Automatic landmark correspondence detection for ImageJ

Stephan Saalfeld and Pavel Tomančák

Max Planck Institute of Molecular Cell Biology and Genetics Dresden, Germany

ABSTRACT

Landmark correspondences can be used for various tasks in image processing such as image alignment, reconstruction of panoramic photographs, object recognition and simultaneous localization and mapping for mobile robots. The computer vision community knows several techniques for extracting and pairwise associating such landmarks using distinctive invariant local image features. Two very successful methods are the Sale Invariant Feature Transform (SIFT)[1] and Multi-Scale Oriented Patches (MOPS).[2]

We implemented these methods in the Java programming language[3] for seamless use in ImageJ.[4] We use it for fully automatic registration of gigantic serial section Transmission Electron Microscopy mosaics. Using automatically detected landmark correspondences, the registration of large image mosaics simplifies to globally minimizing the displacement of corresponding points.

We present here an introduction to automatic landmark correspondence detection and demonstrate our implementation for ImageJ. We demonstrate the application of the plug-in on generic and biological image data.

Keywords: SIFT, Scale-Invariant Feature Transform, Automatic Landmark Detection, Registration, Mosaic, Transmission Electron Microscopy

1. INTRODUCTION

Automatic detection of landmark correspondences is crucial for several computer vision tasks. Such tasks include for example automatic image alignment, reconstruction of panoramic photographs, object recognition and simultaneous localization and mapping for mobile robots. While—in principle—arbitrarily selected sets of landmarks do not provide additional information compared to the image as a whole, they allow to focus on a handful stable and meaningful properties of the image at a low level of processing. Simplifying the abovementioned recognition problems to a moderate number of corresponding points helps to reduce the computational cost significantly while preserving in most cases sufficient generality.*

Landmark correspondence detection requires *detection* and *matching* of *interest points*. Ideally, matching is performed invariantly to the transformation that maps one image (or a part of it) to the other. State of the art techniques[1,5–7] allow both automatic detection of interest points and extraction of affine invariant local descriptors for these points. Accordingly, affine invariant matching is nearest neighbour search in such a local feature descriptor space.

It was shown[8] that the local descriptor as used in the *Scale Invariant Feature Transform* (SIFT)[1] outperforms competing techniques in both distinctiveness and robustness with respect to significant image deformation. The SIFT descriptor was published in combination with the *Difference of Gaussian* (DoG) detector (a scale invariant blob detector) and a robust orientation detection thus realizing similarity invariant feature extraction.[1] The author provides his implementation as closed source binaries for free academic use. Meanwhile, several open source implementations for MATLAB, C, C++ and C# are available however an implementation in the Java programming language is missing.

We developed this Java implementation and integrated it into ImageJ as a plug-in. We carefully separated the scale invariant interest point detection from local descriptor extraction such that different detectors and

Further author information:
E-mail: saalfeld@mpi-cbg.de, Telephone: +49 351 210 2753

*For specific applications, this loss in generality has to be evaluated carefully. Landmark correspondences are not the proper answer to *every* problem even if they may appear tempting at first glance.

descriptors may be combined in future implementations. We demonstrate this flexibility with the implementation of Multi-scale Oriented Patches (MOPS)[2] as an alternative local feature descriptor. We subsume here the implemented Difference of Gaussian interest point detector[1] as well as the SIFT and MOPS local feature descriptors. We demonstrate the functionality of the plug-in for registration of images as an example of biological application.

2. MATERIALS AND METHODS

The Scale Invariant Feature Transform was described for 2d intensity images in the range [0–1]. All incorporated operations on intensities require floating point accuracy. Because our implementation is intended to be used as a library independently from ImageJ, our basic image container is the simplest imaginable—a `float[]` that knows about its width and height. For this container, we implemented the required basic filters: accurate Gaussian smoothing and gradient extraction.

2.1 Interest Point Detection

Intensity invariant interest points in an image $I(x, y)$ are structures that are 2d-localizable. That is, the principal curvatures being the eigenvalues α and β of the Hessian matrix $\mathbf{H}(I(x, y))$ both run through a local maximum. Homogeneous areas are indicated by two low principal curvatures, edges by one high and one low curvature and corners, junctions, corners or blobs by two high curvatures. Both principal curvatures being at a local maximum implies the Laplacian $\nabla^2 I(x, y)$ being at a local extremum. It is thus appropriate to preselect extrema of $\nabla^2 I(x, y)$ as interest point candidates and reject edge responses with one low principal curvatures afterwards instead of evaluating the Hessian eigenvalues explicitly for each pixel location.[1]

Scale invariance is guaranteed by inspecting the image at all scales. This ensures that the detection of a local structure depends on the size of the structure only, regardless of the scale of observation. The *linear scale space*[9] $L(\sigma, x, y)$ is an elegant way to represent the image $I(x, y)$ with multiple scales embedded as an explicit parameter. As this parameter increases, finer details are increasingly suppressed. It was shown[9] that *Gaussian smoothing* is the exclusive operator for this task and thus the sole possible scale space kernel. That is, the scale space $L(\sigma, x, y)$ is constructed by smoothing the image $I(x, y)$ with Gaussians $G(\sigma, x, y)$ with increasing σ

$$L(\sigma, x, y) = G(\sigma, x, y) * I(x, y) \quad \text{with} \quad G(\sigma, x, y) = \frac{1}{2\pi\sigma^2} e^{\left(x^2+y^2\right)/2\sigma^2}$$

The difference of two scales[†] that are separated by a constant factor k approximates a scale normalized Laplacian[9]

$$D(\sigma, x, y) = L(k\sigma, x, y) - L(\sigma, x, y) \approx \sigma^2 \nabla^2 G(\sigma, x, y).$$

Scale invariant interest point candidates are local extrema of $D(\sigma, x, y)$, that is they have a specific location and size. Such a local extremum is either larger or smaller than each of its 26 neighbours $D(k^a\sigma, x + b, y + c)$ with $a, b, c \in \{-1, 0, 1\}$. Each detected extremum is located with sub-pixel accuracy by fitting a 3d-quadratic function to $D(\sigma, x, y)$ and its local neighbourhood. In order to virtually unify the sampling distance in each dimension, the scale index i being the exponent of $k : \sigma_i = k^i\sigma_0$ is used as the scale parameter here. The offset $\vec{o} = (o_i, o_x, o_y)^\mathsf{T}$ from the original extrema detection at $\vec{d} = (d_i, d_x, d_y)^\mathsf{T}$ is given by

$$\vec{o} = -\frac{\partial^2 D}{\partial \vec{d}^2}^{-1} \frac{\partial D}{\partial \vec{d}}.$$

In case that at least one dimension of $\vec{o}$ has an absolute value of more than 0.5, the localization is repeated at the sample point that is closest to the estimated location. Potentially unstable detections with low contrast are rejected using the interpolated value $\hat{D}(d_\sigma + o_\sigma, d_x + o_x, d_y + o_y)$. Our implementation uses the suggested absolute value $|\hat{D}|_{min} = 0.03$.[1]

[†]This is where the term *Difference of Gaussian* comes from.

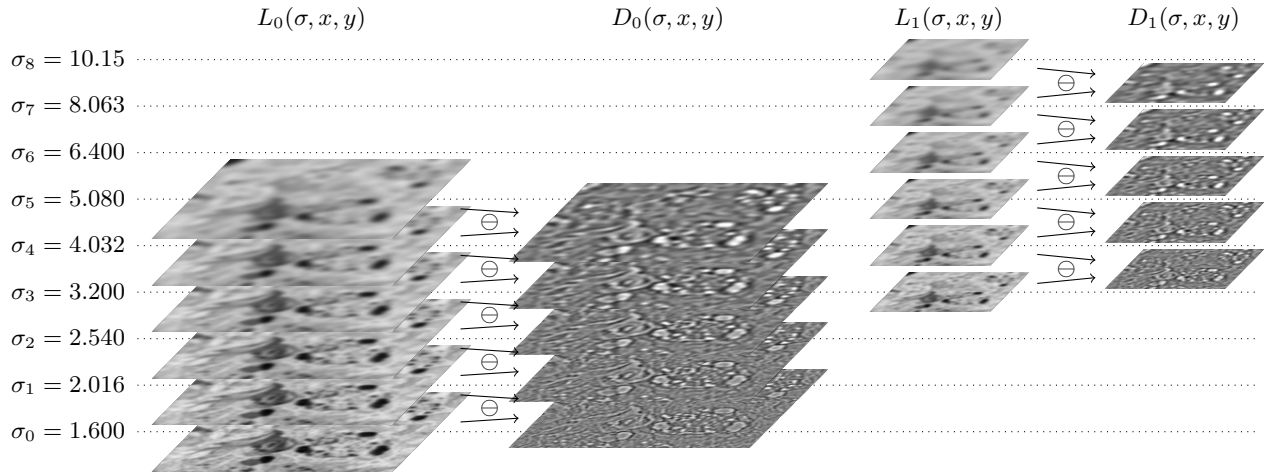


Figure 1. Illustration of the described discrete scale space representation[1] $L_o(\sigma, x, y)$ and its Differences of Gaussians $D(\sigma, x, y)$ showing a Transmission Electron Micrograph. The first two scale octaves with $s = 3$ scale steps per octave are shown. Adjacent scales are separated by constant factor $k : \sigma_{i+1} = k\sigma_i$ with $k = 2^{1/s}$. Each scale octave has $s + 3$ entries with the last three entries representing the same scales as the first three entries of the next octave. This is required for extrema detection in the scale domain of $D(\sigma, x, y)$.

The resulting interest point candidates of these detection steps are either blobs or edges, whereas edge responses are poorly localizable alongside the edge. Therefore, edge responses are rejected by principal curvature comparison.[10] The ratio r of the two eigenvalues $\alpha = r\beta$ is related to the trace $\mathrm{tr}(\mathbf{H})$ and the determinant $\det(\mathbf{H})$ of the Hessian matrix $\mathbf{H}(I(x, y))$[1]

$$\frac{\mathrm{tr}(\mathbf{H})^2}{\det(\mathbf{H})} = \frac{(\alpha + \beta)^2}{\alpha\beta} = \frac{(r\beta + \beta)^2}{r\beta^2} = \frac{(r + 1)^2}{r}.$$

As suggested in the original publication,[1] our implementation rejects detections with $r > 10$ as edge responses.

Having spatial frequencies $< \sigma$ suppressed, an image with $\sigma > 2.0$ can be represented sufficiently with half the sampling frequency of the original image. Lowe[1] suggests to create the scale space as a pyramid of *scale octaves* each covering the scale range $[\sigma_0, 2\sigma_0]$ with a fixed number of *scale steps* s, such that $k = 2^{1/s}$. An octave's first entry $L(\sigma_0, u, v)$ is created from $L(2\sigma_0, u, v)$ of the previous octave by sampling its even pixel coordinates. The octave is then generated by convolving the first image with a fixed set of differential Gaussians. In order to be able to detect DoG-extrema without gaps in the scale domain, each octave has $s+3$ entries in which the last three entries represent the same scales as the first three entries of the next octave (see Fig. 1). In our implementation, scale octaves act independently from each other with each octave being a scale space on its own. The number of scale steps s per octave and the initial Gaussian smoothing σ_0 are parameters of the implementation. Lowe[1] found $s = 3$ and $\sigma_0 = 1.6$ to be appropriate for stable interest point detection.

2.2 Orientation detection

The orientation of a detection is defined as the dominant orientation of the gradients in a local neighbourhood of the detection. This local neighbourhood is given by a Gaussian mask with $\sigma_m = 1.5\sigma$ that of the detection weighting the gradient amplitudes. It is sufficient to sample a region of 9.0σ px sidelength. As suggested,[1] we gather a 36-bin gradient histogram per detection and identify the largest bin being the dominant orientation of the region. For each other bin larger than 80 % of the largest, the detection is duplicated. These dominant orientations are sub-bin localized by fitting a quadratic function to its two neighbouring bins similar to the sub-pixel localization for DoG extrema as described above.

2.3 Local feature description

The above steps result in a set of stable interest points and a specific similarity transformation for each of them. This similarity defines a local feature coordinate frame that guarantees similarity-invariance for the local feature

descriptor to be extracted.

MOPS-descriptors as described by Brown[2] are intensity patches of fixed sidelength l that are sampled around the interest point locations. l is a parameter of the implementation whereas larger descriptors are typically more distinctive while also being more sensitive with respect to occlusion and other artifacts. In order to be more robust against inaccurate detections and minor image deformations, Brown[2] suggests to sample the descriptor with a significantly higher spacing than the detected σ. Instead of the suggested 5σ, we use 4σ which means a spacing of σ in the corresponding entry of the scale octave after the next. By stretching the sampled intensities to the range $[0, 1]$, the descriptor is invariant to exposure variance and contrast changes (see Fig. 2 for an illustration of the MOPS-descriptor).

Instead of sampling smoothed intensities, Lowe[1] proposed to sample $w \times w : w = 4l$ gradients in the original σ-spacing. The amplitudes of the sampled gradients are weighted with a Gaussian mask with $\sigma = l/2$. From each 4×4 sample block, a gradient histogram with b histogram bins is built such that each gradient contributes to its two closest bins linearly. By this, the spatial image information is smoothed comparable to what was described for MOPS while preserving more structural content. Illumination invariance is ensured by stretched all histograms to the range $[0, 1]$ and furthermore cutting all bins above a threshold value $t = 0.2$. The number of 4×4-sample blocks per side l and the number of orientation histogram bins b are parameters of the implementation with $l = 4$ and $b = 8$ being the originally suggested values.[1]

2.4 Local descriptor matching

Corresponding interest points are expected to be nearest neighbours in such a local descriptor space. Whereas more efficient techniques for approximate nearest-neighbour search in high-dimensional spaces exist,[11] we use an exact solution implemented as exhaustive search here. A such constructed set of correspondence candidates $C = \{(\vec{a}_i, \vec{b}_i) : i = \{1, 2, \ldots, |C|\}\}$ typically includes a large number of false matches as not all detections are present in both images. Since no global distance-threshold exists which would separate true from false matches, Lowe[1] suggests to threshold on the ratio r_{NN} of the distances to the nearest and next-nearest neighbours. True matches are expected to have a significantly lower distance than false matches such that r_{NN} is small while going towards 1.0 for false matches. This threshold is a parameter of the implementation with 0.8 being the suggested value.[1]

2.5 Image registration with consistent landmark correspondences

Depending on the images, the remaining set of false positives may still be significantly larger than the set of true matches. In registration applications, the images, and so the set of true matches, are related by an unknown transformation $\mathbf{T} : \mathbb{R}^2 \mapsto \mathbb{R}^2$. The transformation class is a parameter of the implementation with 2d-affine, 2d-rigid and 2d-translation being implemented currently. Regarding this transformation, true and false matches are called *inliers* $C_\mathbf{T} \subset C$ and *outliers* $C \setminus C_\mathbf{T}$. The tool of first choice for robust inlier-/ outlier-separation in presence of many outliers is the *Random Sample Consensus* (RANSAC)[12] that works as follows: For a fixed number of iterations, randomly select a minimal set of correspondence candidates $C_{min} \subset C$ and estimate $\mathbf{T}$ for them. The residual error of all candidates in terms of $\mathbf{T}$ is calculated and candidates with a residual error lower than some ϵ_{max}, which is a parameter of the implementation, are collected as inliers. The largest set of inliers found is used to estimate the optimal $\mathbf{T}$ by means of least square residual errors

$$\arg\min_{\mathbf{T}} \sum_{(\vec{a},\vec{b})\in C} \left\| \mathbf{T}\vec{a} - \vec{b} \right\|^2 .$$

If the number of inliers is smaller than some minimal fraction of all candidates, no transformation is returned. This minimal fraction is a parameter of the implementation with 5 % of all candidates as default.

In our implementation, RANSAC is followed by a a robust regression filter that iteratively removes correspondences with a residual error larger than $3\times$ the median of all residual errors. By this, usually ϵ_{max} can be set to very tolerant values. The plug-in suggests 5 % of the image size by default.

The estimated transformation maps one landmark set to the other and since these landmarks act as a statistic for the whole image, it can be used to map the images into each other. Even in presence of significant artifacts

masking substantial areas of the images as usual in serial section microscopy, local image features provide robust results. Linear transformations from robust landmark correspondences as described above can be used as an initialization for non-linear registration schemes.[13,14]

3. RESULTS

We implemented an ImageJ-plug-in which automatically extracts corresponding landmarks in two images that are approximately related by a 2d-affine, 2d-rigid or 2d-translation transformation. The plug-in uses our implementation of the DoG interest point detector, SIFT and MOPS feature descriptors, RANSAC and the mentioned transformation classes. Corresponding landmarks are added as PointRois to both images whereas corresponding points have the same index. As an example of use, we implemented an ImageJ-plug-in that uses such landmark correspondences for calculating a 2d-affine, 2d-rigid or 2d-translation transformation for an image and mapping it respectively. Meanwhile, there are more plug-ins available[14] that can utilize the extracted landmark correspondences for non-rigid image registration.

Our implementation of SIFT is included into TrakEM2[15] for automatic registration of large serial section microscopy mosaics. We were able to register the *Drosophila first instar larval brain data set* by Albert Cardona fully automatically.[16] The data set consists of 85 sections of 60 nm thickness. Each section was imaged with 9×9 images of $2,048 \times 2,048\,\mathrm{px}^2$ with approximately 6 % image overlap resulting in 6,885 images.[‡]

4. CONCLUSIONS

REFERENCES

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

[2] Brown, M., Szeliski, R., and Winder, S., "Multi-image matching using multi-scale oriented patches," in [*CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*], 510–517, IEEE Computer Society, Washington, DC, USA (2005).

[3] Sun Microsystems, Inc., "The Java programming language [JDK version 6]," (1994–2008).

[4] Rasband, W., "ImageJ: Image processing and analysis in Java [version 1.40g]," (1997–2007).

[5] Mikolajczyk, K. and Schmid, C., "Scale and affine invariant interest point detectors," *International Journal of Computer Vision* **60**(1), 63–86 (2004).

[6] Tuytelaars, T. and Van Gool, L., "Matching widely separated views based on affine invariant regions," *International Journal of Computer Vision* **59**(1), 61–85 (2004).

[7] Matas, J., Chum, O., Urban, M., and Pajdla, T., "Robust wide baseline stereo from maximally stable extremal regions," in [*Proceedings of the British Machine Vision Conference*], Rosin, P. L. and Marshall, D., eds., **1**, 384–393, BMVA, London, UK (September 2002).

[8] Mikolajczyk, K. and Schmid, C., "A performance evaluation of local descriptors," in [*International Conference on Computer Vision and Pattern Recognition*], **2**, 257–263 (June 2003).

[9] Lindeberg, T., "Scale-space theory: A basic tool for analysing structures at different scales," *Journal of Applied Statistics* **21**(2), 224–270 (1994).

[10] Harris, C. and Stephens, M., "A combined corner and edge detector," in [*Proceedings of The Fourth Alvey Vision Conference*], 147–151 (1988).

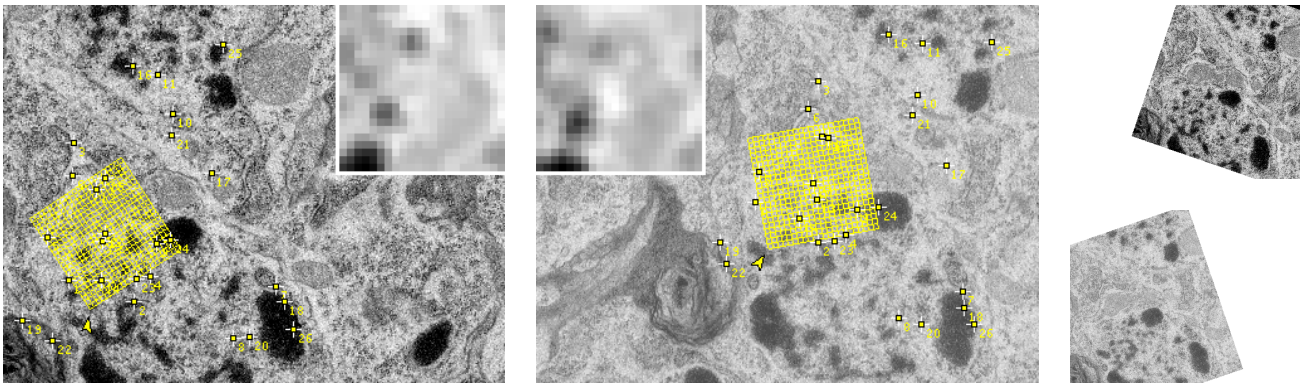
[11] Beis, J. S. and Lowe, D. G., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in [*CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*], 1000–1006, IEEE Computer Society, Washington, DC, USA (June 1997).

[12] Fischler, M. A. and Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM* **24**(6), 381–395 (1981).

[13] Schaefer, S., McPhail, T., and Warren, J., "Image deformation using moving least squares," *ACM Transactions on Graphics* **25**, 533–540 (July 2006).

[‡]The registered data set is available for on-line inspection at `http://fly.mpi-cbg.de/?pid=10&zp=0&yp=39678.1924&xp=41612.2026&sid0=10&s0=5`

Example 1. Serial section TEM, approx. rigid consecutive sections, 22 × 22 MOPS. Rigid mapping



Example 2. Lab shelf photograph, approx. affine viewpoint change, 16 × 16 MOPS. Affine mapping

Figure 2. Results of the ImageJ-plug-in on different types of images. Both examples show two $400 \times 300\,\mathrm{px}^2$ images with common content under different transformation. This transformation is approximated with a rigid/affine mapping. The images are overlaid with automatically extracted consistent landmark correspondences using the MOPS descriptor. One correspondence example is selected and its respective descriptors is shown. The upper left corner of the descriptor patch is marked with an arrow. On the right, both images are mapped into each other with the approximate transformation estimated from the extracted landmark correspondences.
Example 1 shows two TEM images from consecutive serial sections of the Drosophila first instar larval brain data set by Albert Cardona. Both sections are related by an approximately rigid transformation separated by 60nm section thickness. Example 2 shows a lab shelf imaged from different viewpoints approximated by an affine mapping. Note that the implementation performs well even in presence of moderate 3d viewpoint change that introduces significant variance in occlusion.

[14] Arganda-Carreras, I., Sorzano, C. O. S., Marabini, R., Carazo, J. M., de Solorzano, C. O., and Kybic, J., "Consistent and elastic registration of histological sections using vector-spline regularization," in [*Computer Vision Approaches to Medical Image Analysis*], *Lecture Notes in Computer Science* **4241**, 85–95, Springer Berlin / Heidelberg (May 2006).

[15] Cardona, A., "TrakEM2: an ImageJ-based program for morphological data mining and 3d modeling," in [*Proceedings of the 1st ImageJ User and Developer Conference*], (May 18–19th 2006).

[16] Saalfeld, S., *Automatic inter-slice registration of tiled Transmission Electron Microscopy (TEM) images*, Diploma thesis, Technische Universität Dresden (2008).