# Detecting Keypoints with Stable Position, Orientation and Scale under Illumination Changes

Bill Triggs

GRAVIR-CNRS-INRIA, 655 avenue de l'Europe, 38330 Montbonnot, France.
*Bill.Triggs@inrialpes.fr* ◇ *http://www.inrialpes.fr/lear/people/triggs*

**Abstract.** Local feature approaches to vision geometry and object recognition are based on selecting and matching sparse sets of visually salient image points, known as 'keypoints' or 'points of interest'. Their performance depends critically on the accuracy and reliability with which corresponding keypoints can be found in subsequent images. Among the many existing keypoint selection criteria, the popular Förstner-Harris approach explicitly targets geometric stability, defining keypoints to be points that have locally maximal self-matching precision under translational least squares template matching. However, many applications require stability in orientation and scale as well as in position. Detecting translational keypoints and verifying orientation/scale behaviour post hoc is suboptimal, and can be misleading when different motion variables interact. We give a more principled formulation, based on extending the Förstner-Harris approach to general motion models and robust template matching. We also incorporate a simple local appearance model to ensure good resistance to the most common illumination variations. We illustrate the resulting methods and quantify their performance on test images.

**Keywords:** keypoint, point of interest, corner detection, feature based vision, Förstner-Harris detector, template matching, vision geometry, object recognition.

## 1 Introduction

Local-feature-based approaches have proven successful in many vision problems, including scene reconstruction [16,5], image indexing and object recognition [20,21,32, 33,23,24,25]. The basic idea is that focusing attention on comparatively sparse sets of especially salient image points — usually called **keypoints** or **points of interest** — both saves computation (as most of the image is discarded) and improves robustness (as there are many simple, redundant local cues rather than a few powerful but complex and delicate global ones) [37]. However, local methods must be able to find 'the same' keypoints again in other images, and their performance depends critically on the reliability and accuracy with which exactly corresponding points can be found. Many approaches to keypoint detection exist, including 'corners' [2,17,38,28,4], parametric image models [3,31,1], local energy / phase congruency [27,29,30,18], and morphology [35,19]. One of the most popular is that developed by Förstner & Gülch [7,9] and Harris & Stephens [15] following earlier work by Hannah [14] and Moravec [26]. This approach brings

To appear in the 2004 European Conference on Computer Vision. 
This research was supported by the European Union FET-Open research project VIBES.
Reference to K. Rohr corrected 24/Jul/2004.

the accuracy issue to the fore by *defining* keypoints to be points at which the predicted precision of local least squares image matching is locally maximal [14, 22, 6, 10, 12, 11]. Notionally, this is implemented by matching the local image patch against itself under small translations, using one of a range of criteria to decide when the 'sharpness' of the resulting correlation peak is locally optimal. Moravec did this by explicit single-pixel translations [26]; Hannah by autocorrelation [14]; and Förstner by implicit least squares matching, using Taylor expansion to re-express the accuracy in terms of the eigenvalues of the **scatter matrix** or **normal matrix** of the local image gradients, $\int \boldsymbol{\nabla} I^\top \boldsymbol{\nabla} I \, \boldsymbol{dx}$ [7, 9, 8]. All of these methods use rectangular patches, usually with a scale significantly larger than that of the image gradients used. This is problematic for patches that contain just one strong feature, because the self-matching accuracy for these is the same wherever the feature is in the patch, *i.e.* the matching-based approach guarantees good self-matching accuracy, but not necessarily accurate *centring* of the patch on a visible feature. Working independently of Förstner, Harris & Stephens improved the localization performance by replacing the rectangular patches with Gaussian windows (convolutions) with a scale similar to that of the derivatives used [15]. With Gaussian-based derivative calculations and more careful attention to aliasing, the method has proven to be one of the most reliable keypoint detectors, especially in cases where there are substantial image rotations, scalings or perspective deformations [33, 24].

One problem with the Förstner-Harris approach is that it optimizes keypoints only for good *translational* precision, whereas many applications need keypoints that are stable not only under translations, but also under rotations, changes of scale, perspective deformations, and changes of illumination (*c.f.* [34]). In particular, many local feature based object recognition / matching methods calculate a vector of local image descriptors at each keypoint, and later try to find keypoints with corresponding descriptors in other images [20, 21, 32, 23, 24, 25]. This usually requires the extraction of a dominant orientation and scale at each keypoint, and keypoints that have poorly defined orientations or scales tend to produce descriptors that vary too much over re-detections to be useful. Hence, it seems useful to develop keypoint detectors that explicitly guarantee good orientation and scale stability, and also good stability under local illumination variations. This is the goal of the current paper, which generalizes the Förstner-Harris self-matching argument to include non-translational motions, and also provides improved resistance to illumination variations by replacing simple least squares matching with an illumination-compensated matching method related to Hager & Belhumeur's [13].

Much of the paper focuses on the low-level task of *characterizing the local stability of matching under geometric transformations and illumination variations*. The Förstner-Harris approach shows that such analysis is a fruitful route to practical keypoint detection in the translational case, and we argue that this continues to hold for more general transformations. Also note the relationship to invariance: if we use image descriptors based at the keypoints for matching, the more invariant the descriptors are to a given type of transformation, the less accurate the keypoint detection needs to be with respect to these transformations. But exactly for this reason, it is useful to develop detectors whose performance under different types of transformations is quantifiable and controllable, and our approach explicitly does this. We adopt the following basic philosophy:

(*i*) *There is no such thing as generic keypoints.* They should be selected specifically for the use to which they will be put, using a purpose-designed detector and parameters.

(*ii*) *Keypoints are not just positions.* Stability in orientation and scale and resistance to common types of appearance variations are also needed.
(*iii*) *Each image (template) matching method defines a corresponding self-matching based keypoint detector.* If the keypoints will be used as correspondence hypotheses that are later verified by inter-image template matching, the keypoint detector and parameters corresponding to the matching method should be used.

**Contents:** §2 describes our matching based framework for keypoint detection. §3 gives some specific examples and implementation details. §4 gives a few experimental results.

**Notation:** $\boldsymbol{x}$ stands for image coordinates, $\boldsymbol{\nabla}$ for $\boldsymbol{x}$-derivatives, $I, R$ for the images being matched (treated as functions of $\boldsymbol{x}$), $\boldsymbol{t}$ for the image motion/warping model, $c$ for the pixel comparison functional. Derivatives are always row vectors, *e.g.* $\delta I \approx \boldsymbol{\nabla} I \, \boldsymbol{\delta x}$. For most of the paper we assume continuous images and ignore sampling issues.

## 2 General Framework

This section develops a general framework for robust image (template) matching under analytical image deformation and appearance variation models, uses it to derive stability estimates for locally optimal matches, and applies this to characterize keypoint stability under self-matching.

**Template matching model:** We will use the following generalized error model for template matching, explained element-by-element below:

$$Q(\boldsymbol{\mu}, \boldsymbol{\lambda}) \;\equiv\; \int c\left(I(\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{\mu}), \boldsymbol{\lambda}), R(\boldsymbol{x}), \boldsymbol{x}\right) \, \boldsymbol{dx} \tag{1}$$

$I$ is the image patch being matched, $R$ is the reference patch it is being matched against, $\boldsymbol{x}$ is a set of 2D image coordinates centred on $R$, and $c \geq 0$ (discussed further below) is a weighted image pixel comparison functional that is integrated over the patch to find the overall matching quality metric $Q$. $\boldsymbol{x}' = \boldsymbol{t}(\boldsymbol{x}, \boldsymbol{\mu})$ is an image motion / warping model that maps $R$'s coordinates $\boldsymbol{x}$ forwards into $I$'s natural coordinate system, *i.e.*, $I$ is effectively being pulled back (warped backwards) into $R$'s frame before being compared. The motion model $\boldsymbol{t}$ is controlled by a vector of **motion parameters** $\boldsymbol{\mu}$ (2D translation, perhaps rotation, scaling, affine deformation…). Before being compared, $I$ may also undergo an optional appearance correction controlled by a vector of **appearance parameters** $\boldsymbol{\lambda}$ (*e.g.*, luminance or colour shifts/rescalings/normalizations, corrections for local illumination gradients…). Note that we think of the input patch $I$ as an ad hoc function $I(\boldsymbol{x}, \boldsymbol{\lambda})$ of both the position and appearance parameters, rather than as a fixed image $I(\boldsymbol{x})$ to which separate appearance corrections are applied. This allows the corrections to be image-content dependent and nonlocal within the patch (*e.g.* subtracting the mean in Zero Mean Cross Correlation). We assume that $\boldsymbol{\mu} = \boldsymbol{0}$ represents a neutral position or reference transformation for the patch (*e.g.* no motion, $\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{0}) = \boldsymbol{x}$). Similarly, $\boldsymbol{\lambda} = \boldsymbol{0}$ represents a default or reference appearance setting (*e.g.* the unchanged input, $I(\boldsymbol{x}, \boldsymbol{0}) = I(\boldsymbol{x})$).

The patch comparison integral is over a spatial window centred on $R$, but for compactness we encode this in the pixel comparison metric $c$. So $c$ usually has the form:

$$c(I(\boldsymbol{x}), R(\boldsymbol{x}), \boldsymbol{x}) \;\equiv\; w(\boldsymbol{x}) \cdot \rho(I(\boldsymbol{x}), R(\boldsymbol{x})) \tag{2}$$

where $w(\boldsymbol{x})$ is a spatial windowing function (rectangular, Gaussian…) that defines the extent of the relevant patch of $R$, and $\rho(I(\boldsymbol{x}), R(\boldsymbol{x}))$ is a spatially-invariant image pixel comparison metric, *e.g.*, the squared pixel difference $\|I(\boldsymbol{x}) - R(\boldsymbol{x})\|^2$ for traditional unweighted least squares matching. The "pixels" here may be greyscale, colour, multiband, or even pre-extracted edge, feature or texture maps, so $\rho()$ can be quite complicated in general, *e.g.* involving nonlinear changes of luminance or colour space, perceptual or sensitivity-based comparison metrics, robust tailing-off at large pixel differences to reduce the influence of outliers, *etc*. Ideally, $\rho()$ should return the negative log likelihood for the pixels to correspond, so that (assuming independent noise in each pixel) $Q$ becomes the total negative log likelihood for the patchwise match. For practical inter-image template matching, the reliability depends critically on the robustness (large difference behaviour) of $\rho()$. But for keypoint detection, we always start from the self-matching case $I{=}R$, so only the *local* behaviour of $\rho()$ near $I{=}R$ is relevant: keypoint detectors are oblivious to large-difference robustification of $\rho()$. We will assume that $\rho()$ has least-squares-like behaviour for small pixel differences, *i.e.* that it is locally differentiable with zero gradient and positive semi-definite Hessian at $I{=}R$, so that:

$$\frac{\boldsymbol{\delta} c}{\boldsymbol{\delta} I(\boldsymbol{x})}\Big|_{I=R} = \boldsymbol{0}, \qquad \frac{\boldsymbol{\delta}^2 c}{\boldsymbol{\delta} I(\boldsymbol{x})^2}\Big|_{I=R} \geq \boldsymbol{0} \tag{3}$$

Our derivations will be based on 2nd order Taylor expansion at $I{=}R$, so they exclude both non-differentiable $L_1$ matching metrics like Sum of Absolute Differences (SAD) and discontinuous $L_0$ (on-off) style ones. Our overall approach probably extends to such metrics, at least when used within a suitable interpolation model, but their abrupt changes and weak resampling behaviour make general derivations difficult.

Finally, we allow $c$ to be a *functional*, not just a function, of $I, R$. (*I.e.* a function of the local patches, not just their pointwise pixel values). In particular, $c$ may run $I, R$ through convolutional filters ('**prefilters**') before comparing them, *e.g.* to restrict attention to a given frequency band in scale-space matching, or simply to suppress high frequencies for reduced aliasing and/or low frequencies for better resistance to global illumination changes. In general, the resampling implied by $\boldsymbol{t}()$ could significantly change $I$'s spatial frequency content, so prefiltering only makes sense if we do it *after* warping. We will thus assume that prefilters run in $\boldsymbol{x}$-space, *i.e.* they are defined relative to the coordinates of the reference image $R$. For example, for affine-invariant keypoint detection [32, 24, 25], keypoint comparison should typically be done, and in particular prefiltering should be applied, in the characteristic affine-normalized frame of the reference keypoint, so $\boldsymbol{x}$ would typically be taken to be the affine-normalized coordinates for $R$. For any $\boldsymbol{t}()$, derivatives of the unwarped input image $I$ can always be converted to derivatives of its prefilter using integration by parts, so the effective scale of derivative masks always ends up being the $\boldsymbol{x}$-space scale of the prefilter.

**Matching precision:** Now suppose that we have already found a locally optimal template match. Consider the behaviour of the matching quality metric $Q$ under small perturbations $I \to I{+}\delta I$. Under 2nd order Taylor expansion:

$$\delta Q \approx \int \left( \frac{\boldsymbol{\delta} c}{\boldsymbol{\delta} I}\, \delta I + \tfrac{1}{2}\, \delta I^\top \frac{\boldsymbol{\delta}^2 c}{\boldsymbol{\delta} I^2}\, \delta I \right)_{\boldsymbol{x}' = \boldsymbol{t}(\boldsymbol{x})} \boldsymbol{dx} \tag{4}$$

For any perturbation of an exact match, $I(\boldsymbol{t}(\boldsymbol{x})) = R(\boldsymbol{x})$, the first order ($\delta I$) term vanishes identically by (3). More generally, if we are already at a local optimum of $Q$ under

some class of perturbations $\delta I$, the integrated first order term vanishes for this class. Both hold for keypoints, so we will ignore the $\delta I$ term from now on.

Using the parametric model $I(\boldsymbol{t}(\boldsymbol{x}, \boldsymbol{\mu}), \boldsymbol{\lambda})$, the image $I$ changes as follows under first order changes of the motion and appearance parameters $\boldsymbol{\mu}, \boldsymbol{\lambda}$:

$$\delta I \;\approx\; \boldsymbol{L}\,\boldsymbol{\delta\lambda} + \boldsymbol{M}\,\boldsymbol{\delta\mu}\,, \quad \text{where} \quad \boldsymbol{L} \;\equiv\; \frac{\boldsymbol{\partial} I}{\boldsymbol{\partial\lambda}}, \qquad \boldsymbol{M} \;\equiv\; \boldsymbol{\nabla} I \cdot \boldsymbol{T}, \qquad \boldsymbol{T} \;\equiv\; \frac{\boldsymbol{\partial t}}{\boldsymbol{\partial\mu}} \tag{5}$$

Here, $\boldsymbol{\nabla} I \equiv \frac{\boldsymbol{\partial} I}{\boldsymbol{\partial t}}(\boldsymbol{t}(\boldsymbol{x}))$ is the standard gradient of the original unwarped image $I$, evaluated in $I$'s own frame at $\boldsymbol{t}(\boldsymbol{x})$. The columns of the Jacobians $\boldsymbol{L}$ and $\boldsymbol{M}$ can be thought of as appearance and motion basis images, characterizing the linearized first-order changes in $I$ as the parameters are varied. Putting (4, 5) together gives a quadratic local cost model for perturbations of the match around the optimum, based on a positive semidefinite **generalized scatter matrix** $\mathbf{S}$ :[1]

$$\delta Q(\boldsymbol{\delta\lambda}, \boldsymbol{\delta\mu}) \;\approx\; \tfrac{1}{2} \begin{pmatrix} \boldsymbol{\delta\lambda}^\top & \boldsymbol{\delta\mu}^\top \end{pmatrix} \mathbf{S} \begin{pmatrix} \boldsymbol{\delta\lambda} \\ \boldsymbol{\delta\mu} \end{pmatrix} \tag{6}$$

$$\mathbf{S} \;\equiv\; \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{pmatrix} \;\equiv\; \int \begin{pmatrix} \boldsymbol{L}^\top \\ \boldsymbol{M}^\top \end{pmatrix} \frac{\boldsymbol{\delta}^2 c}{\boldsymbol{\delta} I^2} \begin{pmatrix} \boldsymbol{L} & \boldsymbol{M} \end{pmatrix} \boldsymbol{dx} \tag{7}$$

$\mathbf{S}$ generalizes the matrix $\int \boldsymbol{\nabla} I^\top \boldsymbol{\nabla} I \; \boldsymbol{dx}$ that appears in the Förstner-Harris keypoint detector (which assumes pure translation, $\boldsymbol{T} = \boldsymbol{I}, \boldsymbol{M} = \boldsymbol{\nabla} I$, quadratic pixel difference metric $\frac{\boldsymbol{\delta}^2 c}{\boldsymbol{\delta} I^2} = \boldsymbol{I}$, and empty illumination model $\boldsymbol{L}$). To the extent that $c$ gives the negative log likelihood for the match, $\mathbf{S}$ is the maximum likelihood saddle point approximation to the Fisher information matrix for estimating $\boldsymbol{\lambda}, \boldsymbol{\mu}$ from the match. *I.e.*, $\mathbf{S}^{-1}$ approximates the covariance with which the parameters $\boldsymbol{\lambda}, \boldsymbol{\mu}$ can be estimated from the given image data: the larger $\mathbf{S}$, the stabler the match, in the sense that the matching error $\delta Q$ increases more rapidly under given perturbations $\boldsymbol{\delta\lambda}, \boldsymbol{\delta\mu}$.

Now suppose that we want to ensure that the two patches match stably *irrespective of appearance changes*. For a given perturbation $\boldsymbol{\delta\mu}$, the appearance change that gives the best match to the original patch — and hence that masks the effect of the motion as well as possible, thus creating the greatest matching uncertainty — can be found by minimizing $\delta Q(\boldsymbol{\delta\mu}, \boldsymbol{\delta\lambda})$ w.r.t. $\boldsymbol{\delta\lambda}$. By inspection from (6), this is $\boldsymbol{\delta\lambda}(\boldsymbol{\delta\mu}) = -\boldsymbol{A}^{-1}\boldsymbol{B}\,\boldsymbol{\delta\mu}$. Back-substituting into (6) gives an effective quadratic **reduced penalty function** $\delta Q_{\text{red}}(\boldsymbol{\delta\mu}) \;\equiv\; \delta Q(\boldsymbol{\delta\mu}, \boldsymbol{\delta\lambda}(\boldsymbol{\delta\mu})) \;\approx\; \frac{1}{2}\,\boldsymbol{\delta\mu}^\top \boldsymbol{C}_{\text{red}}\,\boldsymbol{\delta\mu}$ characterizing motion-with-best-appearance-adaptation, where the **reduced scatter matrix** is

$$\boldsymbol{C}_{\text{red}} \;\equiv\; \boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B} \tag{8}$$

with $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ as in (7). $\boldsymbol{C}_{\text{red}}$ and $\boldsymbol{C}$ quantify the precision of motion estimation respectively with and without appearance adaptation. Some precision is always lost by factoring out appearance, so $\boldsymbol{C}_{\text{red}}$ is always smaller than $\boldsymbol{C}$. To the extent that the matching error metric

[1] Strictly, to be correct to $\mathcal{O}\big((\boldsymbol{\delta\mu}, \boldsymbol{\delta\lambda})^2\big)$ we should also expand (5) to $2^{\text{nd}}$ order, which introduces a $2^{\text{nd}}$ order 'tensor' correction in the $\delta I$ term of (4). But, as above by (3), the latter term vanishes identically for keypoint detection. Even for more general matching, the correction is usually negligible unless the match is poor *and* the motion / appearance models are very nonlinear. One can think of (7) as a Gauss-Newton approximation to the true $\mathbf{S}$. It guarantees that $\mathbf{S}$ is at least positive semidefinite (as it must be at a locally optimal match). We will adopt it from now on.

$c$ is a statistically valid log likelihood model for image noise, $\boldsymbol{C}^{-1}$ and $\boldsymbol{C}_{\text{red}}^{-1}$ estimate the covariances of the corresponding motion parameter estimates under trials with independent noise samples. More generally, if we also have prior information that appearance variations are not arbitrary, but have zero mean and covariance $\boldsymbol{D}^{-1}$, the optimal $\boldsymbol{\delta\lambda}(\boldsymbol{\delta\mu})$ becomes $-(\boldsymbol{A}+\boldsymbol{D})^{-1}\boldsymbol{B}\,\boldsymbol{\delta\mu}$ and $\boldsymbol{C}_{\text{red}}$ is replaced by the less strongly reduced covariance $\boldsymbol{C}'_{\text{red}} \equiv \boldsymbol{C} - \boldsymbol{B}^{\top}(\boldsymbol{A}+\boldsymbol{D})^{-1}\boldsymbol{B}$.

**Keypoint detection:** Ideally, we want to find keypoints that can be *stably* and *reliably* re-detected under arbitrary motions from the given transformation family $\boldsymbol{t}(\boldsymbol{x},\boldsymbol{\mu})$, despite arbitrary changes of appearance from the appearance family $I(\boldsymbol{x},\boldsymbol{\lambda})$. We focus on the 'stability' aspect[2], which we characterize in terms of the *precision of self-matching* under our robust template matching model. The idea is that the patch itself is its own best template — if it can not be matched stably even against itself, it is unlikely to be stably matchable against other patches. We are interested in stability despite appearance changes, so we use the reduced scatter matrix $\boldsymbol{C}_{\text{red}}$ (8) to quantify geometric precision.

The amount of precision that is needed depends on the task, and we adopt the design philosophy that visual routines should be explicitly parametrized in terms of objective performance criteria such as output accuracy. To achieve this we require keypoints to meet a lower bound on matching precision (equivalently, an upper bound on matching uncertainty). We quantify this by introducing a user-specified **criterion matrix** $\boldsymbol{C}_0$ and requiring keypoints to have reduced precisions $\boldsymbol{C}_{\text{red}}$ greater than $\boldsymbol{C}_0$ (*i.e.* $\boldsymbol{C}_{\text{red}} - \boldsymbol{C}_0$ must be positive semidefinite). Intuitively, this means that for a keypoint candidate to be accepted, its transformation-space motion-estimation uncertainty ellipse $\boldsymbol{C}_{\text{red}}^{-1}$ must be strictly contained within the criterion ellipse $\boldsymbol{C}_0^{-1}$.

In textured images there may be whole regions where this precision criterion is met, so for isolated keypoint detection we must also specify a means of selecting 'the best' keypoint(s) within these regions. This requires some kind of 'saliency' or 'interest' metric, ideally an index of perceptual distinctiveness / reliable matchability modulo our appearance model. But here, following the Förstner-Harris philosophy, we simply use an index of overall matching precision as a crude substitute for this. In the translation-only case, Förstner [7,9] and Harris & Stephens [15] discuss several suitable precision indices, based on the determinant, trace and eigenvalues of the scatter matrix. In our case, there may be several (more than 2) motion parameters, and eigenvalue based criteria seem more appropriate than determinant based ones, owing to their clear links with uncertainty analysis. Different motion parameters also have different units (translations in pixels, rotations in radians, dilations in log units), and we need to normalize for this. The criterion matrix $\boldsymbol{C}_0$ provides a natural scaling, so as our final saliency criterion we will take the *minimum eigenvalue of the normalized reduced motion precision matrix* $\boldsymbol{C}_0^{-1/2}\,\boldsymbol{C}_{\text{red}}\,\boldsymbol{C}_0^{-1/2}$. Intuitively, this requires the longest axis of the motion-estimation covariance ellipse, as measured in a frame in which $\boldsymbol{C}_0$ becomes spherical, to be as small as possible. With this normalization, the keypoint-acceptability criterion $\boldsymbol{C}_{\text{red}} > \boldsymbol{C}_0$ simplifies to the requirement that the saliency (the minimum eigenvalue) must be greater than one. Typically, $\boldsymbol{C}_0$ is diagonal, in which case the normalization matrix $\boldsymbol{C}_0^{-1/2}$ is the

[2] We do not consider other matchability properties [7] such distinctiveness here, as this is more a matter for the descriptors calculated once the keypoint is found. Distinctiveness is usually characterized by probability of mismatch within a population of extracted keypoints (*e.g.* [33]). For a recent entropic approach to image-wide distinctiveness, see [36].

diagonal matrix of maximum user-permissible standard errors in translation, rotation and scale.

As usual, pixel sampling effects introduce a small amount of aliasing or jitter in the image derivative estimates, which has the effect of spreading gradient energy across the various eigenvalues of $\mathbf{S}$ even when the underlying image signal is varies only in one dimension (*e.g.* a straight edge). As in the Förstner-Harris case, we compensate for this heuristically by subtracting a small user-specified multiple $\alpha$ of the maximum eigenvalue of $\boldsymbol{C}_0^{-1/2}\,\boldsymbol{C}_{\text{red}}\,\boldsymbol{C}_0^{-1/2}$ (the 1-D 'straight edge' signal) before testing for threshold and saliency, so our final keypoint saliency measure is $\lambda_{\text{min}} - \alpha\,\lambda_{\text{max}}$.

In practice, the Schur complement in $\boldsymbol{C}_{\text{red}} = \boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B}$ is calculated simply and efficiently by outer-product based partial Cholesky decomposition. A standard symmetric eigendecomposition method is then used to calculate the minimum eigenvalue, except that 2D eigenproblems are handled as a special case for speed.

## 3 Examples of Keypoint Detectors

Given the above framework, it is straightforward to derive keypoint detectors for specific pixel types and motion and appearance models. Here we only consider the simplest few motion and appearance models, and we assume greyscale images.

**Comparison function:** As in the traditional Harris detector, we will use simple squared pixel difference to compare pixels, and a circular Gaussian spatial integration window. So modulo prefiltering, $\frac{\boldsymbol{\delta}^2 c}{\boldsymbol{\delta} I^2}$ in (7) reduces to simple weighting by the window function.

**Affine deformations:** For keypoints, only local deformations are relevant, so the most general motion model that is useful is probably the affine one. We will use various subsets of this, parametrizing affine motions linearly as $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{T}\,\boldsymbol{\mu}$ where:

$$\boldsymbol{T}\,\boldsymbol{\mu} = \begin{pmatrix} 1 & 0 & -y & x & x & y \\ 0 & 1 & x & y & -y & x \end{pmatrix} \begin{pmatrix} u \\ v \\ r \\ s \\ a \\ b \end{pmatrix} = \begin{pmatrix} 1{+}s{+}a & -r{+}b \\ r{+}b & 1{+}s{-}a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} u \\ v \end{pmatrix} \quad (9)$$

Here, $(x, y)$ are window-centred pixel coordinates, $(u, v)$ is the translation, $s$ the scale, and for small motions, $r$ is the rotation and $a, b$ are axis- and $45°$-aligned quadrupole deformations. The resulting $\boldsymbol{M}$ matrix is as follows, where $\boldsymbol{\nabla} I = (I_x, I_y)$:

$$\boldsymbol{M} = \begin{pmatrix} I_x & I_y & -yI_x{+}xI_y & xI_x{+}yI_y & xI_x{-}yI_y & yI_x{+}xI_y \end{pmatrix} \quad (10)$$

If the input image is being prefiltered (which, as discussed, must happen *after* warping, *i.e.* after (10)), we can integrate by parts to reduce the prefiltered $\boldsymbol{M}$ vector to the form:

$$\boldsymbol{M}^p = \begin{pmatrix} I^p_x,\ I^p_y,\ -(yI)^p_x{+}(xI)^p_y,\ (xI)^p_x{+}(yI)^p_y{-}2I^p,\ (xI)^p_x{-}(yI)^p_y,\ (yI)^p_x{+}(xI)^p_y \end{pmatrix} \quad (11)$$

where $I^p \equiv p * I$, $(xI)^p_y \equiv p_y * (xI)$, *etc.*, denote convolutions of $I$, $xI$, *etc.*, against the prefilter $p$ and its derivatives $p_x, p_y$. The $-2I^p$ term in the $s$ entry corrects for the fact that prefiltering should happen after any infinitessimal scale change coded by $\boldsymbol{M}$: without this, we would effectively be comparing patches taken at different derivative

scales, and would thus overestimate the scale localization accuracy. If $p$ is a Gaussian of width $\sigma$, we can use (10) or (11) and the corresponding identities $(xI)^p = xI^p + \sigma^2 I^p_x$ or $(xI)^p_x = x\,I^p_x + \sigma^2 I^p_{xx} + I^p$ (from $(x{-}x')g(x{-}x') = -\sigma^2 g_x(x{-}x')$, *etc.*) to move $x, y$ outside the convolutions, reducing $\boldsymbol{M}^p$ to:

$$\big(I^p_x,\, I^p_y,\, -yI^p_x{+}xI^p_y,\, xI^p_x{+}yI^p_y{+}\sigma^2 I^p_{xx+yy},\, xI^p_x{-}yI^p_y{+}\sigma^2 I^p_{xx-yy},\, yI^p_x{+}xI^p_y + 2\sigma^2 I^p_{xy}\big) \tag{12}$$

**Appearance model:** Class-specific appearance models like [1, 13] can include elaborate models of appearance variation, but for generic keypoint detection we can only use simple generic models designed to improve resistance to common types of local illumination variations. Here, we allow for (at most) a scalar illumination shift, addition of a constant spatial illumination gradient, and illumination rescaling. So our linear appearance model is $I{+}\boldsymbol{L}\,\boldsymbol{\lambda}$ where $\boldsymbol{L}(\boldsymbol{x})$ is a subset of:

$$\boldsymbol{L}(\boldsymbol{x}) \;=\; \begin{pmatrix} 1 & x & y & I(\boldsymbol{x}) \end{pmatrix} \tag{13}$$

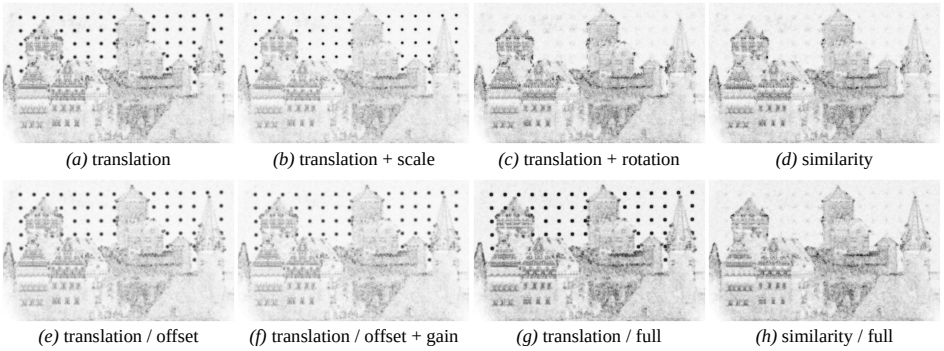
As with $\boldsymbol{M}$, the elements of $\boldsymbol{L}$ must be prefiltered, but $I$ is just smoothed to $I^p$ and $1, x, y$ typically have trivial convolutions (*e.g.*, they are unchanged under Gaussian smoothing, and hence generate a constant diagonal block $\mathrm{diag}(1, \sigma_w^2, \sigma_w^2)$ in $\mathbf{S}$).

**Putting it all together:** The main stages of keypoint detection are: (*i*) prefilter the input image to produce the smoothed image and derivative estimates $I^p, I^p_x, I^p_y, I^p_{xx}, I^p_{xy}, I^p_{yy}$ needed for (12, 13); (*ii*) for each keypoint location $\boldsymbol{x}$, form the outer product matrix of the (desired components of the) combined L/M vector at all pixels in its window, and sum over the window to produce the scatter matrix $\mathbf{S}(\boldsymbol{x})$ (7) (use window-centred coordinates for $x, y$ in (12, 13)); (*iii*) at each $\boldsymbol{x}$, reduce $\mathbf{S}(\boldsymbol{x})$ to find $\boldsymbol{C}_{\mathrm{red}}(\boldsymbol{x})$, normalize by $\boldsymbol{C}_0$, and find the smallest eigenvalue (saliency). Keypoints are declared at points where the saliency has a dominant local maximum, *i.e.* is above threshold and larger than at all other points within a suitable non-maximum-suppression radius. For multiscale detection, processing is done within a pyramid and keypoints must be maxima in both position and scale. As usual, one can estimate subpixel keypoint location and scale by quadratic interpolation of the saliency field near its maximum. But note that, as in the standard Förstner-Harris approach, keypoints do not necessarily contain nameable features (corners, spots) that clearly mark their centres — they may just be unstructured patches with locally maximal matching stability[3].

When calculating $\mathbf{S}$, instead of separate *ab initio* summation over each integration window, one can also use image-wide convolution of quadratic 'energies' as in the standard Förstner-Harris detector, but for the more complicated detectors there are many such maps to be calculated (76 for the full 10-entry L/M model). See the extended version of this paper for details.

In our current implementation, run times for the full 10-L/M-variable detector (which is more than one would normally use in practice) are a factor of about 10 larger than for the original two variable Förstner-Harris detector.

[3] If well-localized centres are needed, specialized locators exist for specific image structures such as spots and corners (*e.g.* [8]), or more generally one could search for *sharp* (high-curvature) and preferably *isolated* maxima of the minimum eigenvalue field or local saliency measure, not just for *high* (but possibly broad) ones. For example, a minimum acceptable peak curvature could be specified via a second criterion matrix.



*(a)* translation *(b)* translation + scale *(c)* translation + rotation *(d)* similarity

*(e)* translation / offset *(f)* translation / offset + gain *(g)* translation / full *(h)* similarity / full

**Fig. 1.** Minimum-eigenvalue strength maps for a popular test image under various motion and illumination models. The saliency differences are much larger than they seem: the maps have been very strongly gamma compressed, normalized and inverted for better visibility. The prefilter and integration windows had $\sigma$=1 pixel, and $\alpha = 0$. Criterion standard deviations were 1 pixel in translation, 1 radian in rotation, $\sqrt{2}$ in scale, but these values are not critical.

**Relation to Zero Mean Matching:** This common matching method compares two image patches by first subtracting each patches mean intensity, then summing the resulting squared pixel differences. We can relate this to the simplest nonempty illumination correction model, $\boldsymbol{L}=\begin{pmatrix}1\end{pmatrix}$, whose reduced scatter matrix over window $w(\boldsymbol{x})$ is:

$$\begin{aligned} \boldsymbol{C}_{\text{red}} \;&=\; \int w\,\boldsymbol{M}^{\top}\boldsymbol{M}\,\boldsymbol{dx} - \overline{\boldsymbol{M}}^{\top}\overline{\boldsymbol{M}} \;=\; \int w\,(\boldsymbol{M}-\overline{\boldsymbol{M}})^{\top}(\boldsymbol{M}-\overline{\boldsymbol{M}})\,\boldsymbol{dx} \\ \overline{\boldsymbol{M}} \;&\equiv\; \textstyle\int w\,(\boldsymbol{M})\,\boldsymbol{dx} \;/\; \left(\int w\,\boldsymbol{dx}\right)^{1/2} \end{aligned} \tag{14}$$

For the translation-only model, $\boldsymbol{T}$ is trivial, so the illumination correction simply has the effect of subtracting from each image gradient its patch mean (*c.f.* (10)). If $w$ changes much more slowly than $I$, $\overline{\boldsymbol{\nabla} I} \approx \boldsymbol{\nabla}\overline{I}$ and hence $\boldsymbol{\nabla} I - \overline{\boldsymbol{\nabla} I} \approx \boldsymbol{\nabla}(I - \overline{I})$, so this is approximately the same as using the gradient of the bandpassed image $I - \overline{I}$. The standard Förstner-Harris detector embodies least squares matching, not zero mean matching. It is invariant to constant illumination shifts, but it does not subtract the gradient of the mean $\boldsymbol{\nabla}\overline{I}$ (or more correctly, the mean of the gradient $\overline{\boldsymbol{\nabla} I}$) to discount the effects of smooth local illumination gradients superimposed on the pattern being matched. It thus systematically overestimates the geometric strength of keypoints in regions with strong illumination gradients, *e.g.* near the borders of smoothly shaded objects, or at the edges of shadows.

## 4 Experiments

Fig. 1 shows that the saliency (minimum eigenvalue) map emphasizes different kinds of image structures as the motion and illumination models are changed. Image $(a)$ is the original Förstner-Harris detector. Images $(b), (c), (d)$ successively add scale, rotation and scale + rotation motions, while images $(e), (f), (g)$ adjust for illumination offset, offset + gain, and offset + gain + spatial gradients. Note the dramatic extent to which

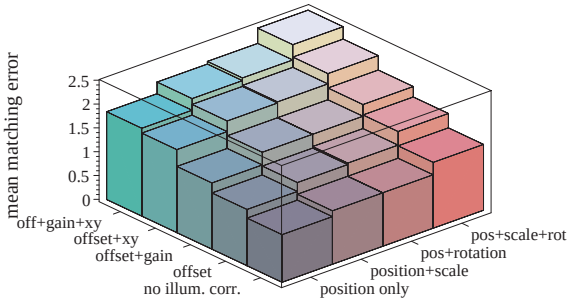



**Fig. 2.** Mean predicted standard error (inverse square root of saliency / minimum eigenvalue in normalized units) for template matching of keypoints under our motion and lighting models, for the model's top 100 keypoints on the Summer Palace image in fig. 3.

enforcing rotational stability in $(a)\rightarrow(c)$ and $(b)\rightarrow(d)$ eliminates the circular dots of the calibration pattern. In comparison, enforcing scale stability in $(a)\rightarrow(b)$ and $(c)\rightarrow(d)$ has more subtle effects, but note the general relative weakening of the points at the summits of the towers between $(a)$ and $(b)$: straight-edged 'corners' are scale invariant, and are therefore suppressed. Unfortunately, although ideal axis- and $45°$-aligned corners are strongly suppressed, it seems that aliasing and blurring effects destroy much of the notional scale invariance of most other rectilinear corners, both in real images and in non-axis-aligned ideal ones. We are currently working on this problem, which also reduces the cross-scale performance of the standard Förstner-Harris detector.

Adding illumination invariance seems to have a relatively small effect in this example, but note the general relative sharpening caused by including $x$ and $y$ illumination gradients in $(a), (e), (f)\rightarrow(g)$. Points on the borders of intensity edges have enhanced gradients owing to the slope alone, and this tends to make them fire preferentially despite the use of the minimum-eigenvalue (most uncertain direction) criterion. Subtracting the mean local intensity gradient reduces this and hence sharpens the results. However a negative side effect of including $x, y$ gradients is that locally quadratic image patches — in particular small dots and ridge edges — become much less well localized, as adding a slope to a quadratic is equivalent to translating it.

Allowing more general motions and/or quotienting out illumination variations always reduces the precision of template matching. Fig. 2 shows the extent of this effect by plotting the relative standard errors of template matching for our complete set of motion and lighting models, where the matching for each model is performed on the model's own keypoints. There is a gradual increase in uncertainty as parameters are added, the final uncertainty for a similarity transform modulo the full illumination model being about 2.5 times that of the original translation-only detector with no illumination correction.

Fig. 3 shows some examples of keypoints selected using the various different motion/lighting models. The main observation is that different models often select different keypoints, and more invariant models generate fewer of them, but beyond this it is difficult to find easily interpretable systematic trends. As in the Förstner-Harris case, keypoints are optimized for matching precision, not for easy interpretability in terms of idealized image events.

## 5 Summary and Conclusions

**Summary:** We have generalized the Förstner-Harris detector [7,9,15] to select keypoints that provide repeatable scale and orientation, as well as repeatable position, over re-detections, even in the face of simple local illumination changes. Keypoints are selected to maximize a minimum-eigenvalue-based local stability criterion obtained from a second order analysis of patch self-matching precision under affine image deformations, compensated for linear illumination changes.

**Future work:** The approach given here ensures accurate re-localizability (by inter-image template matching) of keypoint image patches under various transformations, but it does not always provide accurate 'centres' for them. To improve this, we would like to characterize the stability and localization accuracy of the local maxima of the saliency measure (minimum eigenvalue) under the given transformations. In other words, just as we derived the local transformational-stability matrix $\boldsymbol{C}_{\text{red}}(\boldsymbol{x})$ for *matching* from the scalar matching metric $Q(\boldsymbol{x})$, we need to derive a local transformational-stability matrix for *saliency* from the scalar saliency metric. Only here, the saliency measure is already based on matching stability, so a second level of analysis will be needed.

### References


[1] S. Baker, S. Nayar, and H. Murase. Parametric feature detection. *Int. J. Computer Vision*, 27(1):27–50, 1998.

[2] P.R. Beaudet. Rotationally invariant image operators. In *Int. Conf. Pattern Recognition*, pages 579–583, 1978.

[3] R. Deriche and T. Blaszka. Recovering and characterizing image features using an efficient model based approach. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 530–535, 1993.

[4] R. Deriche and G. Giraudon. A computational approach for corner and vertex detection. *Int. J. Computer Vision*, 10(2):101–124, 1993.

[5] O. Faugeras, Q-T. Luong, and T. Papadopoulo. *The Geometry of Multiple Images*. MIT Press, 2001.

[6] W. Förstner. On the geometric precision of digital correlation. *Int. Arch. Photogrammetry & Remote Sensing*, 24(3):176–189, 1982.

[7] W. Förstner. A feature-based correspondence algorithm for image matching. *Int. Arch. Photogrammetry & Remote Sensing*, 26 (3/3):150–166, 1986.

[8] W. Förstner. A framework for low-level feature extraction. In *European Conf. Computer Vision*, pages II 383–394, Stockholm, 1994.

[9] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop*, Interlaken, June 1987.

[10] A. Grün. Adaptive least squares correlation — concept and first results. Intermediate Research Report to Helava Associates, Ohio State University. 13 pages, March 1984.

[11] A. Grün. Least squares matching: A fundamental measurement algorithm. In *Close Range Photogrammetry and Machine Vision*, pages 217–255. Whittles Publishing, Caithness, Scotland, 1996.

[12] A. Grün and E.P. Baltsavias. Adaptive least squares correlation with geometrical constraints. In *SPIE Computer Vision for Robots*, volume 595, pages 72–82, Cannes, 1985.

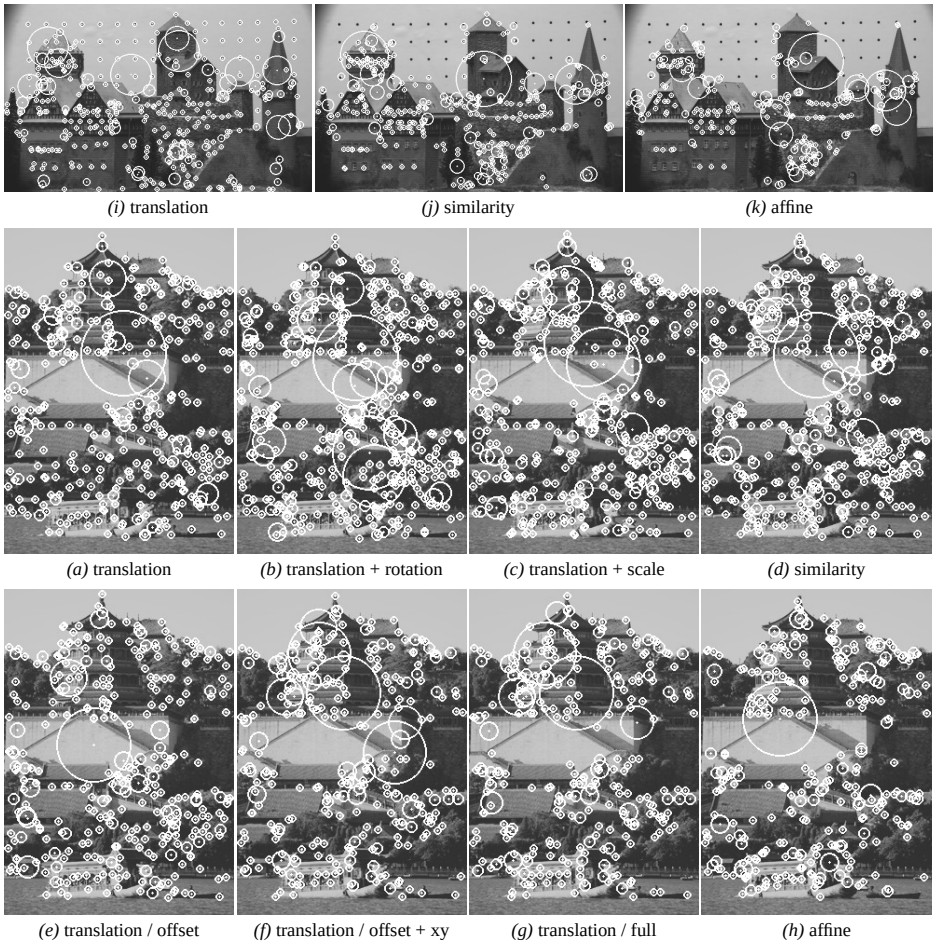



**Fig. 3.** Examples of keypoints from the CMU and Summer Palace (Beijing) test images, under various motion and illumination models. The prefilter and integration windows had $\sigma$=2 pixels, $\alpha = 0$, and non-maximum suppression within 4 pixels radius and scale factor 1.8 was applied. Note that, *e.g.*, 'affine' means 'resistant to small affine deformations', not affine invariant in the sense of [32, 24, 25].

[13] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 20(10):1025–1039, 1998.

[14] M.J. Hannah. *Computer Matching of Areas in Stereo Images*. Ph.D. Thesis, Stanford University, 1974. AIM Memo 219.

[15] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[16] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[17] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Patt. Rec. Lett.*, 1:95–102, 1982.

[18] P. Kovesi. Image features from phase congruency. *Videre: A Journal of Computer Vision Research*, 1(3), 1999.

[19] R. Laganière. Morphological corner detection. In *Int. Conf. Computer Vision*, pages 280–285, 1998.

[20] D. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. Computer Vision*, pages 1150–1157, 1999.

[21] D. Lowe. Local feature view clustering for 3d object recognition. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 682–688, 2001.

[22] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[23] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Int. Conf. Computer Vision*, pages 525–531, 2001.

[24] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conf. Computer Vision*, pages I.128–142, 2002.

[25] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Int. Conf. Computer Vision & Pattern Recognition*, 2003.

[26] H.P. Moravec. Towards automatic visual obstacle avoidance. In *IJCAI*, page 584, 1977.

[27] M. C. Morrone and R. A. Owens. Feature detection from local energy. *Patt. Rec. Lett.*, 6:303–313, 1987.

[28] J.A. Noble. Finding corners. *Image & Vision Computing*, 6(2):121–128, 1988.

[29] D. Reisfeld. The constrained phase congruency feature detector: Simultaneous localization, classification, and scale determination. *Patt. Rec. Lett.*, 17:1161–1169, 1996.

[30] B. Robbins and R. Owens. 2d feature detection via local energy. *Image & Vision Computing*, 15:353–368, 1997.

[31] K. Rohr. Recognizing corners by fitting parametric models. *Int. J. Computer Vision*, 9(3):213–230, 1992.

[32] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *Int. Conf. Computer Vision*, pages 636–643, Vancouver, 2001.

[33] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Computer Vision*, 37(2):151–172, 2000.

[34] J. Shi and C. Tomasi. Good features to track. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 593–600, Seattle, 1994.

[35] S. M. Smith and J. M. Brady. SUSAN - a new approach to low level image processing. *Int. J. Computer Vision*, 23(1):45–78, 1997.

[36] M. Toews and T. Arbel. Entropy-of-likelihood feature selection for image correspondence. In *Int. Conf. Computer Vision*, pages 1041–1047, Nice, France, 2003.

[37] P. H. S. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 278–294, Corfu, Greece, 2000. Springer-Verlag LNCS.

[38] O. Zuniga and R. Haralick. Corner detection using the facet model. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 30–37, 1983.