Non-iterative normalized feature extraction in large viewpoint variances based on PCA of gradient

Jian Zhang[a], Song Cao[b], Di Wen[a]

[a]Dept. of Electronic Engineering, Tsinghua University, Beijing, China
[b]Dept. of Computer Science, University of Southern California, Los Angeles, CA, USA

ABSTRACT

Effective local feature extraction is one of the fundamental tools for retrieval applications in computer vision. However, it is difficult to achieve distinguishable local features in large viewpoint variances. In this paper, we propose a novel non-iterative approach of normalized feature extraction in large viewpoint variances, which adapts local regions to rotation, scale variance and rigid distortion from affine transformation. Our approach is based on two key ideas: 1) Localization and scale selection can be directly achieved with the centroid and covariance matrix of the spatial distribution of pixels in a local region. 2) Principal Component Analysis (PCA) on gradients of intensity gives information on texture, thus it can be used to get a resampled region which is isotropic in terms of variance of gradient. Experiments demonstrate that our normalized approach has significant improvement on matching score in large viewpoint variances.

Keywords: Local Feature, Affine Invariant, Non-iterative

1. INTRODUCTION

Local features are highly effective to represent discriminable information in small image structures. In general, hundreds of valid local interest points or regions can be detected in an image. The features extracted from them can be effectively applied for matching correspondences in image retrieval. However, as the shape and texture of a local region may vary obviously, the robustness of local features is hard to achieve in large viewpoint variances. On the other hand, in most of the practical retrieval applications, large viewpoint variances are inevitably introduced by the camera users. Many of the state-of-the-art local features are extracted from a circular image patch, ignoring the local geometric transformation. Therefore, the matching performance drops dramatically in large viewpoint variances. In order to maintain the effectiveness of local features, extraction methods have to be designed with adaptation to geometric transformation. There are many algorithms that have been proposed to solve this problem, which can be divided into two frameworks.

In the first framework, geometric transformations are estimated under the assumption of global planarity. Segmentation is employed in[1] and the estimated transformation is derived homogeneously for a whole segmented region. Yu and Morel[2] proposed affine scale space which involves affine transformation parameters into the original scale space proposed by Lindeberg.[3] The optimal transformation with the best feature matching result is selected globally for image pairs.

Another framework is based on an iterative local manner. Lindeberg and Garding's approach[4] iteratively refines the shape of Gaussian kernels. They achieved shape-adapted image smoothing after the relationship between the shape of kernel and the second moment matrix of gradient converges. The method in[5] employs the same idea to normalize neighborhoods of Harris corners in image matching. The work of Mikolajczyk and Shmid[6] introduced scale space into this framework for the first time. They achieved both scale and affine transformation invariance.

The approaches discussed above either are dependent on global planarity or show low efficiency as a consequence of the iteration. In this paper, we make efforts to harmonize high efficiency with the localness of normalization. We replace the constraint of global planarity with a week assumption of local planarity, which means that each local region is on an arbitrary planar surface and no local regions are required to be coplanar. To estimate a normalizing transformation, we analyze gradients of intensity in a local region with PCA (Principal Component Analysis). The normalization in our approach guarantees that local image patches corresponding

(a) (b) (c) (d)

Figure 1. Outline of our approach. (a) Detection: Regions filled with red pixels are detected by MSER. (b) Resample scale selection: Ellipses are fitted based on the estimation of Σ_s, the yellow one is described in Section 2.2, while the red one is enlarged to include the whole region. (c) Normalization: Applying our normalization based on PCA of gradient, the two normalized regions only have difference in rotation and scale, thus the invariance to rigid distortion is achieved. With the updated Σ_s described in Section 2.3, we get the scale for feature extraction. The yellow circle shows the range in which dominant gradient orientation is calculated. The red one denotes range for HOG feature extraction. (d) Feature extraction: Oriented HOG features extracted from patches in (c), which satisfy scale, rotation and rigid distortion invariance.

to the same region on a real object only have difference in rotation and scale. In this process, resampling is performed only once for each region, thus iteration is avoided. Assisted by our statistical scale selection and oriented feature extraction, our approach achieves full invariance to rotation, scale variance and rigid distortion simultaneously in large viewpoint variances.

Our approach with specific steps is discussed in Section 2. Matching score comparison and robust estimation results are shown with experiments in Section 3.

2. FEATURE EXTRACTION BASED ON PCA OF GRADIENT

2.1 Outline of our approach

Our proposed approach is a non-iterative framework, which automatically adapt local features to rotation, scale change and rigid distortion in large viewpoint variances. The outline of our approach is described in the following.

1) Initially detect local regions with MSER.[7]

2) Calculate the centroid and covariance matrix of pixels distribution in a local region. The eigenvalue is used to determine a spatial range for resample.

3) Analyse the covariance matrix of gradient in the original region. An affine transformation is estimated to transformthe original region to an isotropy one which has equal variance of gradient in an arbitrary direction. At the same time, our approach can also prevent the normalizing resampling from losing image details.

4) Update the covariance matrix of pixels distribution with the normalizing transformation. To maintain the completeness of a normalized local region, we select a proper scale for feature extraction with the updated covariance matrix in the same way as 2).

5) Extract oriented HOG descriptors on normalized regions with the scale selected in 4).

2.2 Statistical localization and scale selection

Our local region is localized at the centroid of the pixels in detected regions. As shown in Figure 1(a), we initially detect local region Ω with MSER. The localization in our approach has sub-pixel accuracy as it is a statistical average of the spatial distribution of pixels.

To achieve scale selection, the most widely used approach is the scale-space framework[8] involving construction of image pyramid for the detection of local maxima in scale space. In our approach, a more efficient statistical way is employed.

We estimate the covariance matrix Σ_s of the spatial distribution of pixels in the local region with

$$\Sigma_s = \int_{x \in \Omega} (x - E(x))(x^T - E(x^T))dx. \tag{1}$$

The symmetric Σ_s can be decomposed as

$$\Sigma_s = A \begin{bmatrix} \lambda_{s1} & 0 \\ 0 & \lambda_{s2} \end{bmatrix} A^T. \tag{2}$$

Rows of orthogonal matrix A denote two directions $\boldsymbol{a}$ and $\boldsymbol{b}$ in which projected ordinates of pixels in Ω have variances λ_{s1} and λ_{s2} respectively. Supposing that $\lambda_{s1} \geq \lambda_{s2}$, it can be proved $\boldsymbol{a}$ is the direction in which the pixels in Ω are the most scattered with the largest spatial distribution variance. In common cases, if pixels in a local region are all included in direction $\boldsymbol{a}$ in a certain circle, we can guarantee with great probability that pixels are all included in an arbitrary direction so that we will not neglect any parts of the detected region for resample and feature extraction. In our experiment, we use λ_{s1} as the scale. Pixels lying in a circle which is centerd at the centroid and has a radius equalling 2.5 times of λ_{s1} in the original patch are all mapped into the resampled normalized patch. Note that we first calculate Σ_s on the original patch to obtain a scale for the spatial range for resample, then we directly update Σ_s,which will be discussed in Section 2.3 to obtain a new scale for feature extraction.

To illustrate the above description more clearly, we fit yellow ellipses in Figure 1(b). The direction of long axis and short axis are $\boldsymbol{a}$ and $\boldsymbol{b}$, while the half length of the axes are λ_{s1} and λ_{s2}.

2.3 Normalization based on PCA of gradient

Once localization is done and resample scale is selected, we now consider normalizing the local region. In our approach, the first objective of normalization is to obtain a structure isotropic in terms of the variance of gradient. The second objective is to maintain as many details in texture as possible without iteration.

In Lindeberg,[4] the second moment matrix, or the correlation matrix of gradient is employed to adapt Gaussian kernel to the local structure for shape adapted image smoothing. Baumberg,[5] Mikojczyc and Shmid[6] both use a Gaussian-windowed second moment matrix iteratively. Inspired by their methods, we base our normalizing on the PCA of gradients covariance matrix. As MSER is involved, the major difference is we do not need iterative refinement to captured the shape of the region.

For image $I(x)$ and Ω, a region detected by MSER, we define the covariance matrix of gradient as

$$\Sigma_g = \int_{x \in \Omega} (\nabla I(x) - E(\nabla I(x)))(\nabla^T I(x) - E(\nabla^T I(x)))dx. \tag{3}$$

Decomposing Σ_g as

$$\Sigma_g = B \begin{bmatrix} \lambda_{g1} & 0 \\ 0 & \lambda_{g2} \end{bmatrix} B^T, \tag{4}$$

in which λ_{g1} and λ_{g2} ($\lambda_{g1} \geq \lambda_{g2}$) are positive and B is an orthogonal matrix, we can tell that on the directions denoted by the two rows of B, the variances of gradient reach maximal λ_{g1} and minimal λ_{g2} respectively, which is the main property of PCA.

In order to clarify the meaning of the decomposing in Equation (4), we can imagine that when a small patch is shrunk, or downsampled, the value of gradient will increase and so as to the variance of gradient. To achieve the second objective of our normalization method, we have to prevent losing details by avoiding spatial compression. If we guarantee that the variance of gradient does not increase in an arbitrary direction after the normalization, we can avoid spatial compression, thus to maintain as much details as possible in the whole local region. To achieve this goal, we constrain that the variance of gradient equals λ_{g2} in an arbitrary direction in the normalized isotropic region.

Now we will show a general relationship between a patch and its correspondence after a affine transformation. Patch $I_1(x)$ is transformed to $I_2(x) = I_1(Ax)$ with an affine matrix A, the corresponding regions Ω_1 and $\Omega_2 = \{x|Ax \in \Omega_1\}$ are detected respectively by MSER. The relation between gradient of the patches is

$$\nabla I_2(x) = A^T \nabla I_1(x')|_{x'=Ax}. \tag{5}$$

Calculating Σ_{g1} and Σ_{g2} respectively on $I_1(x)$ and $I_2(x)$ with Equation (3) and substitute (5) into it, we will find a general relationship

$$\Sigma_{g2} = det(A)^{-1} A^T \Sigma_{g1} A. \tag{6}$$

As a special case, if we set transformation A to

$$A = \lambda_{g1min} det(\Sigma_{g1}^{-1/2}) \Sigma_{g1}^{-1/2} \tag{7}$$

in which λ_{g1min} denotes the smaller eigenvalue of Σ_{g1}. Then we have

$$\Sigma_{g2} = \begin{bmatrix} \lambda_{g1min} & 0 \\ 0 & \lambda_{g1min} \end{bmatrix}. \tag{8}$$

It means after transforming $I_1(x)$ to $I_2(x) = I_1(Ax)$, we get the desired isotropic resampled patch $I_2(x)$ with a variance of gradient equalling λ_{g1min} in an arbitrary direction.

For each single original patch, we apply the transformation in Equation (7) respectively. To reveal the relationship between two normalized patches originally sampled from different viewpoints, we will discuss the following case.

$I_{o1}(x)$ and $I_{o2}(x)$ are patches from two images taken from different viewpoints . They contain local regions corresponding to the same region on a real object. $I_{n1}(x)$ and $I_{n2}(x)$ are the corresponding normalized patches. Under our week assumption of locally planarity, the relation between $I_{o1}(x)$ and $I_{o2}(x)$ is affine transformation. The normalizations on $I_{o1}(x)$ and $I_{o2}(x)$ are also affine. Thus the relation between $I_{n1}(x)$ and $I_{n2}(x)$ is an affine transformation. We denote the transformation as B with which the relation between the normalized patches can be describe as $I_{n2}(x) = I_{n1}(Bx)$.

Let λ_{go1min} and λ_{go2min} denote the smaller eigenvalues of Σ_{go1} and Σ_{go2} (Σ_g of patches $I_{o1}(x)$ and $I_{o2}(x)$) respectively, we will have

$$\Sigma_{gn1} = \begin{bmatrix} \lambda_{go1min} & 0 \\ 0 & \lambda_{go1min} \end{bmatrix} \qquad \Sigma_{gn2} = \begin{bmatrix} \lambda_{go2min} & 0 \\ 0 & \lambda_{go2min} \end{bmatrix} \tag{9}$$

With Equation (6) we can derive

$$\Sigma_{gn2} = det(B)^{-1} B^T \Sigma_{gn1} B. \tag{10}$$

Substituting Equation (9) to (10), we get

$$B^T B = det(B) \frac{\lambda_{go2min}}{\lambda_{go1min}} I \tag{11}$$

in which I is an identity matrix. Now it emerges that B is orthogonal as a compound of rescaling and rotation. It is proved that normalized with our proposed approach, the two patches originally from different viewpoints now only have difference in rotation and a constant factor for rescaling.

In Figure 1(c), the normalization result is visualized. Reverse mapping and bilinear interpolation are used to get the normalized region.

Additionally, in our experiments, we found that regions detected by MSER tend to have large intensity changes near the region boundary, thus providing rich information about gradient. If we enlarge the detected region with four-neighbours of the boundary pixels, the estimation on Σ_g may have obvious change. Empirically, we can achieve an relatively stable estimation if we repeat the enlargement for $3 \sim 5$ times. This inflation procedure can result in more reliable estimation of the transformation matrix.

2.4 Descriptor extraction

On the normalized region pairs, there is still difference in scale and rotation. Many Descriptors dealing with these two differences have been proposed. Most of the state-of-the-art descriptors keep invariance to scale change, and in Lowe [9] and Shmid and Mohr ,[10] methods to obtain invariance to rotation are proposed. In our experiment, we use the oriented HOG in SIFT as descriptor which shows good rotation and scale invariance.

In the first step of descriptor extraction, we have to guarantee the spatial completeness of the normalized region with a new scale on the normalized region. We may directly calculate Σ_s with Equation(1) inside the normalized region. However, considering to avoid the redetection with MSER and recalculation of Σ_s with Equation (1), we substitute Equation (7) into Equation (1) and the calculation can be simplified as

$$\Sigma_{sn} = \lambda_{gomin}^{-3} det(\Sigma_{go}^{1/2})^4 \Sigma_{go}^{1/2} \Sigma_{so} \Sigma_{go}^{T/2} \tag{12}$$

in which, index s refers to the spatial distribution of pixels, while g is related with gradient. Index o and n mean the original region and normalized region respectively. λ_{gomin} is the smaller eigenvalue of Σ_{go}. In this way, without too much computation, we get the characteristic scale λ_{snmax} (the greater eigenvalue of Σ_{sn}) of the resampled region for feature extraction.

In our experiment, we calculate dominant gradient orientation in a circle (shown with yellow color) in Figure 1(c) with a radius of λ_{snmax} on the normalized patch. Descriptor is formed with the gradient in another circle (shown with red color) in Figure 1(c) with a radius of 2.5 times of λ_{snmax}. In terms of the histograms in calculating descriptors, we used 36 bins for dominant orientations and 128 bins for forming descriptors.

3. EXPERIMENT

In this section, we demonstrate the performance of our normalized extraction approach. Our approach is compared with Harris Affine, Hessian Affine, MSER and SIFT. Implementations of Harris Affine and Hessian Affine are from the author's website.[11] SIFT is implemented with VLfeat.[12]

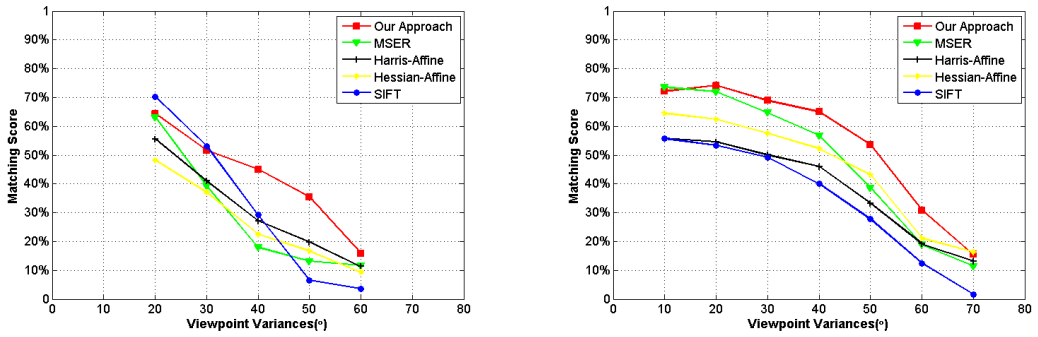
We test the performance on two datasets. The first dataset is 'Graffiti', outdoor scene sequences from website,[11] which is used to test Harris-Affine and Hessian-Affine in viewpoint variances in.[6] We set up the second dataset 'Close-up Shot' to have further test. 10 subsets of object images are collected, and each one has a sequence whose viewpoint angle changes from 10° to 70°.

3.1 Matching score test

To test the extraction performance, we judge different approaches by comparing the putative matching result of features extracted with different approach. Matching score used in Mikolajczyk and Shmid[13] is the metric in our experiment, which is the number of correct matched local region pairs with respect to the total number of putatively matched pairs. A region is putatively matched with its nearest neighbour if its angle in feature space to its nearest neighbour is less than 0.9 times of that to its second nearest neighbour. We verify correctly matched pairs with their consistency with the ground truth transformation.

In Figure 2, Table 1 and Table 2, we show the matching score comparison. The viewpoint variance ranges have a little difference, as variances in 'Graffiti' are from 20° and 60° while 'Close-up Shot' ranges from 10° to 70°. All the methods in the comparison have generally better performance on 'Close-up Shot', as the photos in 'Close-up Shot' are product packages with more surface pattern details and less background interference.

On both datasets, our approach outperforms the others in viewpoint variances from 30° to 60°. Comparing with Harris-Affine and Hessian-Affine, we have 10% to 15% improvements in large viewpoint variances. Our approach successfully slows down the decline of matching score and we can still keep it above 35% and 50% respectively in the two datasets under variance of 50°. It is noticeable that our approach also has significant improvement over MSER, which shows that our normalization based on PCA of gradient shows its contribution to the improved performance.



(a) (b)

Figure 2. Matching score comparison: Comparison on matching score among our approach, Harris-Affine, Hessian-Affine, MSER and SIFT. (a) Comparison on 'Graffiti'. (2) Comparison on 'Close-up Shot' (average on 10 subsets).

	Viewpoint Variance(°)						
Algorithm	10	20	30	40	50	60	70
Our approach	∼	64.42%	51.68%	**45.02%**	**35.47%**	**15.82%**	∼
Harris-Affine	∼	55.60%	41.00%	27.10%	19.80%	11.30%	∼
Hessian-Affine	∼	48.20%	37.10%	22.50%	16.70%	9.40%	∼
SIFT	∼	**70.92%**	**53.16%**	29.22%	6.51%	3.65%	∼
MSER	∼	63.23%	39.24%	17.90%	13.21%	11.65%	∼

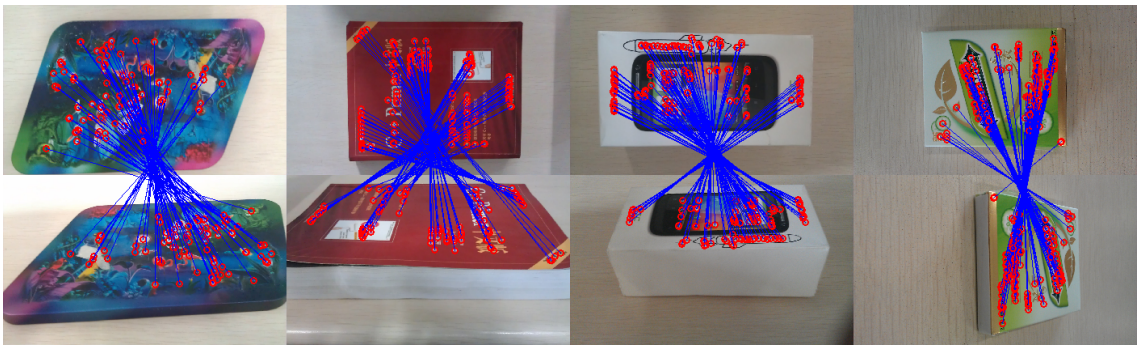
Table 1. Comparison on 'Graffiti'

	Viewpoint Variance(°)						
Algorithm	10	20	30	40	50	60	70
Our approach	72.17%	**74.16%**	**68.94%**	**65.00%**	**53.67%**	**30.79%**	**15.50%**
Harris-Affine	55.80%	54.59%	50.07%	46.01%	33.26%	19.18%	13.18%
Hessian-Affine	64.46%	62.43%	57.48%	52.26%	43.19%	21.14%	16.37%
SIFT	55.73%	53.40%	49.22%	40.01%	27.78%	12.41%	1.63%
MSER	**73.54%**	71.94%	64.79%	56.80%	38.71%	18.98%	11.42%

Table 2. Comparison on 'Close-up Shot' (average on 10 subsets)

3.2 Application in robust estimation

To apply our extraction approach to robust estimation, we conduct experiments with two stages. In the matching stage, we simply match descriptors with their nearest neighbours. A threshold on the correlation of features vectors is involved as an initial elimination of wrong matches. In the estimation stage, we estimate geometric constraints with RANSAC[13] on the initial correspondence set. The fitted models vary according to the content of test images. We estimate affine transformation with normalized DLT[14] algorithm for planar surfaces and fundamental matrix with normalized eight-points algorithm[15] for complex scenes.



(a)

(b)

Figure 3. Robust estimation test: (a) Affine transformation guided matching in 60° viewpoint variance. (b) Estimation of fundamental matrix in complex scenes. Epipolar lines are shown to demonstrate the result of estimation. In both (a) and (b), small red spots are the location of the related local regions. Only a subset of spots and lines are shown for clearness.

In affine transformation guided matching, we use the estimation of RANSAC to limit the correspondence searching area in the image for each descriptor to achieve guided match. In Figure 3(a), robust matching is shown between image pairs in 'Close-up Shot'. We typically have a few hundreds of correct matches on 800×600 images after guided matching.

As discussed in Section 1, our normalized extraction only depends on the week assumption of locally planarity. To show the extraction performance in complex scenes which are not globally planar, we use features extracted with our approach to estimate fundamental matrix. The epipolar lines corresponding to the matches tightly consistent with the estimation are shown in Figure 3(b). In experiments shown in Figure 3, we only display subsets of the spots and lines for clearness.

4. CONCLUSION

In this paper, we proposed a novel non-iterative approach for normalized local feature extraction. We simultaneously adapt local regions to rotation, scale change and rigid distortion from affine transformations in large viewpoint variances. Localization and scale selection are based on analysis of the spatial distribution of pixels. Normalization is performed with PCA on gradient in local regions without any iteration. In our experiments, matching score is significantly improved in the comparison with other feature extraction methods in large viewpoint invariances. Application tests in affine transformation guided matching and fundamental matrix estimation show that our extraction approach works both in global planar cases and complex scenes. As our initial detector is MSER, our approach is more effective on image with many closed segmentable local regions (e.g. letters, small pattern in trademarks). Future work may include methods to combine our approach with other local point based algorithm so that the normalized approach can be more effective both on images mainly providing discriminable local points and images with more closed segmentable local regions.

REFERENCES

[1] Schaffalitzky, F. and Zisserman, A., "Viewpoint invariant texture matching and wide baseline stereo," in [*Proc. ICCV*], **2**, 636–643 (2001).

[2] Yu, G. and Morel, J.-M., "A fully affine invariant image comparison method," in [*Proc. ICASSP*], 1597–1600 (2009).

[3] Lindeberg, T., "Feature detection with automatic scale selection," *Journal of IJCV* **30**(2), 79–116 (1998).

[4] Lindeberg, T. and Garding, J., "Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure," *Journal of Image and Vision Computing* **15**, 415–434 (1997).

[5] Baumberg, A., "Reliable feature matching across widely separated views," in [*Proc. CVPR*], **1**, 774–781 (2000).

[6] Mikolajczyc, K. and Shmid, C., "Scale & affine invariant interest point detectors," *Journal of IJCV* **60**(1), 63–86 (2004).

[7] Matas, J., Chum, O., Urban, M., and Pajdla, T., "Robust wide-baseline stereo from maximally stable extremal regions," in [*Proc. Image and Vision Computing*], **22**, 761–767 (2004).

[8] Lowe, D. G., "Distinctive image features from scale-invariant keypoints," *Journal of IJCV* **60**(2), 91–110 (2004).

[9] Schmid, C. and Mohr, R., "Local grayvalue invariants for image retrieval," *Journal of T-PAMI* **19**, 530–535 (1997).

[10] Visual Geometry Group, O. U., "Affine covariant features." `http://www.robots.ox.ac.uk/~vgg/research/affine/` (2006).

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

[12] Mikolajczyk, K., Tuytelaars, T., Schmid, Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Van Gool, L., "A comparison of affine region detectors," *Journal of IJCV* **65**(1), 43–72 (2005).

[13] Fischler, M. A. and Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM* **24** (1981).

[14] Hartley, R. I. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, ISBN: 0521623049 (2000).

[15] Hartley, R. I., "In defense of the eight-point algorithm," *Journal of T-PAMI* **19**, 580–593 (1997).