1.2 Overview of the paper

Section 2 presents a state of the art on interest point detectors as well as implementation details for the detectors used in our comparison. Section 3 defines the repeatability criterion, explains how to determine it experimentally and presents the results of a comparison under different transformations. Section 4 describes the information content criterion and evaluates results for different detectors. In section 5 we select the detector which gives the best results according to the two criteria, show that the quality of its results is very high and discuss possible extensions.

2 Interest point detectors

By "interest point" we simply mean any point in the image for which the signal changes two-dimensionally. Conventional "corners" such as L-corners, T-junctions and Y-junctions satisfy this, but so do black dots on white backgrounds, the endings of branches and any location with significant 2D texture. We will use the general term "interest point" unless a more specific type of point is referred to. Figure 1 shows an example of general interest points detected on Van Gogh's sower painting.

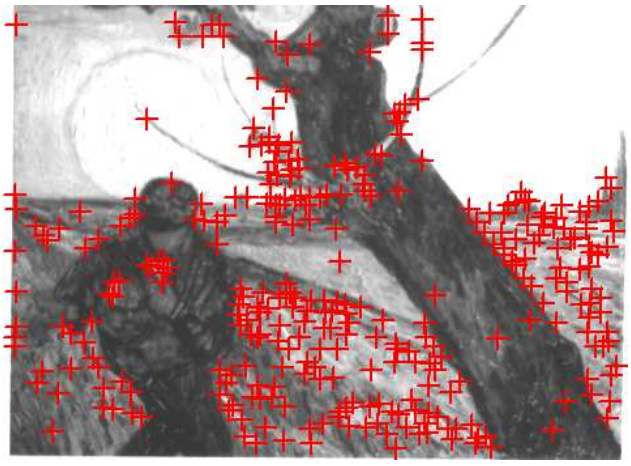

Figure 1: Interest points detected on Van Gogh's sower painting. The detector is an improved version of the Harris detector. There are 317 points detected.

2.1 State of the art

A wide variety of interest point and corner detectors exist in the literature. They can be divided into three categories : contour based, intensity based and parametric model based methods. Contour based methods first extract contours and then search for maximal curvature or inflexion points along the contour chains, or do some polygonal approximation and then search for intersection points. Intensity based methods compute a measure that indicates the presence of an interest point directly from the greyvalues. Parametric model methods fit a parametric intensity model to the signal. They often provide sub-pixel accuracy, but are limited to specific types of interest points, for example to L-corners. In the following we briefly present detection methods for each of the three categories.

2.1.1 Contour based methods

Contour based methods have existed for a long time ; some of the more recent ones are presented. Asada and Brady [1] extract interest points for 2D objects from planar curves. They observe that these curves have special characteristics : the changes in curvature. These changes are classified in several categories : junctions, endings etc. To achieve robust detection, their algorithm is integrated in a multi-scale framework. A similar approach has been developed by Mokhtarian and Mackworth [29]. They use inflexion points of a planar curve.

Medioni and Yasumoto [28] use B-splines to approximate the contours. Interest points are maxima of curvature which are computed from the coefficients of these B-splines.

Horaud et al [23] extract line segments from the image contours. These segments are grouped and intersections of grouped line segments are used as interest points.

Shilat et al [44] first detect ridges and troughs in the images. Interest points are high curvature points along ridges or troughs, or intersection points. They argue that such points are more appropriate for tracking, as they are less likely to lie on the occluding contours of an object.

Mokhtarian and Suomela [30] describe an interest point detector based on two sets of interest points. One set are T-junctions extracted from edge intersections. A second set is obtained using a multi-scale framework : interest points are curvature maxima of contours at a coarse level and

are tracked locally up to the finest level. The two sets are compared and close interest points are merged.

The algorithm of Pikaz and Dinstein [37] is based on a decomposition of noisy digital curves into a minimal number of convex and concave sections. The location of each separation point is optimized, yielding the minimal possible distance between the smoothed approximation and the original curve. The detection of the interest points is based on properties of pairs of sections that are determined in an adaptive manner, rather than on properties of single points that are based on a fixed-size neighborhood.

2.1.2 Intensity based methods

Moravec [31] developed one of the first signal based interest point detectors. His detector is based on the auto-correlation function of the signal. It measures the greyvalue differences between a window and windows shifted in several directions. Four discrete shifts in directions parallel to the rows and columns of the image are used. If the minimum of these four differences is superior to a threshold, an interest point is detected.

The detector of Beaudet [4] uses the second derivatives of the signal for computing the measure "DET" : $DET = I_{xx}I_{yy} - I_{xy}^2$ where $I(x, y)$ is the intensity surface of the image. DET is the determinant of the Hessian matrix and is related to the Gaussian curvature of the signal. This measure is invariant to rotation. Points where this measure is maximal are interest points.

Kitchen and Rosenfeld [24] present an interest point detector which uses the curvature of planar curves. They look for curvature maxima on isophotes of the signal. However, due to image noise an isophote can have an important curvature without corresponding to an interest point, for example on a region with almost uniform greyvalues. Therefore, the curvature is multiplied by the gradient magnitude of the image where non-maximum suppression is applied to the gradient magnitude before multiplication. Their measure is $K = \frac{I_{xx}I_y^2 + I_{yy}I_x^2 - 2I_{xy}I_xI_y}{I_x^2 + I_y^2}$.

Dreschler and Nagel [16] first determine locations of local extrema of the determinant of the Hessian "DET". A location of maximum positive DET can be matched with a location of extreme negative DET, if the directions of the principal curvatures which have opposite sign

are approximatively aligned. The interest point is located between these two points at the zero crossing of DET. Nagel [32] shows that the Dreschler-Nagel's approach and Kitchen-Rosenfeld's approach are identical.

Several interest point detectors [17, 18, 19, 48] are based on a matrix related to the auto-correlation function. This matrix $\mathtt{A}$ averages derivatives of the signal in a window W around a point (x, y) :

$$\mathtt{A}(x,y) = \begin{bmatrix} \sum_{(x_k,y_k)\in W} (I_x(x_k,y_k))^2 & \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) \\ \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) & \sum_{(x_k,y_k)\in W} (I_y(x_k,y_k))^2 \end{bmatrix} \quad (1)$$

where $I(x, y)$ is the image function and (x_k, y_k) are the points in the window W around (x, y).

This matrix captures the structure of the neighborhood. If this matrix is of rank two, that is both of its eigenvalues are large, an interest point is detected. A matrix of rank one indicates an edge and a matrix of rank zero a homogeneous region. The relation between this matrix and the auto-correlation function is given in appendix A.

Harris [19] improves the approach of Moravec by using the auto-correlation matrix $\mathtt{A}$. The use of discrete directions and discrete shifts is thus avoided. Instead of using a simple sum, a Gaussian is used to weight the derivatives inside the window. Interest points are detected if the auto-correlation matrix $\mathtt{A}$ has two significant eigenvalues.

Förstner and Gülch [18] propose a two step procedure for localizing interest points. First points are detected by searching for optimal windows using the auto-correlation matrix $\mathtt{A}$. This detection yields systematic localization errors, for example in the case of L-corners. A second step based on a differential edge intersection approach improves the localization accuracy.

Förstner [17] uses the auto-correlation matrix $\mathtt{A}$ to classify image pixels into categories - region, contour and interest point. Interest points are further classified into junctions or circular features by analyzing the local gradient field. This analysis is also used to determine the interest point location. Local statistics allow a blind estimate of signal-dependent noise variance for automatic selection of thresholds and image restoration.

Tomasi and Kanade [48] motivate their approach in the context of tracking. A good feature

is defined as one that can be tracked well. They show that such a feature is present if the eigenvalues of matrix A are significant.

Heitger et al [21] develop an approach inspired by experiments on the biological visual system. They extract 1D directional characteristics by convolving the image with orientation-selective Gabor like filters. In order to obtain 2D characteristics, they compute the first and second derivatives of the 1D characteristics.

Cooper and al [9] first measure the contour direction locally and then compute image differences along the contour direction. A knowledge of the noise characteristics is used to determine whether the image differences along the contour direction are sufficient to indicate an interest point. Early jump-out tests allow a fast computation of the image differences.

The detector of Reisfeld et al [38] uses the concept of symmetry. They compute a symmetry map which shows a "symmetry strength" for each pixel. This symmetry is computed locally by looking at the magnitude and the direction of the derivatives of neighboring points. Points with high symmetry are selected as interest points.

Smith and Brady [47] compare the brightness of each pixel in a circular mask to the center pixel to define an area that has a similar brightness to the center. Two dimensional features can be detected from the size, centroid and second moment of this area.

The approach proposed by Laganière [26] is based on a variant of the morphological closing operator which successively applies dilation/erosion with different structuring elements. Two closing operators and four structuring elements are used. The first closing operator is sensitive to vertical/horizontal L-corners and the second to diagonal L-corners.

2.1.3 Parametric model based methods

The parametric model used by Rohr [39] is an analytic junction model convolved with a Gaussian. The parameters of the model are adjusted by a minimization method, such that the template is closest to the observed signal. In the case of a L-corner the parameters of the model are the angle of the L-corner, the angle between the symmetry axis of the L-corner and the x-axis, the greyvalues, the position of the point and the amount of blur. Positions obtained by this method

are very precise. However, the quality of the approximation depends on the initial position estimation. Rohr uses an interest point detector which maximizes $\det(\mathtt{A})$ (cf. equation (1)) as well as the intersection of line segments to determine the initial values for the model parameters.

Deriche and Blaszka [14] develop an acceleration of Rohr's method. They substitute an exponential for the Gaussian smoothing function. They also show that to assure convergence the image region has to be quite large. In cluttered images the region is likely to contain several signals, which makes convergence difficult.

Baker et al [2] propose an algorithm that automatically constructs a detector for an arbitrary parametric feature. Each feature is represented as a densely sampled parametric manifold in a low dimensional subspace. A feature is detected, if the projection of the surrounding intensity values in the subspace lies sufficiently close to the feature manifold. Furthermore, during detection the parameters of detected features are recovered using the closest point on the feature manifold.

Parida et al [34] describe a method for general junction detection. A deformable template is used to detect radial partitions. The minimum description length principle determines the optimal number of partitions that best describes the signal.

2.2 Implementation details

This section presents implementation details for the detectors included in our comparison. The detectors are Harris [19], an improved version of Harris, Cottier[10], Horaud [23], Heitger [21] and Förstner [17]. Except in the case of the improved version of Harris, we have used the implementations of the original authors, with the standard default parameter values recommended by the authors for general purpose feature detection. These values are seldom optimal for any given image, but they do well on average on collections of different images. Our goal is to evaluate detectors for such collections.

The standard Harris detector [19] ("Harris") computes the derivatives of the matrix $\mathtt{A}$ (cf. equation (1)) by convolution with the mask [-2 -1 0 1 2]. A Gaussian ($\sigma = 2$) is used to weight the derivatives summed over the window. To avoid the extraction of the eigenvalues of the matrix $\mathtt{A}$, the strength of an interest points is measured by $\det(\mathtt{A}) - \alpha\,\mathrm{trace}(\mathtt{A})^2$. The second term is used

to eliminate contour points with a strong eigenvalue, α is set to 0.06. Non-maximum suppression using a 3x3 mask is then applied to the interest point strength and a threshold is used to select interest points. The threshold is set to 1% of the maximum observed interest point strength.

In the improved version of Harris ("ImpHarris"), derivatives are computed more precisely by replacing the [-2 -1 0 1 2] mask with derivatives of a Gaussian ($\sigma = 1$). A recursive implementation of the Gaussian filters [13] guarantees fast detection.

Cottier [10] applies the Harris detector only to contour points in the image. Derivatives for contour extraction as well as for the Harris detector are computed by convolution with the Canny/Deriche operator [12] ($\alpha = 2$, $\omega = 0.001$). Local maxima detection with hysteresis thresholding is used to extract contours. High and low thresholds are determined from the gradient magnitude (high = average gradient magnitude, low = 0.1 * high). For the Harris detector derivatives are averaged over two different window sizes in order to increase localization accuracy. Points are first detected using a 5x5 window. The exact location is then determined by using a 3x3 window and searching the maximum in the neighborhood of the detected point.

Horaud [23] first extracts contour chains using his implementation of the Canny edge detector. Tangent discontinuities in the chain are located using a worm, and a line fit between the discontinuities is estimated using orthogonal regression. Lines are then grouped and intersections between neighboring lines are used as interest points.

Heitger [21] convolves the image with even and odd symmetrical orientation-selective filters. These Gabor like filters are parameterized by the width of the Gaussian envelope (σ =5), the sweep which increases the relative weight of the negative side-lobes of even filters and the orientation selectivity which defines the sharpness of the orientation tuning. Even and odd filters are computed for 6 orientations. For each orientation an energy map is computed by combining even and odd filter outputs. 2D signal variations are then determined by differentiating each energy map along the respective orientation using "end-stopped operators". Non-maximum suppression (3x3 mask) is applied to the combined end-stopped operator activity and a relative threshold (0.1) is used to select interest points.

The Förstner detector [17] computes the derivatives on the smoothed image ($\sigma = 0.7$). The

derivatives are then summed over a Gaussian window ($\sigma = 2$) to obtain the auto-correlation matrix `A`. The trace of this matrix is used to classify pixels into region or non-region. For homogeneous regions the trace follows approximatively a χ^2-distribution. This allows to determine the classification threshold automatically using a significance level ($\alpha = 0.95$) and the estimated noise variance. Pixels are further classified into contour or interest point using the ratio of the eigenvalues and a fixed threshold (0.3). Interest point locations are then determined by minimizing a function of the local gradient field. The parameter of this function is the size of the Gaussian which is used to compute a weighted sum over the local gradient measures ($\sigma = 4$).

3 Repeatability

3.1 Repeatability criterion

Repeatability signifies that detection is independent of changes in the imaging conditions, i.e. the parameters of the camera, its position relative to the scene, and the illumination conditions. 3D points detected in one image should also be detected at approximately corresponding positions in subsequent ones (cf. figure 2). Given a 3D point X and two projection matrices P_1 and P_i, the projections of X into images I_1 and I_i are $x_1 = P_1 X$ and $x_i = P_i X$. A point x_1 detected in image I_1 is repeated in image I_i if the corresponding point x_i is detected in image I_i. To measure the repeatability, a unique relation between x_1 and x_i has to be established. This is difficult for general 3D scenes, but in the case of a planar scene this relation is defined by a homography [42]:

$$x_i = H_{1i} x_1 \quad \text{where } H_{1i} = P_i P_1^{-1}$$

P_1^{-1} is an abusive notation to represent the back-projection of image I_1. In the case of a planar scene this back-projection exists.

The repeatability rate is defined as the number of points repeated between two images with respect to the total number of detected points. To measure the number of repeated points, we have to take into account that the observed scene parts differ in the presence of changed imaging conditions, such as image rotation or scale change. Interest points which can not be observed in

A Derivation of the auto-correlation matrix

The local auto-correlation function measures the local changes of the signal. This measure is obtained by correlating a patch with its neighbouring patches, that is with patches shifted by a small amount in different directions. In the case of an interest point, the auto-correlation function is high for all shift directions.

Given a shift $(\Delta x, \Delta y)$ and a point (x, y), the auto-correlation function is defined as:

$$f(x,y) = \sum_{(x_k,y_k)\in W} (I(x_k,y_k) - I(x_k+\Delta x, y_k+\Delta y))^2 \tag{5}$$

where (x_k, y_k) are the points in the window W centered on (x, y) and I the image function.

If we want to use this function to detect interest points we have to integrate over all shift directions. Integration over discrete shift directions can be avoided by using the auto-correlation matrix. This matrix is derived using a first-order approximation based on the Taylor expansion:

$$I(x_k+\Delta x, y_k+\Delta y) \approx I(x_k,y_k) + \begin{pmatrix} I_x(x_k,y_k) & I_y(x_k,y_k) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{6}$$

Substituting the above approximation (6) into equation (5), we obtain:

$$\begin{aligned} f(x,y) &= \sum_{(x_k,y_k)\in W} \left(\begin{pmatrix} I_x(x_k,y_k) & I_y(x_k,y_k) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right)^2 \\ &= \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \begin{bmatrix} \sum_{(x_k,y_k)\in W} (I_x(x_k,y_k))^2 & \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) \\ \sum_{(x_k,y_k)\in W} I_x(x_k,y_k) I_y(x_k,y_k) & \sum_{(x_k,y_k)\in W} (I_y(x_k,y_k))^2 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \\ &= \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \mathtt{A}(x,y) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \end{aligned} \tag{7}$$

The above equation (7) shows that the auto-correlation function can be approximated by the matrix $\mathtt{A}(x, y)$. This matrix $\mathtt{A}$ captures the structure of the local neighborhood.