Evaluating keypoint detectors

Wolfgang Förstner

16. 05. 2019

1 Problem

Given is a keypoint detector, leading to a list of points with a covariance matrices

$$\mathcal{X} = \{(\boldsymbol{x}_i, \Sigma_{x_i x_i}), i = 1, ..., I\}\,. \tag{1}$$

e.g., derived from the structure tensor depending on the pixel noise and the image function $g(\boldsymbol{x})$ in the neighbourhood $\mathcal{N}(\boldsymbol{x}_i)$ of a point $\boldsymbol{x}_i$

$$\Sigma_{x_i x_i} = \sigma_{n_i}^2\, T_i^{-1} \quad \text{with} \quad T_i = \sum_{j \in \mathcal{N}(\boldsymbol{x}_i)} \nabla g \nabla^\mathsf{T} g\,. \tag{2}$$

Given is a pair of images $\{\mathcal{A}(\boldsymbol{x}), \mathcal{B}(\boldsymbol{y})\}$ with known geometric relation between corresponding positions $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{G}) = \boldsymbol{0} \tag{3}$$

e.g., for a homography $\mathcal{G} = \mathcal{H}(\mathsf{H})$ applied to non-homogeneous coordinates $\boldsymbol{x} = \mathsf{N}^e(\mathbf{x})$

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{H}(\mathsf{H})) = \boldsymbol{y} - \mathcal{H}(\boldsymbol{d}) = \boldsymbol{0} \quad \text{with} \quad \mathcal{H}(\boldsymbol{x}) = \mathsf{N}^e(\mathsf{H}\mathbf{x})\,. \tag{4}$$

For two images of a 3D scene the criterium would need to also know the 3D geometry $\mathcal{S}$ of the scene in order to check for correspondence, then the relation would be $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}; \mathcal{S}) = \boldsymbol{0}$. If no confusion is to be expected, we omit the given parameters in the function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$.

The task is to evaluate the keypoint detector by comparing the two list $\mathcal{X}$ and $\mathcal{Y}$ of the keypoints in the two images w.r.t. the known geometric relation $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$.

2 Criteria

2.1 Using LSM Covariance Matrices as Proxy

If we want to evaluate the similarity of two corresponding windows the idea of this test is to use some descriptor: Hence this is the idea of the test to evaluate descriptors.

However, we want to evaluate detectors: $I -> L_I^D \boldsymbol{x}_i^D$, which map an image I to a list $\mathcal{L}_I^D$ of keypoint positions $\boldsymbol{x}_i^D = [x_i, y_i]^D$ with some detector D. This can be done by applying the detector to the same image pair $\{I, J\}$, where we know the mutual homography. I.e., if two keypoints dx_i^D and $\boldsymbol{x}_j^D$ correspond and are derived with the same detector (so I omit the D in the following), they should be related by

$$\boldsymbol{x}_j = \mathsf{H}_{ij} \circ \boldsymbol{x}_i \tag{5}$$

(assuming the coordinates here still are non-homogeneous coordinates, i.e. the homography applied to a point $\boldsymbol{x}_i$ and the resulting displaced coordinates $\mathsf{H}_{ij} \circ \boldsymbol{x}_i$ are given in non-homogeneous manner.)

Since the positions of the keypoints are noisy, we need to evaluate the vector-valued difference

$$\boldsymbol{d}_{x_{ij}} = \boldsymbol{x}_j - \mathsf{H}_{ij} \circ \boldsymbol{x}_i \,. \tag{6}$$

The most simple way would use the absolute value

$$d_{x_{ij}} = ||\boldsymbol{d}_{x_{ij}}|| \quad \text{or} \quad d_{x_{ij}}^2 = ||\boldsymbol{d}_{x_{ij}}||^2 \,. \tag{7}$$

If this (Euclidean) distance $||d_{x_{ij}}||$ is small, then we have an argument, that the two points are corresponding (though we cannot be sure); if the distance is large, we have an argument, that the two keypoints do not correspond, though we cannot be sure either.

Moreover, it may be, that one point i may be close to several j. The we might clean the bi-bartite graph with edges $\{i, j\}$ possibly referring to the same point i or j (I have an efficient routine for this, found on the internet)

So we could measure how many - possibly only unique - distances $d_{x_{ij}}$ of an image pair are below a threshold T_d, which could be treated as putative correspondences (therefore the index c)

$$N_c(T_d) = \#ij || d_{x_{ij}} || < T_d \,. \tag{8}$$

We might report $N_c(T_d)$ as a function of T_d, which gives insight on the distribution of the d's, see the discussion below.

Now, up to this point we did not take into account, whether the keypoint promises a precise match or an imprecise match, thus treat all keypoint alike.

However, if the window around a keypoint shows high texture, we would expect, that we could be able to locate it in the other image quite precisely: hence then the distance d of this correspondence is expected to be small: On the other hand; if the image window around the keypoint shows only weak detail or is blurred, then we would expect the distance between corresponding keypoints is large.

Therefore applying a fixed threshold T_d to d would certainly not be appropriate, falsely eliminating correspondences which are weak and falsely accepting correspondences which are precise.

Therefore we could argue the following: if we would know, the expected precision of the point transfer, ie. the precision of the distance d, then we could

- either apply a constant to a modified distance $t_{x_{ij}}$, which takes the expected precision into account or
- apply an individual threshold T_{ij} to each $d_{x_{ij}}$.

Now, we in general do not know how precise the keypoint detectors are. But it appears reasonable to relate the precision of the distance to

1. the precision, i.e. covariance matrix $\Sigma_{x_i x_i}$ of $\boldsymbol{x}_i$ of detecting a point
2. the precision, i.e. covariance matrix $\Sigma_{x_j x_j}$ of $\boldsymbol{x}_j$ of detecting a point, and
3. the precision, i.e. covariance matrix Σ_{HH} of the homography.

These covariance matrices will be related to the image window around the keypoints.

The idea now is, to use the CovM of the expected precision of least squares matching, as proxy for the CovM of the keypoint detector. Of course, in a first step we then have no easy possibility to balance the result of different detectors, which may lead to keypoints of different precision.

However, we may reasonably assume that the CovM resulting from detector D is proportional to the CovM of LSM, using an assumption on the noise variance and the inverse structure tensor of a window of adequate diameter: hence with some unknown detector specific variance factor σ_D^2

$$\Sigma_{x_i x_i} = \sigma_D^2 \, \Sigma_{x_i x_i}^{LSM} \tag{9}$$

This factor could be derived easily by estimating the homography for all putative (inlier) matches, namely from the reprojection errors $\boldsymbol{e}_{ij}^D$, which are the I discrepancies in x- and y-direction after estimating the optimal homography. Then we have an estimate for the variance factor

$$\sigma_D^2 = \sum_{ij} (\boldsymbol{e}_{ij}^D)^\mathsf{T} \, (\Sigma_{x_i x_i}^{LSM})^{-1} \, \boldsymbol{e}_{ij}^D \; / \; (2I - 8) \tag{10}$$

Without going into details of the variance propagation, we then would be able to determine the Mahalanobis distance of the point $\boldsymbol{x}_j$ and the predicted point $\mathsf{H}_{ij} \circ \boldsymbol{x}_i$

$$t_{ij}^{2,D} = \boldsymbol{d}_{x_{ij}}^\mathsf{T} \, \Sigma_{d_{ij} d_{ij}}^{-1} \, \boldsymbol{d}_{x_{ij}} \sim \chi_2^2 \tag{11}$$

which, if all the assumptions hold, is following a χ_2^2-distribution. Hence we would count the inliers then by

$$N_c^M(T) = |\,(i,j)\mid t_{ij} < T_t\,| \tag{12}$$

and take the threshold, e.g., $T_t^2 = \texttt{chi2inv}(0.99, 2) = 9.2103$.

So, we do not need to apply least-squares matching but need to specify the size of the window around the keypoint detector (which depends on the detector, and possibly the scale the detector yields) and derive the CovM of LSM as proxy for the detectors.

2.2 The coverage

Following Shoaib Ehsan (2011) we can evaluate the detector by requiring the points to be as evenly distributed as possible. Of course, the image texture will provide limits, but we may compare two different detectors based on the coverage of the same image.

take the distances $d(i,i') = ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||$ of all points within a single image. Assuming the points are distinct, they first calculate the average distance of all points w.r.t. a given point as the harmonic mean of all distances to that point:

$$d_i = \frac{I-1}{\sum_{i' \neq i} \frac{1}{d_{ii'}}} \,. \tag{13}$$

They take the harmonic mean, in order to penalize small distances. The coverage then is defined as the harmonic mean of these local distances

$$c = \frac{I}{\sum_i \frac{1}{d_i}} \,. \tag{14}$$

In multi-scale systems, keypoints of different scale may be very close. Therefore they only use distances $d_{ii'}$ larger than a small threshold.

2.3 The matching criterium

We start with a criterium for two points are corresponding or matching.[1] We take as ground truth the given geometric relation. Due to the random noise in the images the vector $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ will randomly deviate from $\boldsymbol{0}$. If the uncertainty of the two points approximately follows a norm distribution with the given covariance matrices, we could use the (squared) Mahalanobis distance

$$d^2(\boldsymbol{x},\boldsymbol{y};\mathcal{H}) = \boldsymbol{f}^\mathsf{T} \Sigma_{ff}^{-1} \boldsymbol{f} \quad \text{with} \quad \underline{d}^2 \sim \chi_2^2 \tag{15}$$

which can be evaluated using its distribution under the assumption the two points match.

Observe, the covariance matrix Σ_{ff} depends on the covariance matrices of the two points and the parameters of the geometric relation. Hence, even if the points show the same precision, the difference $\boldsymbol{f}$ will not be the same for all position $\boldsymbol{x}$. However, if the transformation is close to an identity and the points have the same covariance matrix, a the distribution of d^2 will be the same.

Under these conditions, the distance d^2 can be transformed into a P-value using the cumulative probability distribution

$$P(d^2) = \mathbb{P}(\underline{x} > d^2) \tag{16}$$

which is

$$P(d^2) = 1 - F_{\chi^2}(d^2, 2) \quad \text{with} \quad F_{\chi^2}(x,k) = \int_{t=0}^{x} \chi^2(t,k) \mathrm{d}t \,. \tag{17}$$

[1] We use the attributes *corresponding* and *matching* as synonyms.

and in MATLAB is `P(d^2)=chi2cdf(d^2,2)`. If d is small, its P-value is large and thus which can be interpreted as the degree of closeness of the two points.

2.4 Obtaining putative matches

Given two sets of points $\boldsymbol{x}_i, i = 1, ..., I$ and $\boldsymbol{y}_j, j = 1, ..., J$ we want to evaluate how well they are suited for the following matching process.

We need to select all point in both images which may possibly lie in the region $\mathcal{R}_A$ or $\mathcal{R}_B$ of the other image. Hence, we keep those points $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ for which

$$\mathcal{X}_c = \{\boldsymbol{x}_i \mid \mathcal{H}(\boldsymbol{x}_i) \in \mathcal{R}_B\} \quad \text{and} \quad \mathcal{Y}_c = \{\boldsymbol{y}_i \mid \mathcal{H}^{-1}(\boldsymbol{y}_i) \in \mathcal{R}_A\}\,. \tag{18}$$

This will lead to I_c and J_c points in the two images which lie in the common region $\mathcal{R}_A \cap \mathcal{R}_b$.

2.5 Putative matches

We now want to find putative matches and evaluate these.

For the admissable points of an image pair we obtain a $I_c \times J_c$ distance matrix

$$\underset{I_c \times J_c}{\mathsf{D}_0} = [d_{ij}]\ . \tag{19}$$

At this point we might not want to arrive at a unique 1–1 matching, since at a later step we would take the features and the geometry into account to arrive at an optimal matching. Moreover, many of these edges will not represent correspondences. Hence, we could eliminate all edges with $d_{ij} < d_0$, which is equivalent to perform a χ^2-test and reject pairs with a too large Mahalanobis distance (squared).

Therefore, we take the cleaned graph with the distance matrix

$$\mathsf{D} = \mathsf{D}(d_0) = [\mathsf{D}_{ij}] = \begin{cases} [\mathsf{D}_{ij}]\,, \text{ if } d_{ij} < d_0 \\ 0, \text{ else} \end{cases} \tag{20}$$

and the adjacency matrix

$$\mathsf{A} = \mathsf{A}(d_0) = \text{sign}(\mathsf{D}(d_0)) \tag{21}$$

both matrices depending on the threshold d_0 for the normalized distance.

The rows and columns of the adjacency matrix contain useful information:

- The multiplicity of a match can be derived from the row and column sums if A.
- A matching (i, j) is unique, if the sums $\sum_i A_{ij} = \sum_j A_{ij} = 1$.
- A point i or j does have no putative match if the corresponding row or column is zero.

These values depend on the allowed distance d_0.

2.6 Evaluation of putative matches

We now can characterize the keypoint detector by using the adjacency matrix only, or, in addition, take the covariance matrices into account.

2.6.1 Evaluation solely based on the adjacency matrix

We may use the following properties, see Fig. 1

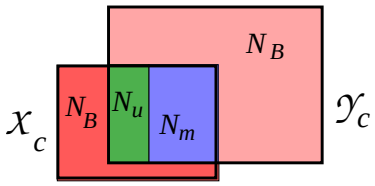

Figure 1: The keypoints selected in the two images and lying in the overlapping region of both images may be characterized by the numbers N_A and N_b of spurious points, the number N_u of unique matches and the number N_m of matches with multiple other points.

- The number $N_u(d_0)$ of unique matches and of the success ratio for unique matches
$$\wp_u(d_0) = \frac{N_u(d_0)}{\min(I_c, J_c)} \tag{22}$$
- The numbers $N_A(d_0)$ and $N_B(d_0$ of spurious points in image $\mathcal{A}$ and image $\mathcal{B}$ or the corresponding ratios
$$\wp_A(d_0) = \frac{N_A(d_0)}{I_c} \quad \text{and} \quad \wp_B(d_0) = \frac{N_B(d_0)}{J_c} \tag{23}$$
- The number $N_m(d_0)$ of points with multiple matching candidates, which will require a resolution in a later step, or the ratio
$$\wp_m(d_0) = \frac{N_m(d_0)}{I_c + J_c} \tag{24}$$

The numbers $N.$ may be averaged over several images. These averages may be normalized by the average number of points I_c and IJ_c of the images.

The numbers may be depend on the degree ε of distortion

The numbers may be documented as functions of d_0 or on ε.

A good keypoint detector would

- have a large percentage of unique matches,
- a small number of multiple matches, and
- a small number of spurious matches.

These numbers may, in addition to other criteria, be also used for the putative matches

- after having applied some similarity check of the keypoint descriptors,
- after having applied some geometric filter, e.g., using the inliers after RANSAC, or
- after boosting the correspondences by reinsertion of previously eliminated correspondences.

2.6.2 Evaluation including the covariance matrices

For the putative correspondences, it appears useful to exploit the covariance matrices for further characterization:

- The higher the precision of the selected points, the better. The precision can be characterized by the *Helmert point error* $\sqrt{\mathrm{tr}\boldsymbol{\Sigma}_{xx}}$ and its distribution.
- The putative matches could be used to estimate the geometric transformation.

  This leads to an *estimated variance factor* $\widehat{\sigma}_0^2$, which essentially depends on the residuals. This variance factor should be 1, if the covariance matrices are realistic, no outliers exist, and the supposed geometric transformation actually holds. This type of analysis could be done only with the unique matches, or, with all putative matches, including the multiple matches, when taking care of the multiplicity in the estimation model.

  Such an estimation would also yield a *covariance matrix for the estimated parameters*, which could be compared when using different detectors applied to the same image. The disadvantage of this measure is, that higher precision of one detector can be compensated by more matches of another detector. Therefore it is difficult to average over images of different complexity, though the precision of the transformation may be used in a follow-up step within processing chain (e.g., within a Kalman filter set-up).

2.7 Relation to previous work

The uniqueness measure r_u already has been proposed by Ehsan et al. (2016), there called *improved repeatability*, eq. (1). The dependency of the measure on compression, blur, and uniform lighing changes also is investigated in this paper.

3 Open questions

Using these measures for the characterization of the detectors, we still need to be sure that they can be generalized, i.e., not depend on specific properties of the data sets:

- What influence does the size of the image have on the performance characteristics?
- What influence does the number of the images have on the performance characteristics?
- What influence does image content have on the performance characteristics?
- What consequences do the answers to the previous questions have onto the design of a bechmark?

References

Ehsan, S., A. F. Clark, A. Leonardis, N. Ur Rehman, A. Khaliq, M. Fasli, and K. D. McDonald-Maier (2016). A Generic Framework for Assessing the Performance Bounds of Image Feature Detectors. *Remote Sensing 8*(11).

Shoaib Ehsan, Nadia Kanwal, A. F. C. K. D. M.-M. (2011). Measuring the Coverage of Interest Point Detectors. In A. C. M. Kamel (Ed.), *Image Analysis and Recognition*, Volume 6753. Springer.

Zhang, D. and C. Liu (2014). A salient object detection framework beyond top-down and bottom-up mechanism. *Biologically Inspired Cognitive Architectures 9*, 1–8.