large" eigenvalues to be an interest point. Since computing eigenvalues is computationally expensive, interest points are obtained using

$$M_c = \lambda_1 \lambda_2 - \kappa \left(\lambda_1 + \lambda_2\right)^2 = \det(A) - \kappa\ \mathrm{trace}^2(A) \tag{5}$$

where κ is a tunable sensitivity parameter.

The SIFT [2] keypoint detector efficiently searches over different scales and image locations using a difference-of-Gaussian function. At each candidate location, key-points are detected based on measures of their stability. Thereby, after the detection step, the scale is known for each key-point. The SIFT descriptor $\phi(\boldsymbol{p})$ is obtained at key-point $\boldsymbol{p}$ using local image gradients measured at the scale obtained from the key-point detection step.

We show experimentally in Sec. 4 that key-points from the above methods do not in general result in reliable point correspondences across photographs dominated by large reflective and homogeneous regions. Fig. 1 shows an example where SIFT [2] key-points and SIFT nearest neighbor matching [2] result in point matches which are concentrated towards a corner of the image which has relatively non-reflective regions. Such point matches are unsuitable to recover the epipolar geometry of the scene as it is not spatially well distributed to describe the 3D scene. Often strong key-points in reflective homogeneous surfaces are caused by reflections which are may not be present in the other view and hence cannot be matched. On the other hand, the homogeneous surface itself does not have strong features that can be detected as key-points, apart from points along edges of the surface. Hence it makes sense to simply focus on points along image edges.

Image edges have been known to be helpful when working with feature impoverished imagery. For example, [27] have used edge features to improve the performance of visual tracking in the presence of motion blur, in a simultaneous localization and mapping (SLAM) application using video sequence of mostly non-reflective scenes. Also, [22] have used an edge based descriptor to obtain wide baseline correspondences to perform structure from motion (SfM) on imagery of scenes mostly dominated by straight line edges. In a similar spirit, we found that the quality of the obtained point correspondences and the structural information obtained subsequently can be greatly improved by restricting the candidate points to points lying along image edges. Therefore, we select image point sets P and P' such that the points lie on edges in the image which are defined as follows.

Edges in the image. We define image edges as sharp changes in contrast occurring in an image which could be caused due to a genuine artifacts on the surface of an object or due to reflections caused from surrounding objects (Fig. 2).

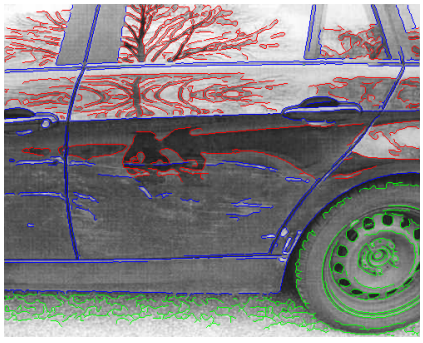


Fig. 2. Various edges in the car door image include edges caused by reflections (red), edges caused by the surface of the car body (blue) and other edges (green). Best viewed in color.

Let the set $e_i = \{\boldsymbol{p_j}, \boldsymbol{p_k}, \ldots\}$ be an image edge segment in image I containing a set of edge points. We obtain the set of all edge points $E = \bigcup_i e_i$ in image I and similarly E' in I'. Our goal then, is to find point matches as per Eq. 1 considering only points which lie on image edges such that $\boldsymbol{p} \in E$ and $\boldsymbol{p'} \in E'$. We used the popular Canny [28] edge detector which has been shown to perform well experimentally [29] with parameters adopted to the data. We used the MATLAB Canny implementation which uses a standard deviation $\sigma = \sqrt{2}$ and computes the two hysteresis thresholds relative to the highest gradient magnitude in the image.

Matching edge points require feature descriptors to be evaluated at each edge point rather than on sparse key-points. It is convenient to use a dense feature descriptor to this end. Owing to its speed and use with wide baseline stereo images, we chose the DAISY [14] feature descriptor for our work.

DAISY descriptor. Inspired by SIFT and GLOH, the DAISY [14] descriptor uses histograms of gradients. However, rather than a Difference of Gaussian (DOG), DAISY uses a Gaussian weighting and a circularly symmetric kernel, making it much faster to compute densely. Gradients are calculated at locations radially distributed around each pixel with larger regions and increasing levels of smoothing as the radial distance increases as shown in Fig. 3. For a

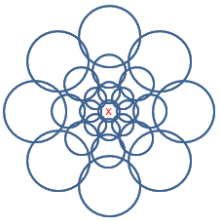


Fig. 3. DAISY [30, 14] descriptor orientation map regions (circles) about pixel 'X' for $H = 8$

given image I, an H number of *orientation maps* $G_o(u, v)$ for $1 \leq o \leq H$ are computed where G_0 is the image gradient norm at location (u, v) in direction o such that $G_o = max\left(\frac{\partial I}{\partial o}, 0\right)$. Each orientation map is convoluted several times with Gaussian kernels G_Σ of different standard deviations Σ to obtain convoluted orientation maps for different sized regions $G_o^\Sigma = G_\Sigma * G_o$. The size of the region is controlled by Σ. As convolutions with a large Gaussian kernel can be obtained by consecutive convolutions with smaller Gaussian kernels, orientation maps at different scales can be obtained very efficiently as $G_o^{\Sigma_2} = G_{\Sigma_2} * G_0 = G_\Sigma * G_{\Sigma_1} * G_o = G_\Sigma * G_o^{\Sigma_1}$ where $\Sigma = \sqrt{\Sigma_2^2 = \Sigma_1^2}$.

Sift on edge points. As SIFT [2] is limited to key-points, we considered Dense SIFT (DSIFT [15]) on edge points. However, the scale which is computed automatically during key-point detection in SIFT [2], needs to given explicitly

with DSIFT. On the contrary, the DAISY descriptor, which is also a dense descriptor, incorporates a range of scales by definition. Additionally, as per the computation complexity evaluation in [30, 14], DAISY is also a lot faster than SIFT. Hence we used DAISY.

Edge feature ambiguities. Indeed edge points on their own are not as discriminative as corners and blobs. However, images of highly reflective objects with large homogeneous regions lack discriminative corners/blobs with sufficient spatial distribution to recover the EPG. For such images, DAISY descriptors evaluated over edge points give better results (Sec. 4). The spatial constraint Sec. 3.2 also reduces the ambiguity of edge point matches. Sophisticated techniques such as graphical models [31] could also be utilized to employ smoothness constraints enforcing points on the same edge in I to match to points on a single edge in I'. In practice however, detected edges are often noisy and tend to fragment in an unpredictable manner. Hence, we found the simple greedy matching in Eq. 1 to be more effective. Next we describe estimating the EPG using putative point correspondences.

3.2 Recovery of the epipolar geometry (EPG)

Given two images that describe a 3D scene, its epipolar geometry (EPG) gives information about the camera setup in a projective sense. The EPG can be used to infer knowledge about the 3D scene via triangulation or stereo matching. In the case of an uncalibrated and unknown camera setup, image rectification may be performed prior to stereo matching.

A given point in one image will lie on its epipolar line in the second image, which is actually the projection of the back projected ray from the first image on to the second image. The epipolar geometry is described algebraically by the *Fundamental Matrix* F [10], which is based on this relationship. To be more specific, suppose two corresponding points $\boldsymbol{p}, \boldsymbol{p'} \in \mathbb{R}^2$ on I and I' have homogeneous coordinates $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{P}^2$ and F is the 3×3 fundamental matrix of rank 2, then $\boldsymbol{x}'^T F \boldsymbol{x} = 0$ for all correct point correspondences $\boldsymbol{x} \leftrightarrow \boldsymbol{x}'$.

RANSAC. Given a set of noisy point correspondences, the EPG may be robustly found using RANdom SAmple Consensus (RANSAC) [32] based methods. The essence of these methods is to find a fundamental matrix F such that $\boldsymbol{x}'^T F \boldsymbol{x} = 0$ for a random subset of the given points such that it agrees with the largest number of the remaining points. This is repeated for a given number of iterations and the best solution is selected. The RANSAC approach is robust in the presence of noisy outliers with considerable errors.

PROSAC has been shown to perform better than RANSAC by assuming that putative matches with higher quality (i.e with a lower matching distance $d(.)$) are more likely to be inliers [33]. In our case however, inter object reflections on the reflective surfaces (*e.g.* reflections of trees on vehicle panels and glass) may generate high quality matches which are outliers to the EPG of the main scene. Therefore we do not consider the matching distance in the RANSAC step.

We use the normalized 8pt algorithm for model fitting in each RANSAC iteration and a distance threshold of 0.01 to filter outliers [10]. We use M-estimator

SAmple Consensus (MSAC) [34] as it is known to converge faster than standard RANSAC. However, the number of samples required to ensure with a given probability, that at least one sample has no outliers for a given sample size, increase exponentially as shown by [10]. This makes images with highly reflective and homogeneous regions which give very noisy point correspondences, very challenging to work with. Selecting points along edges in the image gives more reliable matches for images with largely homogeneous regions and reflections.

Spatial Constraint. To obtain an EPG which is representative of the actual 3D scene, it is important to have matching inlier points which are spatially well distributed across the images. However, naive feature matching of reflective images tend to concentrate correct point matches over areas which are relatively less reflective as shown in Fig. 1. To avoid this problem, we enforce a *spatial constraint* inspired by [35, 36, 37]. The complete matching algorithm with the spatial constraint for obtaining putative point correspondences is given in Alg. 1.

Input : Images I and I'
Output: Putative point correspondences
1) Find the set of edge points E in image I and E' in image I'
2) **Match edge points (asymmetric):** For each edge point $\boldsymbol{p_i} \in E$ find the matched edge point $\boldsymbol{p_j}' \in E'$ as
$p' = m(\boldsymbol{p}) = \text{argmin}_{\boldsymbol{p}' \in E'}\, d\left(\boldsymbol{\phi}(\boldsymbol{p}), \boldsymbol{\phi}(\boldsymbol{p}')\right)$
3) **Enforce spatial constraint:** Consider a rectangular grid of $b_W \times b_H$ spatial buckets over I. Pick the best k matches with the lowest $d(.)$ from each bucket

Algorithm 1: Matching algorithm with the spatial constraint

We are not limited to small camera motions or scale changes as we match edge points asymmetrically from I to I' and only consider buckets over I. As an extreme case, consider the bucket at the top-left corner of the bucketed image I. We may well pick a matching point from the bottom-right corner in the other image I' (which is not bucketed) as long as the matching distance is within the lowest k distance values for the bucket.

We present next, an experimental evaluation of our method along with baseline comparisons.

4 Experiments

We compare our method quantitatively and qualitatively against baseline methods. Experiments were done on the standard DAISY dataset [30, 14] and our own car dataset of over 70 images of highly reflective car images. We experimentally found that for the spatial constraint parameters in Alg. 1, $b_W = I_W/16$, $b_H = I_H/16$ and $k = 2$ gave good results for image pairs of size $I_W \times I_H$ pixels each.

Table 1. Description of methods

Method	Description
deg	DAISY descriptors on edge points - with spatial constraint (ours)
de	DAISY descriptors on edge points - no spatial constraint (ours)
dhg	DAISY descriptors on Harris corner points - with spatial constraint
dh	DAISY descriptors on Harris corner points - no spatial constraint
dsg	DAISY descriptors on SIFT key-points - with spatial constraint
ds	DAISY descriptors on SIFT key-points - no spatial constraint
sg	SIFT descriptors on SIFT key-points and SIFT matching - with spatial constraint
s	SIFT descriptors on SIFT key-points and SIFT matching - no spatial constraint

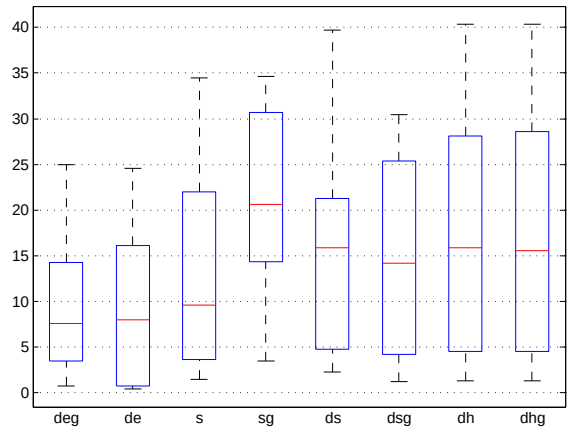


Fig. 4. The box plots show the *comparison measure* of Zhang [36] for our method *de* and baseline methods *s,ds and dh* with *g* at the end indicating tests where the spatial constraint was enforced. We used the DAISY dataset [30, 14]. A lower *comparison measure* indicates better performance. Our method *deg* has the lowest median and inter quartile range (IQR) for the *comparison measure*.

4.1 Quantitative results and comparison with baseline methods

We quantitatively evaluate the quality of the EPG recovered from our method and baseline methods as follows. We use the method adopted by [36] to measure the similarity between the recovered EPG and the ground truth EPG. The method gives a *measure of comparison* between the recovered fundamental matrix F and the ground truth fundamental matrix F_{gt}. As described in [36], the measure is obtained by considering the perpendicular distances between points and corresponding epipolar lines obtained using both fundamental matrices in a symmetric manner. Simillar fundamental matrices give a lower value for the

measure. In our experiments, the better method should recover an EPG closer to the ground truth EPG and therefore give a lower comparison measure. We use the wide baseline *fountain* and *herzjesu* images with ground truth camera calibration information from the DAISY dataset [30, 14]. We use the provided ground truth projection matrices to compute a ground truth fundamental matrix F_{gt} for a given image pair.

Fig. 4 shows box plots of the *comparison measure* using the method by [36] explained above. A summary of the methods evaluated in this paper along with acronyms used are given in Tbl. 1. All images in the datasets have the same dimentions. Although the *comparison measure* evaluated using only the inlier point correspondences is in the order of sub pixels, we evaluate the *comparison measure* [36] over all point correspondences when comparing methods in Fig. 4. This is because even an incorrectly estimated epipolar geometry will still give a very low *comparison measure* for degenerate cases since the inlier point matches satisfy the incorrectly estimated fundamental matrix (*e.g.* coplanar inlier points). We see in Fig. 4 that our method with the spatial constraint *deg* gives the lowest median *comparison measure* (indicated by the horizontal line in the middle of each box) and also has the lowest dispersion or spread as seen by the interquartile range indicated by the ends of each box. The baselines *ds* and *dh* improve marginally with the spatial constraint (*dsg* and *dhg*). However, the spatial constraint causes a significant performance drop with SIFT key-points and SIFT matching (*s* vs *sg*). The SIFT distance ratio (nearest neighbor test) already filters out matches with features that are not very discriminative but could have supported the correct EPG. Enforcing the spatial constraint in *(*sg) further reduces the number of matches which may support the correct EPG. Therefore, enforcing the spatial constraint in *sg* yields poor quality matches which significantly affects the EPG computation. The results in Fig. 4 show that our methods *deg* and *de* continue to perform better than the baselines, even with the spatial constraint. Qualitative results shown in Sec. 4.2 indicate the same.

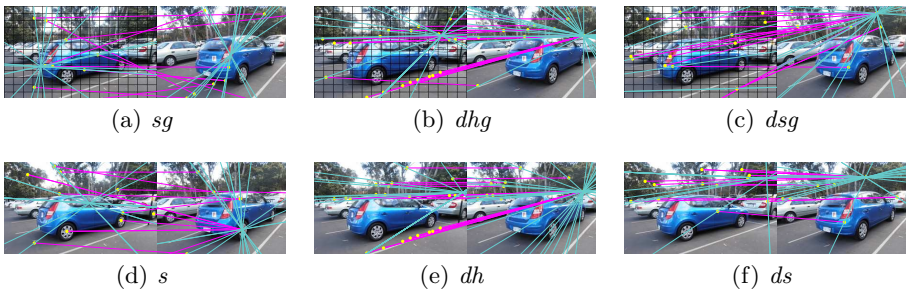
Matching Distance. Putative matches were found using the SIFT distance ratio (nearest neighbor test) [2] for baseline methods *s* and *sg*. Results on relatively non-reflective images were comparable with our method *deg* (Fig. 4). However, such matches are not very reliable with very reflective images (methods *s* and *sg* in Figs. 5 & 8). For matching DAISY descriptors in *de*, *deg*, *ds*, *dsg*, *dh* and *dhg*, we used the L2 norm for $d(.)$ in Eq. 1 as per [30, 14].

Descriptor Scale. The scale was computed automatically from SIFT keypoint detection in baseline *s*. For the other methods, we used the DAISY descriptor [30, 14] which has image differences obtained at radially distributed positions about the initial point/pixel, computed by applying increasingly larger Gaussian kernels when moving away from the point. We used $R = 15, Q = 3, T = 8, H = 8$ as per [30, 14] and Sec. 3.1.

4.2 Qualitative results

To get a sense of the recovered EPG we present some qualitative results. We evaluated our method and baselines qualitatively on our car dataset containing

over 70 image pairs. Results on the entire dataset are provided as supplementary material. A typical result is shown in Fig. 5 and Fig. 6 with methods denoted as per Tbl. 1. The recovered EPG from the baseline methods in Fig. 5 are clearly wrong as the epipolar lines seem to indicate that the photographer has walked towards the car where as in reality the photographer has moved side ways. As the recovered epipoles are incorrectly located inside the images, uncalibrated rectification [10] cannot be performed. On the other hand, our method *deg* in Fig. 6(a) recovers a significantly better EPG in Fig. 6(b) for the same image pair. The near horizontal direction of the epipolar lines correctly reflect the movement of the camera. In fact, it is possible to perform uncalibrated stereo rectification (Fig. 6(e)). Not enforcing the spatial constraint (*de*) gives poorer results (Fig. 6(c) and Fig. 6(d)) in this instance, which are not suitable for stereo rectification.



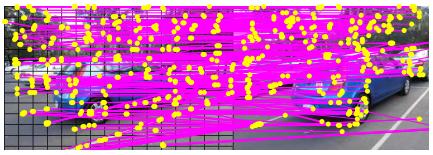
(a) *sg* (b) *dhg* (c) *dsg*

(d) *s* (e) *dh* (f) *ds*

Fig. 5. EPG and inliers for the baseline methods discussed in Sec. 4.2. Notation *c.f.* Tbl. 1 and Sec. 4.1. Color code: cyan lines - epipolar lines, yellow dots - matched points, magenta lines - point correspondences. Best viewed in color. Images may be cropped for clarity.

Note that the uncalibrated rectification process introduces a projective distortion in the transformed images which is as expected [10]. Hence the apparent disparities between the blue points in the rectified left image and the superimposed red points from the rectified right image (Fig. 6(e)) may not correctly indicate inverse depth as with calibrated rectification. EPG and uncalibrated rectification results on the image pair in Fig. 1 are shown in Fig. 7.

We further verify our method on a highly reflective image pair of a building in Fig. 8. The camera motion between the two photographs is clearly horizontal. Hence the recovered epipolar (EP) lines (shown in cyan) should be horizontal. This is reflected correctly in our methods *deg* and *de*. However, the EPG recovered using the baseline methods do not indicate this and is clearly wrong. Among the baselines, *dhg* performs better but EP lines (particularly at the top) are not horizontal.

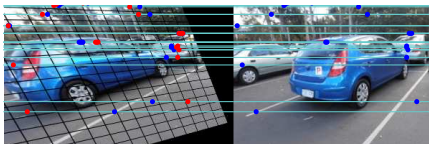


(a) Putative matches with *deg* (with spatial constraint)

(b) Plausible EPG and inliers from (a)

(c) Putative matches with *de* (no spatial constraint)

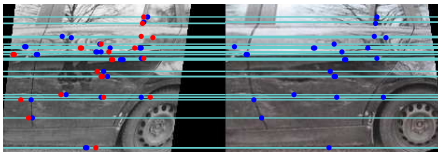
(d) Incorrect EPG and inliers from (c)



(e) Uncalibrated stereo rectification using EPG in (b). Left image has points from right image superimpose in red.

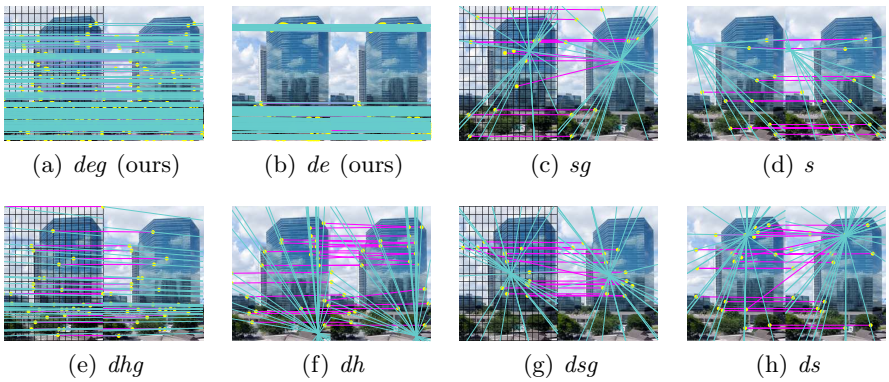
Fig. 6. Results from our method discussed in Sec. 4.2 showing the effect of the spatial constraint.
Notation *c.f.* Tbl. 1 and Sec. 4.1. Color code: cyan lines - epipolar lines, yellow dots - matched points, magenta lines - point correspondences. Best viewed in color. Images may be cropped for clarity.

(a) EPG and inliers



(b) Uncalibrated stereo rectification

Fig. 7. Results using our method *deg* on photographs of a very reflective car door.
Notation *c.f.* Tbl. 1 and Sec. 4.1. Color code: cyan lines - epipolar lines, yellow dots - matched points, magenta lines - point correspondences. Best viewed in color. Images may be cropped for clarity.



(a) *deg* (ours) (b) *de* (ours) (c) *sg* (d) *s*

(e) *dhg* (f) *dh* (g) *dsg* (h) *ds*

Fig. 8. Recovered EPG and inlier point correspondences from a photograph of a highly reflective building. Since the camera motion between the two images is horizontal, the recovered epipolar lines should be horizontal. This is reflected correctly in our methods *deg* and *de*, unlike with the baseline methods. The upper EP lines with *dhg* are not horizontal. The spatial grid is overlaid on the left image where the spatial constraint was enforced.
Notation *c.f.* Tbl. 1 and Sec. 4.1. Color code: cyan lines - epipolar lines, yellow dots - matched points, magenta lines - point correspondences. Best viewed in color. Images may be cropped for clarity.

5 Discussion

We present a method for finding reliable point correspondences in images of scenes dominated by highly reflective and largely homogeneous surfaces. Conventional methods for finding point correspondences are mainly designed for textured and non-reflective surfaces. As such they generate a lot of spurious matches from images with highly reflective and homogeneous surfaces and give poor results when recovering the epipolar geometry of the scene. We have proposed a novel method of combining established computer vision techniques by matching points along image edges and enforcing a spatial constraint to obtain reliable point correspondences from such images, resulting in sufficiently low noise levels. In addition to providing theoretical intuition, we have experimentally showed that our approach gives good results on images with highly reflective and homogeneous surfaces where baseline methods fail. An interesting future direction would be to explore QDEGSAC [38] to avoid potentially degenerate configurations with unknown camera calibration. An interesting application would be to detect the reflections in the images based on depth cues after performing a 3D reconstruction of the scene based on the obtained point correspondences and recovered EPG.

Acknowledgement. This work was supported by Control €xpert.

Bibliography

[1] Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the World from Internet Photo Collections. International Journal of Computer Vision (IJCV) (2008)

[2] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) (2004)

[3] Jayawardena, S., Yang, D., Hutter, M.: 3D model assisted image segmentation. In: Proc. of the International Conference on Digital Image Computing Techniques and Applications (DICTA), IEEE (2011)

[4] Jayawardena, S., Hutter, M., Brewer, N.: A novel illumination-invariant loss for monocular 3D pose estimation. In: Proc. of the International Conference on Digital Image Computing Techniques and Applications (DICTA), IEEE (2011)

[5] Jayawardena, S.: Image Based Automatic Vehicle Damage Detection. PhD thesis, The Australian National University (2013)

[6] Pylvanainen, T., Berclaz, J., Korah, T., Hedau, V., Aanjaneya, M., Grzeszczuk, R.: 3D City Modeling from Street-Level Data for Augmented Reality Applications. In: 3DIMPVT, IEEE (2012)

[7] Greenspan, H., Gordon, S., Zimmerman, G., Lotenberg, S., Jeronimo, J., Antani, S., Long, R.: Automatic detection of anatomical landmarks in uterine cervix images. Medical Imaging, IEEE Transactions on **28** (2009) 454–468

[8] Zimmerman-Moreno, G., Greenspan, H.: Automatic detection of specular reflections in uterine cervix images. In: Medical Imaging, International Society for Optics and Photonics (2006) 61446E–61446E

[9] Wu, T.T., Qu, J.Y.: Optical imaging for medical diagnosis based on active stereo vision and motion tracking. Optics express (2007)

[10] Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)

[11] Swaminathan, R., Kang, S., Szeliski, R., Criminisi, A., Nayar, S.: On the motion and appearance of specularities in image sequences. Proc. of the European Conference on Computer Vision (ECCV) (2002)

[12] Mendonça, P.R., Cipolla, R.: Estimation of epipolar geometry from apparent contours: Affine and circular motion cases. In: Proc. of Computer Vision and Pattern Recog. (CVPR). (1999)

[13] Schmid, C., Zisserman, A.: The geometry and matching of lines and curves over multiple views. International Journal of Computer Vision (IJCV) (2000)

[14] Tola, E., Lepetit, V., Fua, P.: DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2010)

[15] Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)

[16] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. Proc. of the European Conference on Computer Vision (ECCV) (2006)

[17] Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary robust invariant scalable keypoints. In: Proc. of the International Conference on Comp. Vision (ICCV). (2011)

[18] Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference (AVC). (1988)

[19] Bosch, A., Zisserman, A., Muoz, X.: Image classification using random forests and ferns. In: Proc. of the International Conference on Comp. Vision (ICCV). (2007)

[20] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision (IJCV) (2005)

[21] Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and vision computing (IVC) (2004)

[22] Meltzer, J., Soatto, S.: Edge descriptors for robust wide-baseline correspondence. In: Proc. of Computer Vision and Pattern Recog. (CVPR). (2008)

[23] Lin, W., Cheong, L., Tan, P., Dong, G., Liu, S.: Simultaneous Camera Pose and Correspondence Estimation with Motion Coherence. International Journal of Computer Vision (IJCV) (2012)

[24] Mikolajczyk, K., Zisserman, A., Schmid, C., et al.: Shape recognition with edge-based features. In: Proc. of the British Machine Vision Conference (BMVC). (2003)

[25] Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **24** (2002) 509–522

[26] Fischer, J., Ruppel, A., Weißhardt, F., Verl, A.: A rotation invariant feature descriptor O-DAISY and its FPGA implementation. In: Prof. of the International Conference on Intelligent Robots and Systems (IROS). (2011)

[27] Klein, G., Murray, D.: Improving the agility of keyframe-based SLAM. In: Proc. of the European Conference on Computer Vision (ECCV). Springer (2008)

[28] Canny, J.: A computational approach to edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (1986)

[29] Heath, M.D., Sarkar, S., Sanocki, T., Bowyer, K.W.: A robust visual method for assessing the relative performance of edge-detection algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (1997)

[30] Tola, E., Lepetit, V., Fua, P.: A Fast Local Descriptor for Dense Matching. In: Proc. of Computer Vision and Pattern Recog. (CVPR). (2008)

[31] Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2001)

[32] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Association for Computing Machinery (ACM) (1981)

[33] Aghazadeh, O., Sullivan, J., Carlsson, S.: Novelty detection from an egocentric perspective. In: Proc. of Computer Vision and Pattern Recog. (CVPR). (2011)

[34] Rogers, M., Graham, J.: Robust active shape model search. Proc. of the European Conference on Computer Vision (ECCV) (2006)

[35] Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial Intelligence (AI) (1995)

[36] Zhang, Z.: Determining the epipolar geometry and its uncertainty: A review. International Journal of Computer Vision (IJCV) (1998)

[37] Kitt, B., Geiger, A., Lategahn, H.: Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In: Intelligent Vehicles Symposium (IV). (2010)

[38] Frahm, J.M., Pollefeys, M.: RANSAC for (Quasi-)Degenerate data (QDEGSAC). In: Proc. of Computer Vision and Pattern Recog. (CVPR). (2006)