

## Feature detection and description for image matching: from hand-crafted design to deep learning

Lin Chen, Franz Rottensteiner & Christian Heipke

To cite this article: Lin Chen, Franz Rottensteiner & Christian Heipke (2021) Feature detection and description for image matching: from hand-crafted design to deep learning, Geo-spatial Information Science, 24:1, 58-74, DOI: [10.1080/10095020.2020.1843376](https://doi.org/10.1080/10095020.2020.1843376)

To link to this article: <https://doi.org/10.1080/10095020.2020.1843376>



© 2020 Wuhan University. Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 17 Nov 2020.



Submit your article to this journal [↗](#)



Article views: 10620



View related articles [↗](#)






View Crossmark data [↗](#)



Citing articles: 26 View citing articles [↗](#)

## Feature detection and description for image matching: from hand-crafted design to deep learning

Lin Chen , Franz Rottensteiner  and Christian Heipke 

Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover, Hannover, Germany

### ABSTRACT

In feature based image matching, distinctive features in images are detected and represented by feature descriptors. Matching is then carried out by assessing the similarity of the descriptors of potentially conjugate points. In this paper, we first shortly discuss the general framework. Then, we review feature detection as well as the determination of affine shape and orientation of local features, before analyzing feature description in more detail. In the feature description review, the general framework of local feature description is presented first. Then, the review discusses the evolution from hand-crafted feature descriptors, e.g. SIFT (Scale Invariant Feature Transform), to machine learning and deep learning based descriptors. The machine learning models, the training loss and the respective training data of learning-based algorithms are looked at in more detail; subsequently the various advantages and challenges of the different approaches are discussed. Finally, we present and assess some current research directions before concluding the paper.

### ARTICLE HISTORY

Received 16 August 2020  
Accepted 26 October 2020

### KEYWORDS

Image matching; affine shape estimation; feature orientation; descriptor learning; image orientation

### Foreword

This contribution is dedicated to Prof. Gottfried Konecny at the occasion of celebrating his 90<sup>th</sup> birthday with this special issue of Geospatial Information Science, a journal initiated by Wuhan University, China, more than 20 years ago. The paper contains a review of automatic feature matching for bundle adjustment and 3D reconstruction in photogrammetry and computer vision. It has been written by members of IPI, the Institute of Photogrammetry and GeoInformation, led by Gottfried Konecny between 1971 and 1998, including a coauthor from China, who currently pursues his PhD studies at Leibniz University Hannover. The paper thus connects two aspects also important in the scientific work of Gottfried Konecny: the development of aerial triangulation and the relationship of the international scientific community with China.

Indeed, aerial triangulation was the topic of Gottfried Konecny's PhD thesis (Konecny 1962), published shortly after the landmark contributions of Brown (1958) and Schmid 1958/1959 had appeared. The thesis dealt with the use of convergent imagery to improve both, economic use due to enlarged blocks as well as coordinate accuracy due to a larger base-to-height ratio. In the following decades aerial triangulation was predominantly employed using nadir images, and software systems, developed at universities, paved the way to success in practical applications. An early example was a system based on independent models (Ackermann, Ebner, and Klein 1970); well known

bundle adjustment systems such as BLUH (Jacobsen 1980) and Bingo (Kruck 1983) were developed at IPI. Similar developments took place in Vienna (Kager 1980); this system was later extended toward the acquisition of object geometries by the second author of the current paper (Rottensteiner 2000). The focus of aerial triangulation research and development then shifted to the automatic determination of image coordinates of tie points, to which IPI contributed a very successful solution (Wang 1994). The topic was taken up by one of the authors of this paper, too (Tang and Heipke 1993, 1996; Heipke 1997); both solutions found their way into practical use. Lately, oblique and convergent imagery have again been a focus of research and development, this time because of their superior possibilities for façade interpretation in city model applications. Consequently, feature matching for the automatic determination of image coordinates in convergent imagery has attracted attention (e.g., Chen, Rottensteiner, and Heipke 2020) – this is the topic of this contribution.

The second aspect relating this contribution to Gottfried Konecny's work concerns his connection to China. As is evident from this special issue, Gottfried Konecny came to China in the late seventies of the last century already, paving the way to China's integration into ISPRS. Among many other persons he met the late Prof. Shiliu Gao, and the two of them became good friends. A few years later Prof. Gao hosted Christian Heipke, one of the authors of this paper, in Wuhan. Shiliu Gao was also instrumental in

establishing the connection between Christian Heipke and Liang Tang, resulting among other things in the work on automatic relative orientation mentioned already. Finally, in 2013 Shiliu Gao advised Lin Chen, the first author of this contribution, to join IPI for his PhD studies.

## 1. Introduction

Feature based image matching is a method to solve the correspondence problem between two or more images. Conjugate points are a requirement for the estimation of the image orientation parameters (also called pose parameters), which is a prerequisite of all geometric applications in photogrammetry, robotics and computer vision involving three dimensions. The 3D reconstruction from multiple images, Simultaneous Localization And Mapping (SLAM), Structure-from-Motion (SfM) and the generation of image mosaics all rely on image coordinates of matched conjugate features. Therefore, the quality of matching algorithms is vital for the stability and quality of the solution to those problems.

Not surprisingly, there are many publications in the domain of feature based image matching, both from the theoretical and the application perspective. This paper aims at a review of the description of local features and its related topics in feature detection, orientation and affine shape estimation. While we cover a period of 40 years, starting with the seminal paper of Barnard and Thompson (1980), the review has a focus on the period since the publication of the SIFT algorithm (Lowe 2004).

Feature based image matching consists of five steps: feature detection, affine shape estimation, orientation assignment, feature description and matching of the descriptors (Szeliski 2010). The basic pipeline of extracting local image features and descriptors is shown in Figure 1. Before giving a more detailed discussion of each step, we first provide an overview of feature based image matching.

Local features are corners or blobs appearing at some distinctive positions in the image. They are not necessarily salient image corners to the human eye, but are distinctive based on some mathematical

model. As scale differences between overlapping images are common, in particular for close-range imagery, features are normally detected in scale space to reduce the influence of scale change. After those features are determined, their position and range (characteristic scale) are known. In the next step, an affine shape correction of the feature is estimated to decrease the influence of skewness and unequal scales of the two image coordinate axes. Note that the affine transformation is an approximation of central perspective for small image windows and is typically required for large-baseline image pairs with convergent viewing directions. Subsequently, the principal orientation of the detected feature is estimated, taking into account different rotations of the two images around the viewing direction. Through those steps, the position, scale, affine shape and rotation of features are determined. According to these geometric elements, a so called feature support window around the detected feature is resampled from the original image to remove the geometric distortion. The size of the (most frequently square) feature support window is normally several times its characteristic scale. This window is the basis of feature description. During feature description, a high dimensional feature vector is derived from the feature support window; this vector is used to represent the detected feature. Descriptors are normally designed to be invariant against a limited level of geometric and illumination changes.

After descriptors are derived independently for each image, the correspondence problem can be cast as a neighborhood search in the high dimensional feature space for the feature descriptors of the different images. Two related topics are essential: the definition of a measure of similarity between potentially conjugate vectors and the computational complexity of finding these conjugate features. Based on the similarity measure, e.g. the Euclidean distance of the vectors, strategies such as the tree-based search for the nearest neighbor, thresholding the ratio between the distance to the nearest and the second nearest neighbor (Lowe 2004) are employed to find matches. Efficiently finding nearest neighbors in descriptor space is normally solved by spatial indexing of high dimensional data, e.g. Beis and Lowe (1997). As this contribution

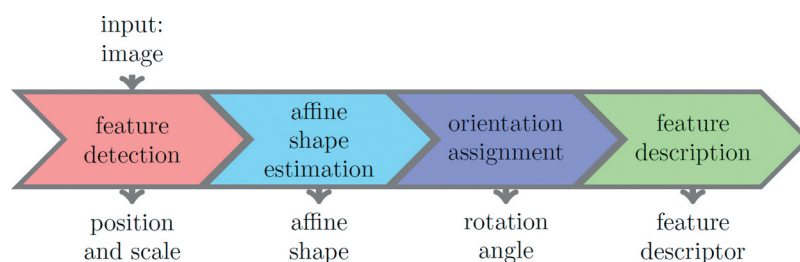


Figure 1. Feature detection and description pipeline.

concentrates on feature detection and description, the actual matching step is not further discussed in the remainder of this paper.

Recently, Ma et al. (2020) and Csurka, Dance, and Humenberger (2018) also reviewed feature based image matching algorithms. However, surveying feature based matching for images and point clouds, Ma et al. (2020) has a different scope than this paper. Csurka, Dance, and Humenberger (2018) investigated feature detection and description developed in the computer vision community. Compared to those two papers, this paper covers feature based image matching methods used in oblique aerial image orientation, which is a standard photogrammetric application. Also, neither Ma et al. (2020) nor Csurka, Dance, and Humenberger (2018) include a detailed review for feature orientation and affine shape estimation, whereas our paper reviews those two issues in detail. Moreover, this paper analyses the limitations in the current feature based image matching algorithms and, accordingly, suggests potential future research directions in this field.

In the following, details of local feature detection, feature affine shape and orientation estimation and feature description are presented in sections 2, 3 and 4, respectively. A brief overview of the applications of feature based image matching in the orientation of oblique aerial images is provided in section 5. Finally, advantages, limitations and future research directions are discussed in section 6, before concluding the paper in section 7.

## 2. Local feature detection

### 2.1. Translation and rotation invariant features

The development of so called interest operators to detect the position of features can be traced back to the work of Moravec (1979) and Dreschler and Nagel (1981). Moravec (1979) assesses average quadratic gradients in the four main directions (horizontal, vertical and both diagonals) of a local window. If the minimum of these four values is larger than a threshold, the window center is chosen as an interesting point. This reflects the simple idea that a local feature should differ from its surroundings. Dreschler and Nagel (1981), on the other side, determine pairs of maximum and minimum curvature of the gray value function in the vicinity of corners. The interest point is then defined as the zero crossing of the curvature between the two points.

The Moravec detector is not rotation invariant because gradients are estimated in four pre-defined directions. A better idea is to analyze the Hessian matrix of the auto-correlation function of the gray values  $M$  (also known as structural tensor) (Lucas et al. 1981). Its two eigenvalues contain information

of the curvature of that function. If both eigenvalues are small, the local region from which  $M$  was determined does not show much gray value variation; if one eigenvalue is large and the other one small, a strong change in one direction and thus an edge is present; if both eigenvalues are large, then the local correlation function peaks sharply and represents either a corner or a certain circular signal (Förstner 1991). Förstner calculates the eigenvalues based on the inverse of the auto-correlation matrix and suggests two indicators related to the size and similarity of the two eigenvalues (Förstner and Gülch 1987). Harris and Stephens (1988) proposes a cornerness value, which is computed as  $Det(M) - \alpha Trace(M)^2$ , where  $\alpha$  is a variable balancing determinant and trace; interest points are found by comparing the computed cornerness with a threshold. According to Rodehorst and Koschan (2006), the Förstner operator behaves slightly better than the Harris operator in terms of localization accuracy, detection and repeatability rate. Instead of the auto-correlation matrix, the Hessian matrix of the gray values in a local window can also be used to detect features (Lindeberg 1998). Based on the determinant and trace of the Hessian matrix, feature selection criteria similar to those applied to the auto-correlation matrix are derived.

The detectors discussed so far are all based on local shift-invariant windows and are therefore invariant to translation of the images. Because of the use of eigenvalues instead of gray value changes in the direction of the image coordinate axes  $x$  and  $y$ , the Förstner and Harris operators are also invariant against rotations. In addition, detectors based on the auto-correlation and the Hessian matrix are robust against small scale change. However, with increasing scale change the performance drops considerably (Rodehorst and Koschan 2006; Aanaes, Dahl, and Steenstrup Pedersen 2012). As is widely known, however, scale differences are common, especially in close-range photogrammetric applications or photo community collections, e.g. downloaded from the internet (Agarwal et al. 2011).

### 2.2. Scale invariant features

Multi-scale interest operators detect features in multiple scales and then match them across scales, see e.g. (Brown, Szeliski, and Winder 2005). However, this approach only works if the scale difference is not too large or the scale ratio is approximately known *a priori*. A more advanced method is to analyze the features using scale-space theory (Lindeberg 1998), which describes the scale space at some scale  $t$  as a convolution of the original image with the two dimensional Gaussian function with variance  $t$ . When changing  $t$  continuously rather than in discrete

steps, scale becomes a variable of a function that maps the original image to scale space. The sum of the second order derivatives of the Gaussian function in  $x$  and  $y$  directions, i.e., the Laplacian of Gaussian (LoG), normalized by the variance of the Gaussian, is used to compute the local extrema in scale-space and those extrema are selected as features, which are now scale-invariant.

In the scale invariant feature transform (SIFT) (Lowe 2004), the normalized LoG is approximated by the Difference of Gaussian (DoG), and sub-pixel localization is obtained by fitting a local 3D quadratic to the surroundings of the extrema in scale space. Today, SIFT is one of the best-known operators for feature detection (and description, see below) and performs well in matching images with scale change; SIFT can also tolerate a certain amount of affine distortion.

In Mikolajczyk and Schmid (2004) a scale selection mechanism is added to the Harris corner detector; the LoG over scale is evaluated at each detected Harris point, and points for which the LoG is an extremum are preserved. This is followed by an optional iterative refinement for both scale and position. The Hessian Laplace detector mikolajczyk2004scale, (Mikolajczyk and Schmid 2004; Mikolajczyk 2002) is similar in spirit to that work, but starts from points detected using the Hessian matrix.

### 2.3. Detectors based on a comparison of gray values or saliency

In this category, the gray value of a pixel is compared with that of the pixels in its neighborhood. If the difference is lower than some threshold, the pixels are considered to be similar. In SUSAN (Smallest Univalued Segment Assimilating Nucleus) (Smith and Brady 1997), a pixel is selected as a feature if the proportion of similar pixels in a local neighborhood is a local minimum and lower than some threshold. A further approach, FAST (Features from Accelerated Segment Test) (Rosten, Porter, and Drummond 2010), uses machine learning to accelerate the comparison process. As comparisons are only run on discrete pixels, the localization accuracy cannot be refined to sub-pixel level. This category of operators is primarily employed in applications, in which speed is essential, but high localization accuracy is not required.

### 2.4. Detectors based on machine learning

Due to different illumination conditions, the 3D shape of the object surface and potentially complex reflection functions, analyzing gray value differences between images using explicit mathematical transformations or designing features in an intuitive manner may become infeasible. An alternative is to consider

feature detection as a machine learning task. The main principle of such methods is to map an input image to a score map in which the value (score) for each pixel can be interpreted as the probability of being a distinctive feature. The parameters of the mapping functions used in this process are determined from training data by machine learning techniques, widely used in photogrammetry and remote sensing today (Heipke and Rottensteiner 2020).

An example for such an approach is Verdie et al. (2015). To take into account changes in illumination, a regressor is trained to predict a score map whose maxima are points with high repeatability under challenging illumination changes; afterward, the features expected to be most stable against illumination change are extracted by non maximum suppression. The LIFT (Learned Invariant Feature Transform) detector (Yi et al. 2016b) contains similar ideas and is designed to better discriminate between matched and non-matched feature pairs in a global fashion.

The core idea of the covariant detector is that features detected in the original image patch that are transformed to another coordinate system using some geometric transformation should have the same positions in that coordinate system as features detected after applying the geometric transformation to the original image patch. This so called covariance constraint is used in Lenc and Vedaldi (2016), in which a regressor is employed as a detector to map an image patch to a feature response.

If an input image is converted to a response map using trainable models, the top/bottom quantiles of the distribution of the response values can be used to define thresholds for selecting the best features according to the criterion implied by the response map. According to Savinov et al. (2017), the order of those top/bottom quantiles should be kept constant before and after the input image is geometrically transformed. In that paper, a quad-network, composed of two original and two transformed image patches, is used to learn an order-preserving feature detection network. As mentioned above, Rosten, Porter, and Drummond (2010) use machine learning to improve the speed and repeatability of feature detection based on a gray value comparison of pixels in the neighborhood.

## 3. Feature orientation and affine shape estimation

After feature location and scale in the image plane are detected, an image patch surrounding the detected feature is typically extracted with a size proportional to its scale. This patch reflects the appearance of the underlying feature and is used as input for descriptor computation. Due to potentially different rotations around the viewing direction and also different

viewing directions, patches of conjugate features can be distorted with respect to each other.

Simply requiring the descriptor to be invariant against these relative distortions of the feature support windows decreases the discriminative power of the descriptors. As mentioned above, these distortions can be modeled as rotation and affine distortion, the latter considering local skewness and scale differences between the two axes of the image coordinate system. The rotation difference can then be determined by finding a principal orientation for each image patch surrounding a detected feature, and computing the descriptor based on that principal direction. Similarly, the affine shape can be estimated and the patches are then re-sampled to compensate rotation and affine distortions between image patches.

### 3.1. Orientation assignment

After features are detected, the orientation of a feature can be estimated by calculating a principal direction using the gradients calculated in a local window surrounding the detected feature. In SIFT (Lowe 2004), a histogram of oriented gradients is calculated; the bin with the maximum count is then picked, and the corresponding direction is refined by fitting a parabola to the peak and the two adjacent bins. Other bins with high values, i.e. larger than 80% of that of the maximum bin, are retained as secondary principal orientations. Features with multiple peaks in the support window can be better matched in this way. In SURF (Speeded-Up Robust Feature) (Bay et al. 2008), Haar wavelet responses in horizontal and vertical directions inside a circular window surrounding the detected feature are computed and plotted as points in 2D. Responses within a rotating cone of size  $\pi/3$  are summed and the principal direction is assigned to the cone direction with the highest result. Also, the mean gradient in a small window surrounding the detected feature has proven to be helpful in aligning features (Brown, Szeliski, and Winder 2005).

In Yi et al. (2016a) and Yi et al. (2016b), the orientation is estimated by deep learning. The principal direction for an input patch surrounding a detected feature is predicted by a Siamese Convolutional Neural Network (CNN), which maximizes the similarity of descriptors calculated for pairs of conjugate input feature patches. A similar idea is used in Mishkin, Radenovic, and Matas (2018) and Chen, Rottensteiner, and Heipk (2020) to learn the orientation of local features. This strategy shows significantly better performance than the aforementioned methods based on hand crafted features.

### 3.2. Affine shape estimation

Detecting local features in scale space and assigning them an orientation is basically equivalent to normalizing rotation and scaling of local features before description. However, this transformation is not sufficient to model the geometric transformations between local image patches in case of large changes in view-point and viewing direction between images. Therefore, perspective changes, which for small windows can be compensated by an affine transformation, should also be estimated and taken into account before feature description.

Invariance against affine transformations has also been investigated. The method of Edge-Based Regions (EBR) (Tuytelaars et al. 1999) uses edges starting from detected Harris points to construct affine invariants. This method can only be applied to features surrounded by edges, consequently the application range is limited. The approach called Intensity Based Region (IBR) (Tuytelaars and Van Gool 2000, 2004) starts from one detected feature point and constructs lines of different directions; in each direction the line ends at the local maximum of gray value change in a pre-defined neighborhood. Then, an ellipse is fitted through all end points, and the ellipse is used to represent the underlying feature. In Matas et al. (2004), the watershed algorithm is used to find local extrema employed for ellipse fitting. As reported in Mikolajczyk et al. (2005) the features extracted in this way are sensitive to scale change.

Affine shape estimation theory using the second moment matrix is studied in Lindeberg and Garding (1997); Baumberg (2000); Mikolajczyk (2002); Mikolajczyk and Schmid (2004). Affine Gaussian scale-space is generated by convolution of the image patch with a non-uniform Gaussian kernel, which is represented by a  $2 \times 2$  covariance matrix  $\Sigma$ . As indicated in Lindeberg and Garding (1997), in affine Gaussian scale-space the response of two image patterns related by a particular affine transformation  $B$  is equal, if for the underlying Gaussian kernels  $\Sigma_1$  and  $\Sigma_2$  the following relation holds:  $\Sigma_2 = B^T \Sigma_1 B$ . The second moment matrix  $M$  measures the level of isotropy of a feature and is thus used to describe the covariance matrices, thus  $M_2 = B^T M_1 B$ . Based on that, an iterative procedure is proposed to derive features for which the shape is approximately circular. In Baumberg (2000) the patch surrounding the features is normalized by geometrically transforming the patch using  $M^{-1/2}$  as the transformation matrix.

After that, the second moment matrix for the normalized patch is iteratively calculated and the patch is yet again transformed using  $M^{-1/2}$ , until the two eigenvalues of  $M$  for the normalized patch are close enough to each other. As a result, the affine transformation between two image patches is removed

and only a rotation remains. In Mikolajczyk and Schmid (2004), this approach is extended to Gaussian scale-space and Harris-Affine and Hessian-Affine detectors are employed to select the features.

However, a considerable amount of features is removed after the iterative affine adaptation algorithm. According to Mikolajczyk and Schmid (2004), only 20–30% of the initially detected features are preserved for further feature matching. To tackle this problem, affine shape estimation based on a deep neural network is proposed in Mishkin, Radenovic, and Matas (2018), where the shape parameters are estimated by minimizing the distance between the descriptors of matching point pairs. A model to predict affine shape is also learned in Chen, Rottensteiner, and Heipk (2020). In addition, it is used to match images taken from oblique aerial cameras, resulting in a notable performance improvement compared to using hand-crafted algorithms for removing affine distortion. ASIFT (Affine SIFT) (Morel and Yu 2009), on the other hand, first simulates different versions of the input image by applying different sets of affine transformation parameters. In a second step, the DoG features and SIFT descriptors detected in each transformed image are combined for descriptor matching, a process that makes ASIFT computationally expensive. ASIFT thus does not take algorithmic measures to estimate local affine shape for each feature, but makes matching more invariant toward affine distortions by applying a standard algorithms to different simulated views.

An affine invariant measure combining points and lines in a local neighborhood is proposed in Chen and Shao (2013). The method starts by extracting points using the DoG detector (Lowe 2004) and finding potential candidates in the other image based on the DAISY descriptor (Tola, Lepetit, and Fua 2009). After

that, point pairs  $X_1$  and  $X_2$  that are spatially close to each other in the first image and their correspondences  $Y_1$  and  $Y_2$  are considered. For a line  $L_1$  located in the neighborhood of  $X_1, X_2$  in the first image, the ratio between the distances of  $X_1, X_2$  from  $L_1$  is considered to be an affine invariant feature. If there is a line  $L_2$  located in the neighborhood of  $Y_1$  and  $Y_2$  that is a good match for  $L_1$ , the ratio of the distances of  $Y_1, Y_2$  from  $L_2$  must be close to the distance ratio calculated in the first image. By applying this affine invariant feature as similarity measure, Chen and Shao (2013) successfully match high resolution remote sensing images of scenes that are planar or dominated by planar objects.

A graph summarizing the development of feature detection, orientation and affine shape estimation is shown in Figure 2.

#### 4. Local feature description

In essence, the description of features is a problem of representation. It transforms the detected features into a new space called feature descriptor space, where different features can be more easily discriminated and matched. A local context window surrounding the feature is typically used to build the descriptor. Designing descriptors is difficult, because there are many factors influencing the gray values in the feature support window. A descriptor should be invariant with respect to different limited changes, e.g., illumination change, rotation or affine distortions. The central problem of descriptor design is to achieve invariance against these transformations. The simplest descriptor is the combination of pixel gray values in this feature support window. However, this description is sensitive to photometric and geometric changes. Thus, many alternative descriptors have been proposed.

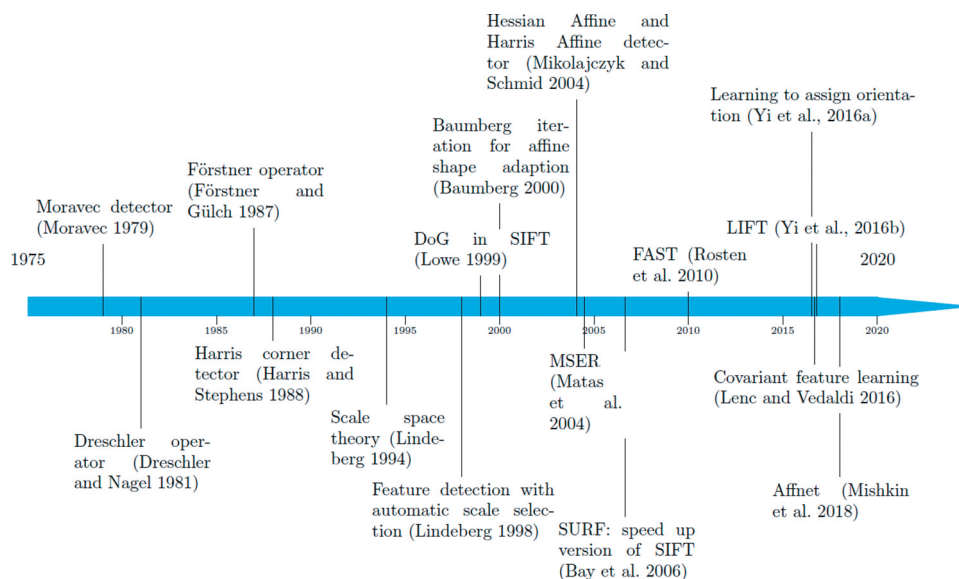


Figure 2. A timeline highlighting some landmark papers for feature detection, orientation assignment and affine shape adaption.

There are two groups of descriptors: floating point descriptors and binary descriptors, depending on numerical type used to represent the elements of the descriptors. Floating point descriptors are designed for better discriminability, but are normally computationally more expensive. Binary descriptors are designed for applications where computational resources are sparse, e.g. in real time SLAM or tracking.

According to Brown, Hua, and Winder (2011), the determination of descriptors consists of several steps: transformation, aggregation, normalization, and dimension reduction. Transformation is often some basic filter operation, e.g. calculating gradients or the Haar feature responses. Filters are often designed to preserve and magnify some special local patterns; they are applied either to the whole feature support window or only to special positions. Aggregation comprises an integration procedure, e.g. maximum or mean value computation or histogram generation. It is based on small regions in the feature support window, the arrangement of which needs to be determined *a priori*. Aggregation increases the robustness against noise and results in some invariance against a limited level of transformations between images. Normalization transforms a particular value to a fixed range and alleviates the influence of the absolute response; this improves the robustness against illumination change. As some descriptors are high-dimensional and different dimensions can have strong correlations, e.g., because of spatial correlation in the original feature support window, dimension reduction, e.g. based on principal component analysis (PCA), is applied in a number of algorithms to obtain a more compressed and less redundant representation.

This review starts with giving an overview over hand-crafted descriptors in section 4.1 before discussing approaches based on machine learning in section 4.2.

## 4.1. Hand crafted descriptors

### 4.1.1. Classical feature descriptors

This group contains the well-known descriptors SIFT (scale invariant feature transform [Lowe 2004]) and SURF (speeded-up robust feature [Bay, Tuytelaars, and Van Gool 2006]). These descriptors integrate a large scope of knowledge researchers had accumulated in feature description earlier. SIFT first calculates the gradients in the features support window, followed by Gaussian filtering to assign the central pixel a larger weight, and then aggregates the gradients in square grids aligned with the main feature orientation. SURF uses the Haar wavelet response as basic transformation, aggregation also takes place in grids. Variants of these descriptors are DAISY (Tola, Lepetit, and Fua 2009), which uses steerable filters and performs

accumulation in circular patterns, and PCA-SIFT (Ke, Sukthankar, and Society 2004), which decreases the correlation between SIFT descriptor dimensions by employing PCA, where the projection basis is trained using on a large number of collected SIFT descriptors.

### 4.1.2. Binary descriptors based on gray value comparison

The comparison of pairs of pixel gray values inside the feature support window, followed by storing the result of the comparison as a binary number, is the basis of binary descriptors. BRIEF (Binary Robust Independent Elementary Features) (Calonder et al. 2010, 2012) employs different pixel positions in the smoothed feature support window under different distributions. ORB (Oriented BRIEF) (Rublee et al. 2011) adds orientation estimation and computes the descriptor in a greedy search selecting 256 pairs of pixel positions that can best discriminate homologous features in a training set.

BRISK (Binary Robust Invariant Scalable Keypoints) (Leutenegger, Chli, and Siegwart 2011) identifies the orientation of keypoints detected in scale-space and conducts the gray value comparison in concentric circles, a pattern similar to DAISY. The radius of the circles increases with the distance of the sampling point from the center of the patch. The comparison is performed between pairs of circles, the distance of which is smaller than some threshold (short-distance pairing). Finally, FREAK (Fast Retina Keypoint) (Alahi, Ortiz, and Vandergheynst 2012) selects the pairs for comparison based on the knowledge of the human retina. FREAK also samples in a circular pattern, but comparisons concentrate in the region near the center of the feature support window.

## 4.2. Descriptors based on machine learning

The design of a descriptor can be defined as a machine learning problem, where the descriptor is computed by a function with trainable parameters and the learning objective is to increase the similarity of conjugate pairs of features, while decreasing the similarity of non-conjugate pairs. The way to map the feature support window to descriptor space can vary widely. In this section, some commonly used types of functions are reviewed.

### 4.2.1. Transformation-embedding-pooling form

Descriptors trained for feature-based matching were first proposed in Winder and Brown (2007), where different combinations of transformation, embedding and pooling are learned jointly to achieve a discriminative descriptor. The experiments reported in these papers indicate promising performance



improvements compared to hand-crafted features. However, for each of the four aforementioned components, a separate loss to be optimized in training is designed, which leads to difficulties in the training procedure. Consequently, rather complex optimization strategies are required to find a good solution. A convex objective function was proposed in Simonyan, Vedaldi, and Zisserman (2014) to tackle that problem; as a result a notable performance improvement was achieved.

Another important contribution of Brown, Hua, and Winder (2011); Winder and Brown (2007) is a training dataset, called the Brown dataset, which is widely employed in later works. In some publications the term Photo Tourism dataset (Snavely, Seitz, and Szeliski 2008) is used as a synonym for the Brown dataset, although the former contains many more images than the latter. The Brown dataset relies on 3D reconstruction from multiple view images. For each image in the dataset, dense stereo matching and image orientation results are utilized to retrieve ground-truth matches.<sup>1</sup> Therefore, realistic uncertainties for the conjugate features are contained in the training data.

#### 4.2.2. Comparison based feature descriptors

In Lepetit and Fua (2006) multiple random trees are trained to recognize matched features, where gray value comparisons at different positions of the feature support window are used as node tests. The so-called random fern (Ozuysal et al. 2010) uses a naive Bayesian combination of classifiers to achieve even better performance. Another category of descriptors is based on boosting. In Trzcinski, Christoudias, and Lepetit (2015), (2013)), boosting is used to select carefully designed weak features which rely on gray value gradients over rectangular image regions. Then, each of those features is compared to a trainable threshold and converted to a binary value that forms one dimension of an output descriptor. Chen, Rottensteiner, and Heipke (2014) learn Haar-like features which best classify the matching and non-matching pairs using adaptive boosting (Viola and Jones 2004). The descriptors proposed in these papers can also be classified as binary learned descriptors, because they deliver descriptors in a binary form.

#### 4.2.3. Descriptors based on deep neural networks

For learning descriptors based on matching and non-matching pairs, Siamese CNN have proven to be very suitable. The term "Siamese" refers to the fact that two branches of a CNN, each serving as a feature extractor for one of the input image patches, share the same parameters. Siamese CNNs were already proposed for the extraction of descriptors for the verification of human signatures in Bromley et al. (1994). A Siamese CNN is also used in Hadsell, Chopra, and

LeCun (2006) to extract a compact descriptor for the recognition of hand-written digits. In descriptor space, identical digits form clusters, removing the effect of appearance change caused by different writing styles and thus achieving invariance for representing those digits.

Jahner, Grabner, and Bischof (2008) treated a Siamese CNN as a recognition network which delivers a class label for the support window of each detected feature in one image; thus, the number of classes is determined by the amount of detected features in the image. Geometric transformations are simulated for both branches of the Siamese CNN to encourage the CNN to become invariant against geometric distortions. However, when matching images from a new scene, the class labels change and therefore the Siamese CNN must be retrained.

Similar to images of digits written by different people, which may be affected by person-specific variations, feature support windows of conjugate features from different views can also contain complex geometric and/or radiometric differences against which the descriptor should be invariant. Therefore, Siamese CNNs fit naturally to the task of feature matching. To the best of our knowledge, Osendorfer et al. (2013) is the first work to use a Siamese CNN to train descriptors for feature matching, although the authors concentrates on comparing four different types of loss functions. Carlevaris-Bianco and Eustice (2014) employ a Siamese CNN to achieve illumination invariance. Images with severe illumination changes are fed into the network branches; the obtained invariance exceeds the one of hand-crafted descriptors. Siamese CNNs are further used to learn descriptors in Zagoruyko and Komodakis (2015); Han et al. (2015); Simo-Serra et al. (2015); Chen, Rottensteiner, and Heipke (2016).

Instead of measuring the similarity of descriptors using their Euclidean distance, an additional metric network can be employed to directly predict the probability of a correct match for an input patch pair (Zagoruyko and Komodakis 2015; Han et al. 2015). Not surprisingly, the work based on metric learning performs better in discriminating feature pairs, as its similarity measure is not a simple distance measure, but a more advanced trained similarity score. However, for an actual matching task the metric network needs to be run for every possible combination of feature pairs computed from the different images. As a consequence, the computation is expensive and its usage is restricted. This is the reason why most of the current work in this field concentrates only on learning the descriptors, i.e., on finding a good embedding that can discriminate features by simple distance measures. As stated in Simo-Serra et al. (2015), most of the negative pairs cannot contribute to the loss after the training process has run for a while. The solution

given in Simo-Serra et al. (2015) is hard mining, which means that only a fraction of the samples that contribute a higher amount of loss are selected for parameter updating in each iteration step during training.

The triplet architecture, proposed by Chechik et al. (2010), has been used for training descriptors in Kumar et al. (2016) for the first time. Triplet networks are composed of three branches, namely an anchor ( $a$ ), a positive ( $p$ ) and a negative ( $n$ ) branch. Anchor and positive branch correspond to a matching pair, anchor and negative branch to a non-matching pair. The distances  $d(a, p)$  and  $d(a, n)$  are used to build the loss function. Compared to a Siamese CNN that optimizes matched and unmatched parts independently, the triplet architecture pushes unmatched features away from similar features in the descriptor space and thus equally considers similar and dissimilar features. The triplet loss used in Hoffer and Ailon (2015) motivates  $d(a, n)$  to be larger than  $d(a, p)$ , followed by adding a margin (Balntas et al. 2016b) between  $d(a, n)$  and  $d(a, p)$ ; this strategy was already suggested in Chechik et al. (2010). Balntas et al. (2016a) propose to use either  $d(a, n)$  or  $d(p, n)$  as negative loss, depending on which of the two is smaller.

Instead of checking each triplet separately, Kumar et al. (2016) build a global loss function to separate the distribution of distances for matched and unmatched pairs. In a global loss function, the variance of the distances for matched and unmatched pairs is minimized; the mean of the distances for matched pairs is also minimized, while the mean for unmatched pairs is maximized. The authors test the proposed loss for different architectural details and achieve noticeable improvements compared to normal loss functions.

One of the major concerns in Siamese and triplet CNNs for descriptor learning is that very few unmatched pairs are seen in training. This is in contrast to the typical application scenario for a trained descriptor, e.g. feature matching or image retrieval, where much larger numbers of unmatched pairs need to be checked in comparison to matched pairs. To improve the situation, progressive sampling in L2-Net (Tian, Fan, and Wu 2017) uses the hardest unmatched pair for calculating the loss. The main loss is the ratio between  $d(a, p)$  and  $d(a, n_{hardest})$ , for which a small distance is desired for a matched pair and a large one for an unmatched pair. Another constraint is to minimize the correlation between different dimensions of descriptors. Also, the similarity of feature maps in the descriptor network is encouraged to be high for matched features, but low for unmatched pairs. L2-Net achieves a remarkable performance improvement. Similar ideas of finding the hardest unmatched pair are explored in Mishchuk et al. (2017), where the closet non-matching patch to  $a$

and  $p$  in the triplet is found and a margin between the distance for a matched pair and the closet unmatched pair is included in the loss. In Mishchuk et al. (2017), a slightly better result than the one achieved for L2-Net is reported, though the additional constraint used by L2-Net (Tian, Fan, and Wu 2017) is ignored.

Although the distance of a matched pair and the hardest unmatched pair in a triplet is restricted by either setting a relative ratio or a margin between them, training based on such a strategy might result in a large cluster radius for matched features in descriptor space. To avoid such a large cluster radius for matched features, Keller et al. (2018) propose to balance the distance for both matched and hardest unmatched pairs and the “soft” margin or ratio between the distance for matched and hardest unmatched pairs in triplets. The result of their method shows a consistent improvement compared to Tian, Fan, and Wu (2017); Mishchuk et al. (2017).

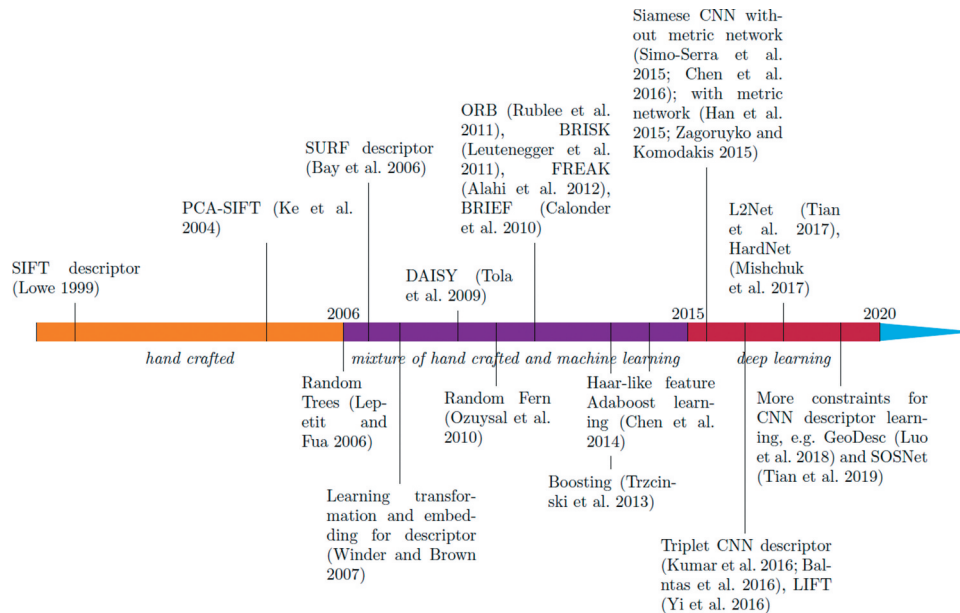
Luo et al. (2018) concentrates on generating more realistic yet challenging matched pairs for L2-Net. The angle between two intersecting rays pointing at the same 3D point from different views (corresponding to matching features) and the incidence angle (angle between local normal and the ray) difference between each pair of rays are used to model the difficulty of matching homologous features. The authors discard easier pairs and only use more difficult matching pairs for sampling. Better performance in matching and retrieval benchmarks is achieved by that method.

A graph showing the timeline of the main developments in feature description is shown in Figure 3.

## 5. An application: orientation of oblique aerial images

In this section, the application of feature based image matching is reviewed with a focus on the orientation of imagery taken from oblique aerial camera systems. For nadir or near nadir images, classical feature based image matching algorithms, e.g. SIFT and SURF, work well. The focus on image orientation for oblique imagery comes from the fact that this is a more challenging task because of large changes in viewing direction and viewpoint between the different images, exceeding the invariance of classical feature based image matching algorithm, as reported in Verykokou and Ioannidis (2018, 2016); Jacobsen and Gerke (2016).

In Smith et al. (2008), conjugate points between the oblique and vertical images are still collected interactively. The reason given by the authors is that the illumination and viewing direction differences led to the failure of automatic tie point generation. An early attempt to match oblique and nadir images automatically was based on SURF (Verykokou and Ioannidis 2016); the results showed that most of the resulting



**Figure 3.** The literature timeline for feature description. For better viewing of the whole graph, some closely linked papers are combined into one block with an +1 year difference, e.g. all the binary descriptors including ORB, BRISK, BRIEF and FREAK are combined into one block.

conjugate points were situated in a planar surface within a limited degree of viewpoint change only. The method is bound to fail if the scene contains larger elevation differences. The idea of running multiple homographies is used in Onyango et al. (2017) to obtain tie points between UAV and oblique camera images. In this strategy, a first homography is computed, outliers of this result are then iteratively used to compute further homographies.

Kim, Nam, and Lee (2019) propose a deep neural network to estimate the rotation and affine transformation between a pair of images. A qualitative performance between nadir and oblique images from the ISPRS multi-platform photogrammetry dataset (Nex et al. 2015) shows that the method can roughly align images from different view points. However, the estimated transformation angles are discretized into large intervals, e.g.  $45^\circ$ , which restricts its practical usage. Moreover, from a theoretical point of view, the method can yield correct results only for scenes containing a low amount of height change.

Matching of nadir and oblique images can also rely on view sphere simulation of images using ASIFT (Affine Scale Invariant Feature Transform), e.g. Wang et al. (2018). Similarly, Shao et al. (2020) use ASIFT to match a digital orthophoto and image frames from videos taken from oblique viewing directions. The idea of view sphere simulation is to simulate a lot of images with different levels of affine transformation between two images which need to be matched. Feature detection, orientation and description are then performed based on all transformed images. Not surprisingly, for some pairs, the geometric distortions are decreased to a level which can be handled by classical matching algorithms. Once these pairs are

successfully matched, their conjugate points are transformed back to the original images, thus obtaining matches in the original image frames.

Finally, a recently published approach successfully matches images acquired using an oblique penta camera system by optimizing the affine shape estimation, orientation assignment and feature description directly in a deep neural network (Chen, Rottensteiner, and Heipk 2020). A comparison to other feature-based matching frameworks shows that the new method delivers more matches between image pairs having very different viewing direction and, thus, a more stable block geometry.

Unfortunately, only limited information is reported for the matching algorithms used in commercial software, e.g. Pix4D or Photoscan. Open source software for 3D image reconstruction from multiple images, e.g. Bundler<sup>2</sup> (Snavely, Seitz, and Szeliski 2006), VisualSFM<sup>3</sup> (Wu et al. 2011; Wu 2013), MicMac<sup>4</sup> and COLMAP<sup>5</sup> (Schönberger and Frahm 2016), mainly relies on SIFT for feature based image matching.

## 6. Discussion

In this section, an analysis of the different modules of feature based image matching is provided. Along with the discussion, some research directions for advancing image matching based on deep learning are suggested.

### 6.1. Detector learning

Detector learning algorithms are built on the core idea of enhancing the repeatability of detecting features

against geometric or illumination changes. In photogrammetry and remote sensing, the invariance of feature detection against radiometric changes is in urgent demand not only for classical image matching. Related topics that rely on stable feature detectors comprise matching of remote sensing images from different sensors (e.g. SAR and optical images) and of multi-modal sensor data (e.g. LiDAR and RGB or multi-spectral images). To solve these problems, multi-sensor and multi-modal datasets which contain ground truth geometric registration are needed. These datasets can then be employed to learn the required detectors in ways similar to what was discussed in this paper.

## 6.2. Orientation assignment and affine shape estimation

When assigning an orientation to a feature support window for image matching, only the relative orientation between the two images is relevant, as they can be matched in any absolute orientation. The same is true for the two parameters of affine distortion (different scale in the two axes of the image coordinate system and skewness). Therefore, orientation assignment and affine shape estimation for a single patch does not have a unique solution. In Yi et al. (2016a) and Yi et al. (2016b), it is claimed that orientation and affine parameters optimized only by using descriptor distance loss motivates the method to find the best possible solution for image matching. However, an underlying assumption to ensure that this idea works is that there is a distinctive descriptor distance minimum when two feature support windows are aligned, i.e., a unique solution exists for the problem. Of course, as far as the computation of image coordinates of tie points for image orientation is concerned, any of the solutions provided by the network is as good as any other. However, the numeric stability and convergence properties of the computations should be investigated.

In current work for the orientation and affine shape estimation, descriptors are needed to build a descriptor distance based loss. However, in classical feature based image matching, e.g. SIFT (Lowe 1999, 2004), orientation assignment and affine shape estimation (Baumberg 2000; Mikolajczyk and Schmid 2004) depend only on geometric measures calculated on individual patches surrounding features. Thus, patches are transformed into some canonical form in which matching can be performed unambiguously and independently of descriptor distance. An open question is whether this idea can be transferred to deep learning based approaches as well.

Although joint learning of orientation and affine shape estimation is theoretically possible, it has not been conducted in published work. Mishkin, Radenovic, and Matas (2018) as well as Chen, Rottensteiner, and Heipk (2020) report on unsuccessful attempts to do so. It is currently unclear what the specific difficulty for training the orientation and affine shape module jointly is. Through the analysis of how descriptor distance changes for pairs of patches simulated with different amounts of rotation and affine distortions, it is possible to further investigate this issue.

## 6.3. Descriptor learning

To learn descriptors from data, pairs or triplets of matching or non-matching image patches are fed into the learning framework. The loss function is designed to make the similarity, often measured by the inverse Euclidean distance between descriptors, a maximum for matching pairs and a minimum for negative pairs, i.e. pairs not corresponding to a correct match. An identical number of positive and negative pairs are typically used in the training procedure to update the parameters of the network for predicting the descriptors. However, in real applications, for instance in image matching or image retrieval, far more negative pairs must be compared than positive pairs, because finding correspondences requires a "one against many others" processing strategy. Taking this imbalance into consideration, the negative pairs are mined by comparing and selecting difficult ones from a pool containing large amounts of such pairs. Therefore, the network sees a significantly larger number of negative samples (pairs not corresponding to a correct match) than positive ones (matching pairs) during training. Methods trying to include more negative pairs are presented in Mishchuk et al. (2017) and Simo-Serra et al. (2015); the results reported in these papers indicate that the discriminability of the learned descriptors can be improved notably in this way.

On the other hand, the appearance of matching patches has not been explored in-depth in descriptor learning. Compared to "seeing" unmatched pairs, the descriptor has limited chance to explore the intra-class variability of matching pairs. In other words, for each patch, the descriptor has a much lower chance to see conjugate patches containing a limited level of distortion. In training data currently in use, only few conjugate patches are contained for each feature and during training the matching pairs are sampled from these correspondences. These patches can only cover a small part of the space of possible viewing directions that would result in patches for which the descriptor should also be similar to the one for a given reference patch.

#### 6.4. An aerial photogrammetric benchmark

To evaluate the different variants of descriptors, results from 3D reconstruction from images are employed, e.g. Fan et al. (2019) and Jin et al. (2020). The test dataset used in Fan et al. (2019) consists of images containing consecutive changes of viewpoint and viewing directions. Although some overlapping images are indeed wide baseline pairs, for each image an overlapping partner can be found with only small changes in viewing direction and viewpoint position. In Jin et al. (2020), the Photo Tourism dataset is used as wide baseline dataset for the 3D image reconstruction task. A check revealed that this dataset contains a large number of consecutive views also.

In aerial photogrammetry, matching of oblique images has been one of the more attractive research areas in the last years, partly due to its practical importance. While the images taken by different cameras of an oblique camera system contain distinct and large viewpoint and viewing direction changes, consecutive images as in the datasets mentioned before typically do not exist. Unfortunately, the only publicly available dataset comprising oblique aerial imagery (Nex et al. 2015) does not contain a dense ground truth depth map in order to derive as many conjugate points as necessary for the evaluation in a way similar to Brown, Hua, and Winder (2011). This requirement largely comes from the fact that correct matches can be removed during the computation of the image orientation parameters, e.g. when the triangulation angle of a match is small. While such matches do not contribute to the computation of the image orientation parameters, it is still interesting to investigate the matching quality of such conjugate points. In this way, besides the bundle adjustment output, which of course forms the primary result for image orientation, evaluation criteria for feature based image matching, e.g. recall, precision, matching points distribution, repeatability of detected features (Mikolajczyk et al. 2005) and matching score (Mikolajczyk and Schmid 2005) can be analyzed in detail as well, in order to assess the influence of the matching algorithm on the image orientation results.

#### 6.5. Transferability of methods based on machine learning

The possibility to transfer training-based methods for descriptor computation, orientation assignment and affine shape estimation between different imaging domains has not been properly investigated yet. Thus, unlike for other machine learning tasks, e.g. semantic segmentation and person re-identification, it is unclear how well methods for image matching such as those based on deep learning can be

transferred from the domain from which the training dataset was generated, e.g. a close range scene, to another domain, e.g., a set of aerial images.

On the one hand, machine learning tasks related to feature based image matching, especially affine shape estimation, orientation assignment and descriptor learning, are based on simple networks and a simple form of distance based loss functions, which should be beneficial for the generalization ability of the learned models. In addition, all the context windows involved in these tasks are "local", having a size several times that of the detected characteristic scale of the features. Thus, a limited range of context is involved, which also increases the possibility for a wide generalization. Nevertheless, in order to obtain a better understanding of this question, the transferability of feature matching algorithms based on deep learning across imaging domains should be systematically analyzed based on a series of datasets containing significant differences, e.g. sets of street view and aerial images.

### 7. Conclusion

In this paper, the four steps of the feature detection and description workflow for feature based image matching, namely feature detection, affine shape estimation, orientation assignment and feature description, are reviewed. During the past decade these modules have gradually evolved from a hand-crafted design to prediction by trained deep neural networks. After the technical review of the different algorithms available in the literature and the orientation of oblique aerial images as an application example, an analysis of the limitations and gaps between the current possibilities of feature based image matching and the requirements of photogrammetry and remote sensing is provided, and potential interesting research directions are suggested.

We believe that the potential of deep learning in solving matching problems in photogrammetry and remote sensing has been proven, but is still far away from being fully discovered. On the one hand, automation and in particular the use of deep learning decreases the dependence on experience and knowledge of specific application domains. On the other hand, there is a danger of using these methods as black boxes without proper checks of the results. In our point of view, a further important issue is thus how to integrate common and expert domain knowledge with deep neural networks to act as a guidance or regularization.

### Notes

1. For a feature  $f_L$  in image  $I_L$ , a small grid surrounding  $f_L$  in  $I_L$  is extracted and transferred to image  $I_R$  using the depth map estimated from the stereo image pair  $I_L, I_R$ . The transferred grid is then used to estimate the scale and pixel localization for the transferred

feature point in  $I_R$ . If the difference of estimated scale and pixel localization for the transferred features point is close to the scale and localization of a feature point  $f_R$  in  $I_R$ , then  $f_R$  and  $f_L$  are considered to be a ground truth match.

2. <http://www.cs.cornell.edu/snavey/bundler/>
3. <http://ccwu.me/vsfm/>
4. <https://micmac.ensg.eu/index.php/Accueil>
5. <https://demuc.de/colmap/>

## Acknowledgments

The authors would like to thank NVIDIA Corp. for donating the GPU used in this research through its GPU grant program. The first author Lin Chen would also like to thank the China Scholarship Council (CSC) for financially supporting his PhD study.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Notes on contributors

**Lin Chen** is a PhD student at the Institute of Photogrammetry and GeoInformation (IPI) of Leibniz University Hannover. He received his Bachelor and Master degrees in Geodesy and Geomatics Engineering from Wuhan University, China. His current research interests are feature matching based on deep learning, dense image matching, and semantic/instance segmentation of remote sensing images.

**Franz Rottensteiner** is an associate professor and leader of the research group “Photogrammetric Image Analysis” at the Institute of Photogrammetry and GeoInformation (IPI) of Leibniz University Hannover. He received the Dipl.-Ing. degree in surveying and the PhD degree and *venia docendi* in photogrammetry, all from Vienna University of Technology (TUW), Vienna, Austria. His research interests include all aspects of image orientation, image classification, automated object detection and reconstruction from images and point clouds, and change detection from remote sensing data. Before joining LUH in 2008, he worked at TUW and the Universities of New South Wales and Melbourne, respectively, both in Australia. He has authored or coauthored more than 150 scientific papers, 36 of which have appeared in peer-reviewed international journals. He received the Karl Rinner Award of the Austrian Geodetic Commission in 2004 and the Carl Pulfrich Award for Photogrammetry, sponsored by Leica Geosystems, in 2017. Since 2011, he has been the Associate Editor of the ISI-listed journal “Photogrammetrie Fernerkundung Geoinformation”. Being the Chairman of the ISPRS Working Group II/4, he initiated and conducted the ISPRS benchmark on urban object detection and 3D building reconstruction.

**Christian Heipke** is a professor of photogrammetry and remote sensing and head of the Institute of Photogrammetry and GeoInformation (IPI) at Leibniz University Hannover, where he currently leads a group of about 25 researchers. His professional interests comprise all aspects of photogrammetry, remote sensing, image understanding and their connection to computer vision and GIS.

His has authored or coauthored more than 300 scientific papers, more than 70 of which appeared in peer-reviewed international journals. He is the recipient of the 1992 ISPRS Otto von Gruber Award, the 2012 ISPRS Fred Doyle Award, and the 2013 ASPRS Photogrammetric (Fairchild) Award. He is an ordinary member of various learned societies. From 2004 to 2009, he served as vice president of EuroSDR. From 2011–2014 he was chair of the German Geodetic Commission (DGK), from 2012–2016 Secretary General of the International Society of Photogrammetry and Remote Sensing (ISPRS). Currently he serves as ISPRS President.

## ORCID

Lin Chen  <http://orcid.org/0000-0002-9839-476X>

Franz Rottensteiner  <http://orcid.org/0000-0003-1942-8210>

Christian Heipke  <http://orcid.org/0000-0002-7007-9549>

## Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## References

- Aanaes, H., A. Dahl, and K. Steenstrup Pedersen. 2012. “Interesting Interest Points.” *International Journal of Computer Vision* 97 (1): 18–35. doi:10.1007/s11263-011-0473-8.
- Ackermann, F., H. Ebner, and H. Klein. 1970. “Ein Programmpaket für die Aerotriangulation mit unabhängigen Modellen.” *Bildmessung und Luftbildwesen* 38 (4): 218–224.
- Agarwal, S., Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. 2011. “Building Rome in a Day.” *Communications of the ACM* 54 (10): 105–112. doi:10.1145/2001269.2001293.
- Alahi, A., R. Ortiz, and P. Vandergheynst. 2012. “Freak: Fast Retina Keypoint.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 510–517, Rhode Island, USA, June 18–20.
- Balntas, V., E. Johns, L. Tang, and K. Mikolajczyk. 2016a. “PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors.” *arXiv preprint arXiv:1601.05030*.
- Balntas, V., E. Riba, D. Ponsa, and K. Mikolajczyk. 2016b. “Learning Local Feature Descriptors with Triplets and Shallow Convolutional Neural Networks.” *Paper presented at the Proceedings of the British Machine Vision Conference*, 1–11, York, UK, September 19–22.
- Barnard, S. T., and W. B. Thompson. 1980. “Disparity Analysis of Images.” *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-2* (4): 333–340. doi:10.1109/TPAMI.1980.4767032.
- Baumberg, A. 2000. “Reliable Feature Matching across Proceidings of IEEE Conference on Computer Vision and Pattern Recognition, 774–781, Hilton Head, USA, June 13–15.
- Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool. 2008. “Speeded-up Robust Features (SURF).” *Computer Vision and Image Understanding* 110 (3): 346–359. doi:10.1016/j.cviu.2007.09.014.

- Bay, H., T. Tuytelaars, and L. Van Gool. 2006. "SURF: Speeded up Robust Features." Paper presented at the *Proceedings of the European Conference on Computer Vision*, 404–417, Graz, Austria, May 7–13.
- Beis, J. S., and D. G. Lowe. 1997. "Shape Indexing Using Approximate Nearest-neighbour Search in High-dimensional Spaces." Paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1000–1006, San Juan, Puerto Rico, June 17–19.
- Bromley, J., I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. 1994. "Signature Verification Using a "Siamese" Time Delay Neural Network." Paper presented at the *Proceedings of the Advances in Neural Information Processing Systems*, 737–744, Denver, Colorado, USA, November 28–December 1.
- Brown, D. C. 1958. "A Solution to the General Problem of Multiple Station Analytical Stereotriangulation." *RCA Data reduction technical report*. D. Brown Associates, Incorporated.
- Brown, M., G. Hua, and S. Winder. 2011. "Discriminative Learning of Local Image Descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (1): 43–57. doi:10.1109/TPAMI.2010.54.
- Brown, M., R. Szeliski, and S. Winder. 2005. "Multi-image Matching Using Multi-scale Oriented Patches." Paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. I, 510–517, San Diego, USA, June 20–26.
- Calonder, M., V. Lepetit, C. Strecha, and P. Fua. 2010. "BRIEF: Binary Robust Independent Elementary Features." Paper presented at the *Proceedings of the European Conference on Computer Vision*, 778–792, Heraklion, Crete, Greece, September 5–11.
- Calonder, M., V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua. 2012. "BRIEF: Computing a Local Binary Descriptor Very Fast." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (7): 1281–1298. doi:10.1109/TPAMI.2011.222.
- Carlevaris-Bianco, N., and R. M. Eustice. 2014. "Learning Visual Feature Descriptors for Dynamic Lighting Conditions." Paper presented at the *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2769–2776, Chicago, USA, September 14–18.
- Chechik, G., V. Sharma, U. Shalit, and S. Bengio. 2010. "Large Scale Online Learning of Image Similarity through Ranking." *Journal of Machine Learning Research* 11 (3): 1109–1135.
- Chen, L., F. Rottensteiner, and C. Heipke. 2020. "Deep Learning Based Feature Matching and Its Application in Image Orientation." In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences V-2*: 25–33. doi:10.5194/isprs-annals-V-2-2020-25-2020.
- Chen, L., F. Rottensteiner, and C. Heipke. 2014. "Learning Image Descriptors for Matching Based on Haar Features." In *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 61–68. Vol. XL-3. doi: 10.5194/isprsarchives-XL-3-61-2014
- Chen, L., F. Rottensteiner, and C. Heipke. 2016. "Invariant Descriptor Learning Using a Siamese Convolutional Neural Network." In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 11–18. Vol. III-3. doi:10.5194/isprsannals-III-3-11-2016
- Chen, M., and Z. Shao. 2013. "Robust Affine-invariant Line Matching for High Resolution Remote Sensing Images." *Photogrammetric Engineering & Remote Sensing* 79 (8): 753–760. doi:10.14358/PERS.79.8.753.
- Csurka, G., C. R. Dance, and M. Humenberger. 2018. "From Handcrafted to Deep Local Features." *arXiv Preprint arXiv:1807.10254*, 1–41.
- Dreschler, L., and -H.-H. Nagel. 1981. "On the Frame-to-frame Correspondence between Greyvalue Characteristics in the Images of Moving Objects." Paper presented at the *Proceedings of the German Workshop on Artificial Intelligence*, 18–29, Bad Honnef, Germany, January 26–31.
- Fan, B., Q. Kong, X. Wang, Z. Wang, S. Xiang, C. Pan, and P. Fua. 2019. "A Performance Evaluation of Local Features for Image-based 3D Reconstruction." *IEEE Transactions on Image Processing* 28 (10): 4774–4789. doi:10.1109/TIP.2019.2909640.
- Förstner, W., and E. Gülch. 1987. "A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features." In *Proceedings of ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, 281–305, Interlaken, Switzerland, June 2–4.
- Föstner, W. 1991. "Statistische Verfahren für die automatische Bildanalyse und ihre Bewertung bei der Objekterkennung und -vermessung." Habilitation Thesis, Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften, Nr. 370.
- Hadsell, R., S. Chopra, and Y. LeCun. 2006. "Dimensionality Reduction by Learning an Invariant Mapping." Paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1735–1742, New York, USA, June 17–22.
- Han, X., T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. 2015. "Matchnet: Unifying Feature and Metric Learning for Patch-based Matching." Paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3286, Boston, USA, June 27–30.
- Harris, C., and M. Stephens. 1988. "A Combined Corner and Edge Detector." Paper presented at the *Proceedings of Alvey Vision Conference*, vol. 15, 147–151, Manchester, UK, September.
- Heipke, C. 1997. "Automation of Interior, Relative, and Absolute Orientation." *ISPRS Journal of Photogrammetry and Remote Sensing* 52 (1): 1–19. doi:10.1016/S0924-2716(96)00029-9.
- Heipke, C., and F. Rottensteiner. 2020. "Deep Learning for Geometric and Semantic Tasks in Photogrammetry and Remote Sensing." *Geo-spatial Information Science* 23 (1): 10–19. doi:10.1080/10095020.2020.1718003.
- Hoffer, E., and N. Ailon. 2015. "Deep Metric Learning Using Triplet Network." Paper presented at the *Proceedings of the International Workshop on Similarity-Based Pattern Recognition*, 84–92, Copenhagen, Denmark, October 12–14.
- Jacobsen, K. 1980. "Vorschläge zur Konzeption und zur Bearbeitung von Bündelblockausgleichungen." PhD diss., Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover No. 102.
- Jacobsen, K., and M. Gerke. 2016. "Sub-camera Calibration of a Penta-camera." In *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35–40. Vol. XL-3/W4. doi:10.5194/isprsarchives-XL-3-W4-35-2016

- Jahrer, M., M. Grabner, and H. Bischof. 2008. "Learned Local Descriptors for Recognition and Matching." *Paper presented at the Proceedings of the Computer Vision Winter Workshop*, 39–46, Moravske Toplice, Slovenia, February 4–6.
- Jin, Y., D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls. 2020. "Image Matching across Wide Baselines: From Paper to Practice." *arXiv preprint arXiv:2003.01587*.
- Kager, H. 1980. "Das interaktive Programmsystem ORIENT im Einsatz." In *International Archives of Photogrammetry*, 390–401. Vols. XXIII–B5. Published by the Committee of the XIV International Congress for Photogrammetry and Remote Sensing Hamburg 1980, Hamburg, Germany, F. Ackermann, H. Bauer, G. Konecny, G. Kupfer (Editors).
- Ke, Y., R. Sukthankar, and I. C. Society. 2004. "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 506–513, Washington, USA, June 27–July 2.
- Keller, M., Z. Chen, F. Maffra, P. Schmuck, and M. Chli. 2018. "Learning Deep Descriptors with Scale-aware Triplet Networks." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2762–2770, Salt Lake City, USA, June 18–22.
- Kim, D.-G., W.-J. Nam, and S.-W. Lee. 2019. "A Robust Matching Network for Gradually Estimating Geometric Transformation on Remote Sensing Imagery." *Paper presented at the Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 3889–3894, Bari, Italy, October 6–9.
- Konecny, G. 1962. "Aerotriangulation mit Konvergentaufnahmen." PhD diss., Deutsche Geodätische Kommission Reihe C, No. 47, München.
- Kruck, E. 1983. "Lösung großer Gleichungssysteme für photogrammetrische Blockausgleichungen mit erweitertem funktionalen Modell." PhD diss., Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover No. 128.
- Kumar, B., G. Carneiro, and I. Reid. 2016. "Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimising Global Loss Functions." *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5385–5394, Las Vegas, USA, July 21–26.
- Lenc, K., and A. Vedaldi. 2016. "Learning Covariant Feature Detectors." *Paper presented at the Proceedings of the European Conference on Computer Vision Workshops*, 100–117, Amsterdam, The Netherlands, October 11–14.
- Lepetit, V., and P. Fua. 2006. "Keypoint Recognition Using Randomized Trees." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (9): 1465–1479. doi:10.1109/TPAMI.2006.188.
- Leutenegger, S., M. Chli, and R. Y. Siegwart. 2011. "BRISK: Binary Robust Invariant Scalable Keypoints." *Paper presented at the Proceedings of the IEEE International Conference on Computer Vision*, 2548–2555, Barcelona, Spain, November 6–13.
- Lindeberg, T. 1998. "Feature Detection with Automatic Scale Selection." *International Journal of Computer Vision* 30 (2): 79–116. doi:10.1023/A:1008045108935.
- Lindeberg, T. 1998. "Scale-space Theory: A Basic Tool for Analyzing Structures at Different Scales." *Journal of Applied Statistics* 21 (1–2): 225–270. doi:10.1080/757582976.
- Lindeberg, T., and J. Garding. 1997. "Shape-adapted Smoothing in Estimation of 3-D Shape Cues from Affine Deformations of Local 2-D Brightness Structure." *Image and Vision Computing* 15 (6): 415–434. doi:10.1016/S0262-8856(97)01144-X.
- Lowe, D. G. 1999. "Object Recognition from Local Scale-invariant Features." *Paper presented at the Proceedings of the International Conference on Computer Vision*, 1150–1157, Kerkyra, Corfu, Greece, September 20–25.
- Lowe, D. G. 2004. "Distinctive Image Features from Scale-invariant Keypoints." *International Journal of Computer Vision* 60 (2): 91–110. doi:10.1023/B:VISI.0000029664.99615.94.
- Lucas, B. D., and T. Kanade. 1981. "An Iterative Image Registration Technique with an Application to Stereo Vision." *Paper presented at the Proceedings of the DARPA Image Understanding Workshop*, 121–130, Vancouver, Canada, August 24–28.
- Luo, Z., T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. 2018. "Geodesc: Learning Local Descriptors by Integrating Geometry Constraints." *Paper presented at the Proceedings of the European Conference on Computer Vision*, 168–183, Munich, Germany, September 8–14.
- Ma, J., X. Jiang, A. Fan, J. Jiang, and J. Yan. 2020. "Image Matching from Handcrafted to Deep Features: A Survey." *International Journal of Computer Vision*. Advance online publication. doi:10.1007/s11263-020-01359-2.
- Matas, J., O. Chum, M. Urban, and T. Pajdla. 2004. "Robust Wide-baseline Stereo from Maximally Stable Extremal Regions." *Image and Vision Computing* 22 (10): 761–767. doi:10.1016/j.imavis.2004.02.006.
- Mikolajczyk, K. 2002. "Interest Point Detection Invariant to Affine Transformations." PhD diss., Institut National Polytechnique de Grenoble.
- Mikolajczyk, K., and C. Schmid. 2004. "Scale & Affine Invariant Interest Point Detectors." *International Journal of Computer Vision* 60 (1): 63–86. doi:10.1023/B:VISI.0000027790.02288.f2.
- Mikolajczyk, K., and C. Schmid. 2005. "A Performance Evaluation of Local Descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (10): 1615–1630. doi:10.1109/TPAMI.2005.188.
- Mikolajczyk, K., T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. 2005. "A Comparison of Affine Region Detectors." *International Journal of Computer Vision* 65 (1–2): 43–72. doi:10.1007/s11263-005-3848-x.
- Mishchuk, A., D. Mishkin, F. Radenovic, and J. Matas. 2017. "Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss." *Paper presented at the Proceedings of the Advances in Neural Information Processing Systems*, 4826–4837, Long Beach, USA, December 4–9.
- Mishkin, D., F. Radenovic, and J. Matas. 2018. "Repeatability Is Not Enough: Learning Affine Regions via Discriminability." *Paper presented at the Proceedings of the European Conference on Computer Vision*, 284–300, Munich, Germany, September 8–14.
- Moravec, H. P. 1979. "Visual Mapping by a Robot Rover." *Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 1, 598–600, Tokyo, Japan, August 20–23.
- Morel, J.-M., and G. Yu. 2009. "ASIFT: A New Framework for Fully Affine Invariant Image Comparison." *SIAM*



- Journal on Imaging Sciences* 2 (2): 438–469. doi:10.1137/080732730.
- Nex, F., F. Remondino, M. Gerke, H.-J. Przybilla, M. Bäumker, and A. Zurhorst. 2015. “ISPRS Benchmark for Multi-platform Photogrammetry.” In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 135–142. Vol. II-3/W4. doi:10.5194/isprsannals-II-3-W4-135-2015
- Onyango, F., F. Nex, M. Peter, and P. Jende. 2017. “Accurate Estimation of Orientation Parameters of Uav Images through Image Registration with Aerial Oblique Imagery.” In *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 599–605. Vol. XLII-1/W1. doi:10.5194/isprs-archives-XLII-1-W1-599-2017
- Osendorfer, C., J. Bayer, S. Urban, and P. Van Der Smagt. 2013. “Convolutional Neural Networks Learn Compact Local Image Descriptors.” *Paper presented at the Proceedings of the International Conference on Neural Information Processing*, 624–630, Daegu, Korea, November 3–7.
- Ozuysal, M., M. Calonder, V. Lepetit, and P. Fua. 2010. “Fast Keypoint Recognition Using Random Ferns.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (3): 448–461. doi:10.1109/TPAMI.2009.23.
- Rodehorst, V., and A. Koschan. 2006. “Comparison and Evaluation of Feature Point Detectors.” *Paper presented at the 5th International Symposium Turkish-German Joint Geodetic Days*, Berlin, Germany: Technical University of Berlin .
- Rosten, E., R. Porter, and T. Drummond. 2010. “Faster and Better: A Machine Learning Approach to Corner Detection.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (1): 105–119. doi:10.1109/TPAMI.2008.275.
- Rottensteiner, F. 2000. “Semi-automatic Building Reconstruction Integrated in Strict Bundle Block Adjustment.” *The International Archives of Photogrammetry and Remote Sensing XXXIII-B2*, 461–468, Amsterdam, The Netherlands, July 16–23.
- Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. 2011. “ORB: An Efficient Alternative to SIFT or SURF.” *Paper presented at the Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571, Barcelona, Spain, November 6–13.
- Savinov, N., A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. 2017. “Quad-networks: Unsupervised Learning to Rank for Interest Point Detection.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3929–3937, Honolulu, USA, July 21–26.
- Schmid, H. 1958/1959. “Eine allgemeine analytische Lösung für die Aufgabe der Photogrammetrie.” *Bildmessung und Luftbildwesen* 4 (1958): 103–113. (1/1959): 1–12, (2/1959): 71.
- Schönberger, J. L., and J.-M. Frahm. 2016. “Structure-from-motion Revisited.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113, Las Vegas, USA, June 27–30.
- Shao, Z., C. Li, D. Li, O. Altan, L. Zhang, and L. Ding. 2020. “An Accurate Matching Method for Projecting Vector Data into Surveillance Video to Monitor and Protect Cultivated Land.” *ISPRS International Journal of Geo-Information* 9 (7): 448. doi:10.3390/ijgi9070448.
- Simonyan, K., A. Vedaldi, and A. Zisserman. 2014. “Learning Local Feature Descriptors Using Convex Optimisation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (8): 1573–1585. doi:10.1109/TPAMI.2014.2301163.
- Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. 2015. “Discriminative Learning of Deep Convolutional Feature Point Descriptors.” *Paper presented at the Proceedings of the IEEE International Conference on Computer Vision*, 118–126, Santiago, Chile, December 7–13.
- Smith, M., N. Kokkas, A. Hamruni, D. Critchley, and A. Jamieson. 2008. “Investigation into the Orientation of Oblique and Vertical Digital Images.” *Paper presented at the European Calibration and Orientation Workshop (EUROCOW)*. Barcelona, Spain.
- Smith, S. M., and J. M. Brady. 1997. “SUSAN - A New Approach to Low Level Image Processing.” *International Journal of Computer Vision* 23 (1): 45–78. doi:10.1023/A:1007963824710.
- Snavely, N., S. M. Seitz, and R. Szeliski. 2006. “Photo Tourism: Exploring Photo Collections in 3D.” *Paper presented at the Proceedings of the International Conference on Computer Graphics and Interactive Techniques (ACM Siggraph)*, 835–846, New York, USA, July 30–August 3.
- Snavely, N., S. M. Seitz, and R. Szeliski. 2008. “Modeling the World from Internet Photo Collections.” *International Journal of Computer Vision* 80 (2): 189–210. doi:10.1007/s11263-007-0107-3.
- Szeliski, R. 2010. *Computer Vision: Algorithms and Applications*. London, UK: Springer Science and Business Media.
- Tang, L., and C. Heipke. 1993. “Approach for Automatic Relative Orientation.” In *Optical 3D Measurement Techniques II: Applications in Inspection, Quality Control, and Robotics*, 347–354, Zurich, Switzerland, October 4–7.
- Tang, L., and C. Heipke. 1996. “Automatic Relative Orientation of Aerial Images.” *Photogrammetric Engineering and Remote Sensing* 62 (1): 47–55.
- Tian, Y., B. Fan, and F. Wu. 2017. “L2-net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 661–669, Honolulu, USA, July 21–26.
- Tola, E., V. Lepetit, and P. Fua. 2009. “Daisy: An Efficient Dense Descriptor Applied to Wide-baseline Stereo.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (5): 815–830. doi:10.1109/TPAMI.2009.77.
- Trzcinski, T., M. Christoudias, P. Fua, and V. Lepetit. 2013. “Boosting Binary Keypoint Descriptors.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2874–2881, Portland, USA, June 23–28.
- Trzcinski, T., M. Christoudias, and V. Lepetit. 2015. “Learning Image Descriptors with Boosting.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (3): 597–610. doi:10.1109/TPAMI.2014.2343961.
- Tuytelaars, T., and L. Van Gool. 2004. “Matching Widely Separated Views Based on Affine Invariant Regions.” *International Journal of Computer Vision* 59 (1): 61–85. doi:10.1023/B:VISI.0000020671.28016.e8.
- Tuytelaars, T., L. Van Gool, L. D’haene, and R. Koch. 1999. “Matching of Affinely Invariant Regions for Visual Servoing.” *Paper presented at the Proceedings 1999 IEEE International Conference on Robotics and Automation*, 1601–1606, Detroit, USA, May 10–15.
- Tuytelaars, T., and L. J. Van Gool. 2000. “Wide Baseline Stereo Matching Based on Local, Affinely Invariant

- Regions.” *Paper presented at the Proceedings of the British Machine Vision Conference*, 38.1–38.14, Bristol, UK, September 11–14.
- Verdie, Y., K. Yi, P. Fua, and V. Lepetit. 2015. “TILDE: A Temporally Invariant Learned Detector.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5279–5288, Boston, USA, June 7–12.
- Verykokou, S., and C. Ioannidis. 2016. “Automatic Rough Georeferencing of Multiview Oblique and Vertical Aerial Image Datasets of Urban Scenes.” *The Photogrammetric Record* 31 (155): 281–303. doi:10.1111/phor.12156.
- Verykokou, S., and C. Ioannidis. 2018. “Oblique Aerial Images: A Review Focusing on Georeferencing Procedures.” *International Journal of Remote Sensing* 39 (11): 3452–3496. doi:10.1080/01431161.2018.1444294.
- Viola, P., and M. J. Jones. 2004. “Robust Real-time Face Detection.” *International Journal of Computer Vision* 57 (2): 137–154. doi:10.1023/B:VISI.0000013087.49260.fb.
- Wang, C., J. Chen, J. Chen, A. Yue, D. He, Q. Huang, and Y. Zhang. 2018. “Unmanned Aerial Vehicle Oblique Image Registration Using an ASIFT-based Matching Method.” *Journal of Applied Remote Sensing* 12 (2): 025002. doi:10.1117/1.JRS.12.025002.
- Wang, Y. 1994. “Strukturzuordnung zur automatischen Oberflächenrekonstruktion.” Ph. D. thesis, Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover No. 207.
- Winder, S. A., and M. Brown. 2007. “Learning Local Image Descriptors.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8, Minneapolis, Minnesota, USA, June 18–23.
- Wu, C. 2013. “Towards Linear-time Incremental Structure from Motion.” *Paper presented at the Proceedings of the International Conference on 3D Vision*, 127–134, Seattle, USA, June 29–July 1.
- Wu, C., S. Agarwal, B. Curless, and S. M. Seitz. 2011. “Multicore Bundle Adjustment.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3057–3064, Colorado Springs, CO, USA, June 20–25.
- Yi, K. M., E. Trulls, V. Lepetit, and P. Fua. 2016b. “LIFT: Learned Invariant Feature Transform.” In *Proceedings of the European Conference on Computer Vision*, 467–483, Amsterdam, The Netherlands, October 11–14.
- Yi, K. M., Y. Verdie, P. Fua, and V. Lepetit. 2016a. “Learning to Assign Orientations to Feature Points.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 107–116, Las Vegas, NV, USA, June 27–30.
- Zagoruyko, S., and N. Komodakis. 2015. “Learning to Compare Image Patches via Convolutional Neural Networks.” *Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4353–4361, Boston, MA, USA, June 7–12.