An Interest Point Detector and Local Image Descriptor for 3D Rigid Scenes

Dissertation thesis proposal

Arne Jacobs*

Center for Computing Technologies, University of Bremen
jarne@tzi.de

Abstract. Approaches for the detection of points of interest (POI) have been applied to several different problems in computer vision, like motion tracking, image registration, object recognition, video indexing, and even texture classification. Since accompanied by corresponding local image descriptors, they have received growing attention. Current approaches completely remain in the two dimensions of the image plane, although they are used in applications like wide baseline stereo matching, or multiple viewpoint object recognition, where 3D transformations occur, which can only be approximated by 2D transformations in the case of highly planar scenes. Accordingly, it has been shown, that current approaches have problems with complex 3D scenes. In the proposed thesis, an interest point detector and a corresponding local image descriptor will be developed that are invariant against (local) 3D rigid transformations. Currently, the feasibility of a shape from shading approach is investigated. The thesis is expected to be finished in 2008.

In image processing, a *point of interest* (POI) is a point in an image, that has special properties which make it stand out in comparison to its neighbouring points. What exactly these properties are differs between POI detection approaches. Pritchett and Zisserman refer to interest points as "reliable generic visual primitives" [1]. They are often also called *salient points*. Most POI detection approaches share the following common properties:

- Only a small fraction of all image points is regarded as points of interest.
- Points of interest denote image points with high information content. Here, the notion of information content is mostly relatively low-level, based on information theoretical considerations. Because a single image point does not contain much information, the direct neighbourhood of a point is also considered during the detection process. Thus, in practice, an interest point denotes an image point as the center of a local image patch.
- Interest points are therefore more distinctive than the average image point. This property is exploited by several applications of interest point detectors.

* Advisor: Prof. Dr. Otthein Herzog

– Approaches to detect points of interest are often accompanied by corresponding *local image descriptors*, which comprise the information of interest points and their neighbourhood image patch in a compact form.

POI detection approaches have long been used in image processing. Because of their distinctive nature, they are used in motion estimation and tracking, minimizing the so-called *aperture problem* [2].

In shape from stereo approaches, points of interest are used for matching two widely separated views [1, 3].

They also play an important role in object recognition [4, 5], by describing object classes as sets of local features with spatial relations between them. Due to their local nature, the use of interest points and local image descriptors makes it possible to deal with occlusions. Furthermore, the need for a preceding segmentation, which is a very difficult problem of its own, is omitted.

Recently, local image descriptors have been used for efficient indexing and retrieval of images and videos [6]. There, instances of a local image descriptor for a huge set of interest points on real world images and videos are clustered to form a kind of visual vocabulary. Each cluster corresponds to a *word*. Then, established and efficient text retrieval algorithms are used to index video data, using an inverted filesystem approach.

Another recent application for local image descriptors is the representation and classification of texture [7].

In the proposed thesis, a new approach for detection of points of interest and a corresponding local image descriptor shall be developed, that can be applied to 3D rigid scenes. I.e., the approach shall be able to cope with 3D rigid transformations of the image data.

In the next section, the thesis' goals and preconditions will be formulated in further detail. Existing approaches related to the thesis will be covered in Sec. 2. Section 3 introduces the proposed approach. The paper concludes with a summary and an outlook (Sec. 4).

1 Preconditions and problem formulation

Two problems are addressed in this thesis:

– The problem of creating an interest point detector that is repeatable under 3D rigid transformation of the underlying scene
– The problem of creating a local image descriptor that is invariant against said transformations

Input to the algorithm shall be single, twodimensional colour or grayscale images of real world scenes. Alternatively, if image sequences are used, no 3D rigid transformation between successive images is presumed, because, in the general case, it is not always existent. In many cases, there will be only 2D motion induced by camera motion, like pan or zoom. This means that the algorithm can not rely on 3D rigid motion to be present, and thus will not be able to

use the additional information possible to extract given the motion, i.e., depth information.

The requirement for the interest point detector is repeatability under 3D rigid transformations. In [8], repeatability is defined as follows:

> "Repeatability is the variation in measurements taken by a single person or instrument on the same item. A measurement is said to be repeatable when this variation is small."

This means, if any point in one input image is detected as a point of interest, the same point should also be detected in another image which shows the same scene undergoing any 3D rigid transformation.

More precisely: Given a detected interest point p in an image I that is a projection of a point w in the world onto the image plane of I, $p = t(I, w)$, and given a second point $p' = t(I', w)$, which is a projection of w onto the image plane of a second image I', taken from another viewpoint. Then, if p' is visible in I', it should also be detected as a point of interest in I'.

The repeatability criterion is crucial for applications like image registration or feature-based motion estimation. Given two images, if matching is only done between the two sets of interest points detected in two images, failing to meet the repeatability requirement will inevidently result in false matches. The same is true for image similarity matching, where missing interest points result in a decrease of the similarity measure.

The requirement for the local image descriptor is invariance against 3D rigid transformations. In [9], invariance is defined as follows:

> "An invariant is something that does not change under a set of transformations. The property of being an invariant is invariance."

In this case, considering the local nature of the image descriptors, the set of transformations consists of local 3D rigid transformations.

The local image descriptor will be used for comparison of two points of interest by means of a similarity measure between their descriptors. Invariance therefore means, that two projections of the same point in the world should be similar to each other, and that two projections of two points in the world should be similar, if the neighbourhood (in the world) of the two points is similar.

2 Related work

Early approaches for the detection of interest points, like the so-called Harris Corner Detector [10], see POI as corners, junctions, end edges, i.e., as spatial discontinuities in the image signal that are intrinsically twodimensional. Harris and Stephens build up a (2×2) covariance matrix of the two partial derivatives of the grayvalue image data in a local image patch. The lower of the two eigenvalues of the covariance matrix then is a measure of the intrinsic two-dimensionality of the patch. The selected points of interest correspond to points where two nearly

orthogonal gradient directions are present. The approach is invariant against rotation. It is not scale invariant, however.

A scale invariant interest point detector and a corresponding image descriptor, called the *SIFT descriptor*, were introduced in [4]. Since then, the SIFT descriptor has been used in several other approaches, although the interest point detector used varied.

The SIFT descriptor has, however, been modified and extended. In [11], principal component analysis (PCA) is applied to the image patch gradient distribution. The resulting local image descriptor has 20 or even less dimensions, in contrast to the 160 dimensions of the original SIFT descriptor. In [12], the GLOH descriptor is introduced, which is an extension of SIFT that also uses PCA to reduce the dimensionality of the local descriptor to 36.

The rotation and scale invariance of interest point detectors, which corresponds to invariance against a four parameter transform, eventually lead to the development of interest point detectors that are invariant against affine transformations [13, 14], which are the next (and most) complex class of linear 2D transformations. While the approach in [13] is based on the same twodimensionality criterion for a salient point as the original approach of Harris et al. [10], the approach proposed in [14] is based on a more information theoretical measure, basically a local entropy measure. It also has a scale invariant predecessor in [15].

Most interest point detection algorithms have been tested for repeatability, yet in most cases, this is done with highly planar objects or even completely planar scenes. This might be due to the twodimensional nature of the four parameter and the 2D affine transform the approaches are based upon. Fraundorfer et al. [16] build up a test scenario comprising more complex threedimensional real world scenes. The repeatibility of the tested approaches reduces drastically with viewpoint changes. Surprisingly, the most common Harris detector [10], which is not even scale invariant, seems to perform very well under the given circumstances, compared with more complex approaches.

This raises the question, if variance against twodimensional transformation is the right way to tackle complex 3D scenes. Hopefully, this thesis, with the development of a 3D rigid invariant POI detector, will be a step towards the answer to this question.

3 Proposed approach

This section introduces the basic idea to approach the thesis' problems and gathers some early thoughts.

3.1 Interest point detector

One of the implications of the preconditions is, that structure from motion approaches can not be used to extract depth information from the input images: In general, there will be only one input image, and if an image sequence is supplied, the 3D rigid motion needed for structure from motion approaches can not be

presumed. Thus, such depth maps can not be used to, e.g., extract 3D corners as interest points, which would be a consequential development of the 2D corner extracting interest point detectors. Depth from stereo approaches are not applicable, either.

Instead, the idea is to use a structure from shading (SFS) approach to detect points that could be, based on the intensity variation in their neighbourhood, prominent 3D structures, e.g., 3D corners.

Without additional a priori knowledge, common SFS approaches are not suitable for this purpose, because they set strict requirements for the input data, e.g., on the prevalent lighting conditions. Furthermore, a complete depth map for the whole image is not necessarily required to identify prominent 3D structures. Instead, a local description of depth variance might be sufficient.

However, several SFS-innate problems apply for the proposed interest point detector, too:

1. The direction of the light source is not known a priori. Most SFS approaches assume a known light source direction [17]. There are, however, approaches to estimate the light source direction on a given image under certain assumptions on the scene's surface properties [18, 19].
2. There might be more than one light source. Because only a local description of depth has to be recovered here, the problem of multiple lightsources can probably be neglected, assuming that one light source will be prevalent at a given location. A more often occurring problem, particularly with outdoor images, could be the case of diffuse lighting, e.g., induced by clouds.
3. Grayvalue changes can be induced not only by shading, but also by shadows induced by other objects that occlude the light source.
4. Grayvalue variation occurs also due to variations of surface properties like color and brightness, mostly what is called texture. This is probably a bigger problem than the previous one. However, the human visual system is capable of distinguishing between shading due to depth variations and "normal" texture [20].

Research questions. The problems stated above directly raise corresponding questions that have to be tackled in this thesis.

To tackle the first problem, a testbed of real world images showing sufficiently complex depth shall be assembled and manually annotated with the ground truth light source direction. This testbed will be used to examine existing approaches to light source direction estimation. For this purpose, a state of the art on this topic has to be assembled, in addition to the few examples mentioned in this paper. The question is, if the light source direction can be reliably estimated by existing approaches, and how these approaches could be improved, if necessary.

For now, the problem of multiple light sources will be neglected, and a single light source will be assumed. The above testbed shall, however, be enhanced with outdoor images with diffuse lighting, e.g., induced by clouds.

Expected to be the least problematic of the four, the third problem will not be tackled in this thesis. Problems will probably only arise at shadow boundaries,

which will normally only cover a small fraction of an image, due to the lower dimensionality of an object's silhouette compared to the object's body.

When considering real world images, the fourth problem can not be neglected. The thesis will probably face one of the two following cases:

1. Shading and thus surface depth can not be separated from surface texture. In this case, a pseudo shape could be computed for textured regions. It will result from (the wrong) assumption that grayvalue variations due to texture are induced by the 3D surface structure. It has to be investigated if such a pseudo shape is feasible for the detection of interest points: Will a point in a textured neighbourhood be detected in several images showing that same point under different lighting conditions? This might be a difficult problem, because lighting conditions and surface texture are independent and the combination might lead to totally different pseudo shapes under different lighting conditions. A simple example: Consider a flat textured surface shown in two images with the same light source direction, but rotated by 180 degrees in one of the images. In one image, a shape from shading approach would then infer a depth map for the textured surface inverse to the other image's surface.
2. Shading due to surface depth can be separated from surface texture. In principle, this is supported by the literature [21, 20]. In this case, the problem of finding points of interest reduces to specifying a saliency measure on the computed local surface 3D structure.

To see, which of the above two is the case, a testbed of real world images with known depth information shall be assembled. Probably, a shape from stereo approach will be used to generate the depth images. The depth data can then be used to generate a "best case" separation of grayvalue variation due to shading from variation due to texture. The result can then be used to determine if such a separation can be done based on single images alone.

As for problem one, a state of the art has to be prepared, that covers shape from shading approaches for textured imagery. Using the testbed, it has to be determined, wether existing approaches are capable of doing the required separation.

A simple case and a corresponding detector. As a little thought experiment, a certain type of 3D structure, namely 3D corners, and its possible appearances in 2D images, shall receive closer attention here. Furthermore, a corresponding detector will be sketched, and it will be made plausible, why it will also be applicable to textured planar surfaces.

Figure 1 shows six sketched examples of 2D projections of 3D corners. Note that the different appearances can, in general, not be modeled by 2D linear transformations, like the four parameter transform or the affine model used in current POI detection approaches.

If the object exposing the corner is untextured in the vicinity of the 3D edges, the latter can be observed as 2D intensity edges in the projected image.

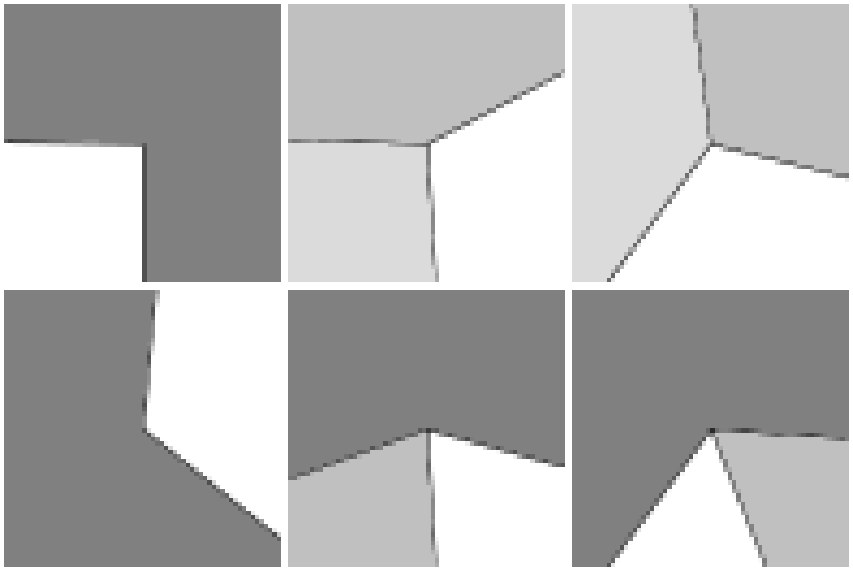

Fig. 1. Several possible appearances of 3D corners in an image. Dark gray denotes background. Black lines denote expected grayvalue discontinuities. Depending on lighting conditions, the surfaces surrounding the corner will be of varying intensity

The problem of finding such a corner then reduces to finding the junction of multiple edges. This problem has been intensively studied, and there are several approaches, e.g., in the form of steerable filters [22, 23], or using so-called *polar averaging* [24].

Similar structures may be observed as parts of a textured surface, of course. This means, that the detector will return points in the image that do not correspond to 3D corners. As long as these points are also detected under 3D rigid transformations of said surface, this is perfectly ok. Assuming the textured surface can be locally approximated by a plane, every 3D rigid transformation results in a locally 2D perspective transformation between the projected images. Thus, as the perspective transform preserves straight lines and edges as well, the transformed structure will also be a junction of intensity edges. I.e., the point will be detected in the transformed image, too.

3.2 Local image descriptor

The fourth problem stated in Sec. 3.1 will heavily influence the definition of a local image descriptor, due to the following reasions:

- If depth structure can not be separated from surface texture, the (pseudo) depth will depend solely on the grayvalue distribution in the regarded local image patch. Thus, only one of both may be used to compute the descriptor.

Probably, the pseudo depth will be more appropriate to achieve the goal of invariance against 3D rigid transformations, because it already contains 3D information.
- If depth structure can be separated from surface texture, there will be two independent types of data available for the image descriptor to base upon. First the descriptor could be based on the depth information, i.e., the local 3D structure. Such a descriptor could be easily made invariant against 3D rigid transformations, analogous to 2D descriptors. Second, the descriptor could be based on the texture's grayvalue distribution, which could also be made 3D rigid invariant using the depth information. There could even be two descriptors, which would make it possible, e.g., to index objects based on their 3D structure alone, or based on its texture, or both.

Due to the dependence upon the answers to the interest point detector research questions, the problem of defining the local image descriptor will be tackled at a later stage in the thesis. Thus, no further investigations have been made into this topic up to now.

4 Summary and outlook

Interest point detectors are a useful tool in early computer vision. Recent results have suggested, that current approaches based on invariance against common 2D transformations might not be sufficient to handle complex 3D scenes. The thesis proposal introduced here accounts for that and formulates the problem of developing an interest point detector and a corresponding local image descriptor for 3D scenes undergoing (local) 3D rigid transformations. A first idea of using shape from shading to solve the given problem has been introduced. A simple, but encouraging example of prominent 3D structures as points of interest has been shown.

The next steps are to assemble a testbed of real world images, along with proper ground truth data, and to test the state of the art in shape from shading, based on the particular requirements of the interest point detector. This will reveal if a shape from shading POI detector is feasible, and, hopefully, how existing SFS approaches have to be modified and further developed for the purpose of the thesis.

The thesis is expected to finish in 2008.

References

1. Pritchett, P., Zisserman, A.: Wide baseline stereo matching. In: ICCV. (1998) 754–760
2. Shi, J., Tomasi, C.: Good features to track. In: Proc. of the Conf. on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, IEEE Computer Society Press (1994) 593–600
3. Baumberg, A.: Reliable feature matching across widely separated views. In: CVPR. (2000) 774–781

4. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. of the International Conference on Computer Vision ICCV, Corfu. (1999) 1150–1157
5. Leibe, B., Schiele, B.: Scale invariant object categorization using a scale-adaptive mean-shift search. In: DAGM04. Springer LNCS, Vol. 3175, Tuebingen, Germany (2004) 145–153
6. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. (2003)
7. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Trans. Pattern Anal. Mach. Intell. **27** (2005) 1265–1278
8. Wikipedia: Repeatability. Wikipedia, the free encyclopedia (2005) [Online; accessed 19-August-2005: `http://en.wikipedia.org/wiki/Repeatability`].
9. Wikipedia: Invariance. Wikipedia, the free encyclopedia (2005) [Online; accessed 19-August-2005: `http://en.wikipedia.org/wiki/Invariance`].
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of the 4th Alvey Vision Conf. (1988) 189–192
11. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: CVPR. Volume 1., Washington, DC, USA, IEEE Computer Society (2004) 511–517
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Accepted to PAMI (2005)
13. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I, London, UK, Springer-Verlag (2002) 128–142
14. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: ECCV (1). (2004) 228–241
15. Kadir, T., Brady, M.: Scale, saliency and image description. International Journal of Computer Vision **45** (2001) 83–105
16. Fraundorfer, F., Bischof, H.: Evaluation of local detectors on non-planar scenes. In: Proc. of the 28th Workshop AAPR. (2004) 125–132
17. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape from shading: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **21** (1999) 690–706
18. Pentland, A.: Finding the illuminant direction. JOSA **72** (1982) 448–455
19. Zheng, Q., Chellappa, R.: Estimation of illuminant direction, albedo, and shape from shading. IEEE Trans. Pattern Anal. Mach. Intell. **13** (1991) 680–702
20. Gipsman, D.: Critical points of shading: on intensity maxima. Master's thesis, School of Computer Science, McGill University (2003)
21. Courteille, F., Crouzil, A., Durou, J., Gurdjos, P.: Towards shape from shading under realistic photographic conditions. In: ICPR04. (2004) II: 277–280
22. Freeman, W., Adelson, E.: The design and use of steerable filters. PAMI **13** (1991) 891–906
23. Simoncelli, E., Farid, H.: Steerable wedge filters for local orientation analysis. IP **5** (1996) 1377–1382
24. Yu, W., Daniilidis, K., Sommer, G.: Low-cost junction characterization using polar averaging filters. In: ICIP (3). (1998) 41–44