International Journal of Computer Vision

https://doi.org/10.1007/s11263-020-01359-2



**Image Matching from Handcrafted to Deep Features: A Survey**

**Jiayi Ma1** · **Xingyu Jiang1** · **Aoxiang Fan1** · **Junjun Jiang2** · **Junchi Yan3**



Received: 9 January 2020 / Accepted: 15 July 2020

© The Author(s) 2020

**Abstract**

As a fundamental and critical task in various visual applications, image matching can identify then correspond the same or similar structure/content from two or more images. Over the past decades, growing amount and diversity of methods have been proposed for image matching, particularly with the development of deep learning techniques over the recent years. However, it may leave several open questions about which method would be a suitable choice for specific applications with respect to different scenarios and task requirements and how to design better image matching methods with superior performance in accuracy, robustness and efficiency. This encourages us to conduct a comprehensive and systematic review and analysis for those classical and latest techniques. Following the feature-based image matching pipeline, we first introduce feature detection, description, and matching techniques from handcrafted methods to trainable ones and provide an analysis of the development of these methods in theory and practice. Secondly, we briefly introduce several typical image matching-based applications for a comprehensive understanding of the significance of image matching. In addition, we also provide a comprehensive and objective comparison of these classical and latest techniques through extensive experiments on representative datasets. Finally, we conclude with the current status of image matching technologies and deliver insightful discussions and prospects for future works. This survey can serve as a reference for (but not limited to) researchers and engineers in image matching and related fields.

**Keywords** Image matching·Graph matching·Feature matching·Registration·Handcrafted features·Deep learning

Communicated by V. Lepetit.

B Junchi Yan

yanjunchi@sjtu.edu.cn

Jiayi Ma

jyma2010@gmail.com

Xingyu Jiang

jiangx.y@whu.edu.cn

Aoxiang Fan

fanaoxiang@whu.edu.cn

Junjun Jiang

junjun0595@163.com

1. Electronic Information School, Wuhan University, Wuhan 430072, China
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
3. Department of Computer Science and Engineering, and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China



**1 Introduction**

Vision-based artificial systems, as widely used to guide machines to perceive and understand the surroundings for better decision making, have been playing a significant role in the age of global automation and artificial intelligence. However, how to process the perceived information under specific requirements and understand the differences and/or relationships among multiple visual targets are crucial topics in various fields, including computer vision, pattern recog-nition, image analysis, security, and remote sensing. As a critical and fundamental problem in these complicated tasks, *image matching, also known as image registration or cor-respondence, aims to identify then correspond the same or similar structure/content from two or more images.* This tech-nique is used for high-dimensional structure recovery as well as information identification and integration, such as 3-D reconstruction, visual simultaneous localization and map-ping (VSLAM), image mosaic, image fusion, image retrieval, target recognition and tracking, as well as change detection, etc.

123

International Journal of Computer Vision

Image matching has rich meaning in pairing two objects, thus deriving many specific tasks, such as sparse feature matching, dense matching (like image registration and stereo matching), patch matching (retrieval), 2-D and 3-D point set registration, and graph matching. Image matching in general consists of two parts, namely, the nature of the matched fea-tures and the matching strategy, which indicate what are used to match and how to match them, respectively. The ultimate goals are to geometrically warp the sensed image into the common spatial coordinate system of the reference image and align their common area pixel-to-pixel (i.e., image registra-tion). To this end, a direct strategy, also known as *area-based* *method*, registers two images by using the similarity mea-surement of the original image pixel intensity or information after pixel-domain transformation in the sliding windows of predefined size or even the entire images, without attempting to detect any salient image structure.

Another classic and widely adopted pipeline called *feature-based method*, i.e., feature detection and description, featurematching, transform model estimation, image resampling and transformation, has been introduced in the prestigious survey paper (Zitova and Flusser [2003](#page57)) and applied in var-ious fields. The feature-based image matching is popular due to its flexibility and robustness and the capability of wide range applications. In particular, feature detection can extract the distinctive structure from an image, and fea-ture description may be regarded as an image representation method that is widely used in image coding and similarity measurements such as image classification and retrieval. In addition, due to the strong ability in deep feature acquisition and non-linear expression, applying deep learning techniques for image information representation and/or similarity mea-surement, as well as parameter regression of image pair transformation, are hot topics in nowadays image matching community, which have been proven to achieve better match-ing performance and present greater potential compared with traditional methods.

In real-world settings, images for matching are usually taken from the same or similar scene/object while captured at different times, from different viewpoints or imaging modalities. In particular, a robust and efficient matching strat-egy is desirable to establish correct correspondences, thus stimulating various methods for achieving better efficiency, robustness and accuracy. Although numerous techniques have been devised over the decades, developing a unified framework remains a challenging task in terms of the fol-lowing aspects:

– Area-based methods that directly match images often depend on an appropriate patch similarity measurement for creating pixel level matches between images. They can be computational expensive and are sensitive to image distortion, appearance changes by noise, vary-

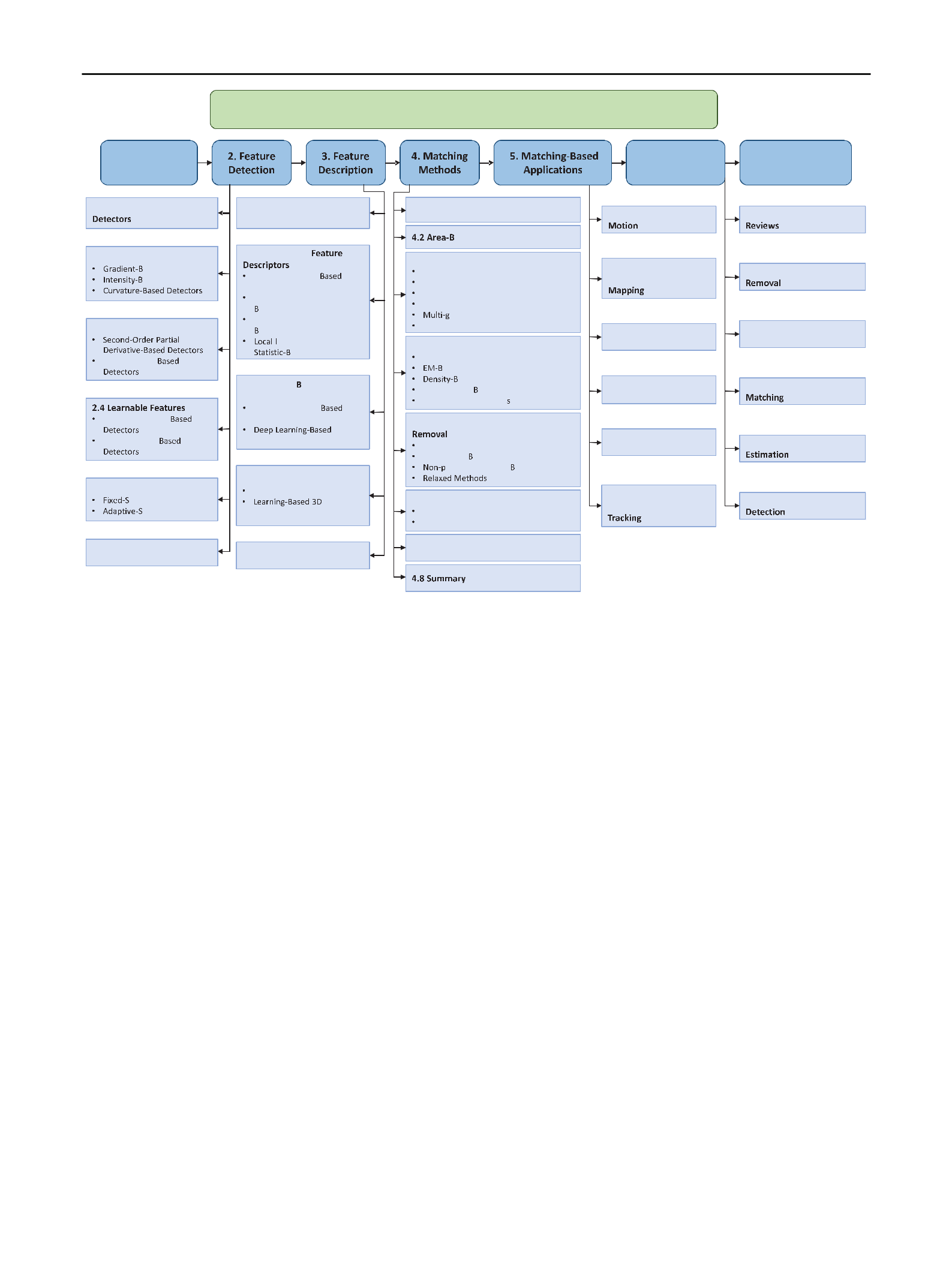
ing illumination, and different imaging sensors, which can have negative impact on similarity measurement and match searching. As a result, usually these methods can only work well under small rotation, scaling, and local deformation.

– Feature-based matching methods are often more efficient and can better handle geometrical deformation. But they are based on salient feature detection and description, fea-ture matching, and geometrical model estimation which can also be challenging. On the one hand, in feature-based image matching, it is difficult to define and extract a high percentage and a large number of features belong-ing to the same positions in 3-D space in the real world to ensure the matchability. On the other hand, matching *N* feature points to *N* feature points detected in anotherimage would create a total of *N* ! possible matchings, and thousands of features are usually extracted from high-resolution images and dominated outliers and noise are typically included in the points sets, which lead to signifi-cant difficulties for existing matching methods. Although various local descriptors have been proposed and cou-pled with detected features to ease the matching process, the use of local appearance information will unavoid-ably result in ambiguity and numerous false matches, especially for images with low quality, repeated contents, and those undergoing serious nonrigid deformations and extreme viewpoint changes.

– A predefined transformation model is often required to indicate the geometrical relation between two images or point sets. But it may vary on different data and is unknown beforehand thus hard to model. A sim-ple parametric model is often insufficient for image pairs that involve non-rigid transformations caused by ground surface fluctuation and image viewpoint varia-tions, multi-targets with different motion properties, and also local distortions.

– The emergence of deep learning has provided a new way and has shown great potential to address image match-ing problems. However, it still faces several challenges. The option of learning from images for direct registra-tion or transformation model estimation is limited when applied to wide baseline image stereo or registration under complex and serious deformation. The application of convolutional neural networks (CNNs) onto sparse point data for matching, registration, and transforma-tion model estimation is also difficult, because the points to be matched–known as unstructured or non-Euclidean data due to their disordered and dispersed nature–make it difficult to operate and extract the spatial relation-ships between two or more points (e.g., neighboring elements, relative positions, and length and angle infor-mation among multi-points) using a deep convolutional technique.

123

International Journal of Computer Vision

**Fig. 1** Structure of this survey

Existing surveys are focused on different parts of image matching tasks and fail to cover the literature from the last decade. For instance, the early reviews (Zitova and Flusser [2003](#page57); Tuytelaars and Mikolajczyk [2008](#page54); Strecha et al. [2008](#page53); Aanæs et al. [2012](#page43); Heinly et al. [2012](#page48); Awrangjeb et al. [2012](#page44); Li et al. [2015](#page50)) typically focus on handcrafted meth-ods, which are not sufficient to provide a valuable reference for investigating CNN-based methods. Most recent reviews involve trainable techniques, but they merely cover a single part of image matching community, either focus on detectors (Huang et al. [2018](#page48); Lenc and Vedaldi [2014](#page49)) or descriptors (Balntas et al. [2017](#page44); Schonberger et al. [2017](#page53)) or specific matching tasks (Ferrante and Paragios [2017](#page47); Haskins et al. [2020](#page48); Yan et al. [2016b](#page55); Maiseli et al. [2017](#page50)), and many others pay more attention on related applications (Fan et al. [2019](#page46); Guo et al. [2016](#page47); Zheng et al. [2018](#page56); Piasco et al. [2018](#page52)). In this survey, we aim to provide an up-to-date and comprehensive summary and assessment of existing image matching meth-ods, especially for the recently introduced learning-based methods. More importantly, we have provided a detailed evaluation and analysis for mainstream methods which are missing in existing literature.

This survey mainly focuses on feature-based matching, although patch matching, point set registration, and other

related matching tasks are also reviewed. The overall orga-nization is presented in Fig. [1](#page3); Sects. [2](#page3) and [3](#page9) describe the feature detection and description techniques respectively, from handcrafted methods to trainable ones. Patch match-ing is classified as a feature description domain, and 3-D point set features are also reviewed. In Sect. [4](#page16), we present different matching methods, including area-based image matching, pure point set registration, image descriptor sim-ilarity matching and mismatch removal, graph matching, and learning-based methods. Sections [5](#page28) and [6](#page30) respectively introduce the image matching-based visual applications and evaluation metrics, including the performance comparison. In Sect. [7](#page42), we conclude and discuss possible future develop-ments.

**2 Feature Detection**

Early image features are annotated manually, which are still used in some low-quality image matching. With the development of computer vision and the requirement for auto-matching approaches, many feature detection methods have been introduced to extract stable and distinct features from images.

123

International Journal of Computer Vision

**2.1 Overview of Feature Detectors**

Detected features represent specific semantic structures in an image or the real world and can be divided into corner feature (Moravec [1977](#page51); Harris et al. [1988](#page47); Smith and Brady [1997](#page53); Rosten and Drummond [2006](#page53); Rublee et al. [2011](#page53)), blob feature (Lowe [2004](#page50); Bay et al. [2006](#page44); Agrawal et al. [2008](#page44); Yi et al. [2016](#page56)), line/edge (Harris et al. [1988](#page47); Smith and Brady [1997](#page53); Canny [1987](#page45); Perona and Malik [1990](#page52)), and morpho-logical region feature (Matas et al. [2004](#page51); Mikolajczyk et al. [2005](#page51)). However, the most popular features that are used for matching are the points (a.k.a. keypoints or interest points). The points are easy to extract and define with a simplified form compared with the line and region features, which can be roughly classified into corner and blob.

A good interest point must be easy to find and ideally fast to compute, as an interest point at a good location is crucial for further feature description and matching. To promote (i) matchability, (ii) the capability for subsequent applications, and (iii) matching efficiency and reduction of storage requirements, many required properties have been proposed for reliable feature extraction (Zitova and Flusser [2003](#page57); Tuytelaars and Mikolajczyk [2008](#page54)), including repeata-bility, invariance, robustness and efficiency. The common idea for feature detection is to construct a feature response to distinguish salient point, line, and region from one another, along with flat and nondistinctive image areas. This idea can be subsequently classified into *gradient-, intensity-, second-order derivative-, contour curvature-, region segmentation-, and learning-based detectors*. In the following, we providea comprehensive introduction of feature detectors with these methods, focusing more on learning-based methods to guide researchers on how the traditional and trainable detectors work and give insights on their strengths and weaknesses.

**2.2 Corner Features**

A corner feature can for example be defined as the crossing point of two straight lines with the forms of “L”, “T”, “X”, or a high curvature point of a contour. The common idea of cor-ner detection is to compute a corner response and distinguish it from edge, flat, or other less distinctive image areas. Differ-ent strategies can be utilized for traditional corner searching, namely, gradient-, intensity-, and contour curvature-based. Refer to Zitova and Flusser ([2003](#page57)), Li et al. ([2015](#page50)), Tuyte-laars and Mikolajczyk ([2008](#page54)) and Rosten et al. ([2010](#page53)) for details.

**2.2.1 Gradient-Based Detectors**

A gradient-based corner response prefers the use of the first-order information in image to distinguish the corner feature. The earliest automatic corner detection method could be

traced to Moravec detector (Moravec [1977](#page51)), which first intro-duced the concept of “interest points” to define the distinct feature points, which are extracted based on the autocor-relation of the local intensity. This method calculates and searches the minimum intensity variation of each pixel from a shifted window in eight directions, and the interest point is detected if the minimum is superior to the given threshold.

However, the Moravec detector is not invariant to the direction or image rotation due to the discontinuous compar-ing directions and sizes. The famous Harris corner detector (Harris et al. [1988](#page47)) was introduced to address the anisotropy and computation complexity problem. The goal of the Har-ris method is to find the directions of the fastest and lowest grey-value changes using a two-order moment matrix or an auto-correlation matrix; thus, it is invariant to orientation and illumination and has reliable repeatability and distinctive-ness. Harris was further improved in Shi and Tomasi ([1993](#page53)) for better tracking performance by making the features more “spread out” and locating more accurately.

**2.2.2 Intensity-Based Detectors**

Several template- or intensity comparison-based corner detectors have been proposed by comparing the intensity of the surrounding pixels with that of the center pixel to simplify the image gradient computing. Due to their binary nature, they are widely used in many modern applications, particularly some with storage and real-time requirements.

The intensity-based corner detector, namely, smallest uni-value segment assimilating nucleus (SUSAN) (Smith and Brady [1997](#page53)), is based on the brightness similarity between the local radius region pixels and the nucleus. SUSAN can be implemented rapidly because it does not require gradient computation. Many analogous methods have been proposed based on the concept of brightness comparison, the most famous of which is the FAST detector (Trajkovi´c and Hedley [1998](#page54)). FAST uses binary comparison with each pixel along a circle pattern against the central pixel and then determines more reliable corner features using a machine learning (i.e., ID3 tree Quinlan [1986](#page52)) strategy, which is trained on a large number of similar scene images and can generate the best criteria for corner selection.

As an improvement of SUSAN, FAST is extremely effi-cient with high repeatability and is used more widely. To improve FAST without loss of efficiency, FAST-ER (Ros-ten et al. [2010](#page53)) was introduced to enhance the repeatability by generalizing the detector based on further pixel intensity comparison centered on the nucleus. Another improvement is the AGAST (Mair et al. [2010](#page50)), in which two more pixel brightness comparison criteria are defined, after which an optimal and specialized decision tree is trained in an extended configuration space, thus rendering the FAST detector more generic and adaptive. To combine the efficiency of FAST

123

International Journal of Computer Vision

and the reliability of the Harris detector, Rublee et al. ([2011](#page53)) proposed an integrated feature detector and descriptor for matching called ORB. The ORB uses the Harris response to select a certain number of FAST corners as the final detected features. The gray-scale centroid of the local patch and the center pixel itself are formed as a vector to represent the main direction of the ORB feature, which helps calculate the similarity of the binary descriptor in ORB. Recently, a Sadder-like detector (Aldana-Iuit et al. [2016](#page44)) has been pro-posed to extract interest points. In this detector, the saddle condition is verified efficiently by intensity comparisons on two concentric rings with certain geometric constraints. The Sadder detector can achieve higher repeatability and greater spread out than traditional methods even modern trainable ones (Komorowski et al. [2018](#page48)).

**2.2.3 Curvature-Based Detectors**

Another strategy for corner feature extraction is based on detected high-level image structures, such as edges, con-tours, and salient regions. Corner features can be defined immediately as the midpoint/endpoint or sparse sampling from an edge or contour (Belongie et al. [2002](#page44)). These are subsequently used for shape matching or point registration, especially for an image pair of less texture or binary type. The curvature-based strategy aims to extract the corner point with the maximum curvature searching based on the detected image curve-like edges. This strategy starts with an edge extraction and selection method, and the two subsequent steps are the curve smoothing and curvature estimation. The corners are finally determined by selecting the curvature extremum points. In general, an edge detector is often first in need for contour curvature-based corner detection.

In curve smoothing, the slope and curvature are difficult to evaluate due to the quantized position of a curve point. Noise and local deformation in a curve may also lead to a serious impact on the feature stability and distinctiveness. There-fore, smoothing methods should be implemented before or during the curvature calculation to make the curvature extremum points more distinct from other curve points. Two smoothing strategies, namely, direct and indirect methods, are generally utilized. A direct smoothing, such as Gaus-sian smoothing (Mokhtarian and Suomela [1998](#page51); Pinheiro and Ghanbari [2010](#page52)), removes noise and may change curve locations to a certain extent. In comparison, in the indirect smoothing strategy, e.g., the region of support method or the chord-length-based method (Ramer [1972](#page52); Awrangjeb and Lu [2008](#page44)), may preserve the curve point locations.

As for curvature estimation, for each point of the smoothed curve, a significance response measure is needed for corner searching, i.e., curvature. Curvature estimation methods are also generally classified as direct and indirect. The former is based on an algebraic or geometric estimation, such as

cosine, local curvature, and tangential deflection (Mokhtar-ian and Suomela [1998](#page51); Rosenfeld and Weszka [1975](#page52); Pinheiro and Ghanbari [2010](#page52)). The latter estimates the curvature in an indirect way and is often used as a significance measure, such as counting the number of curve points through several mov-ing rectangles along the curve (Masood and Sarfraz [2007](#page50)), using the perpendicular distances from the chord connecting the two endpoints of the curve to curve points (Ramer [1972](#page52)), and other alternatives (Zhang et al. [2010](#page56), [2015](#page56)). Compared with indirect estimation methods, the direct ones are more sensitive to noise and local variation due to the less neigh-boring point consideration.

Finally, corners can be determined with threshold strat-egy to remove false and indistinctive points (Mokhtarian and Suomela [1998](#page51); Awrangjeb and Lu [2008](#page44)). Additional details can be obtained from a contour curvature-based corner survey (Awrangjeb et al. [2012](#page44)). In addition and more recently, a mul-tiscale segmentation-based corner detector, named MSFD (Mustafa et al. [2018](#page51)), has been proposed for wide-baseline scene matching and reconstruction. Feature points in MSFD are detected at the intersection of the boundaries of three or more regions by using off-the-shelf segmentation meth-ods. MSFD can generate rich and accurate corner features for wide-baseline image matching and high reconstruction performance.

The above-mentioned corner feature detectors are easily located in the contour or edge structures of an image (i.e., not such spread-out or uneven distribution), and are limited by the scale and affine transformation between two images. Among the three types of corner detection strategies, the gradient-based methods are able to locate more accurately, whereas the intensity-based methods show advantage for efficiency. The contour curvature-based methods require more compu-tation but they are a better choice for processing textureless or binary images, such as infrared and medical images, because the image cue-based feature descriptors are unworkable for these types of images and the point-based descriptors are often coupled for the matching task (i.e., point set registra-tion or shape matching). Please refer to Sects. [3](#page9) and [4](#page16) for details.

**2.3 Blob Features**

A blob feature is commonly indicated as a local closed region (e.g., with a regular shape of circle or ellipse), inside which the pixels are considered similar to one another and are dis-tinct from the surrounding neighborhoods. The blob feature can be written in the form of *(x,* *y, θ )*, with *(x,* *y)* being the pixel coordinate of the feature location and *θ* indicat-ing the blob shape information of the feature, including scale and/or affine. Numerous blob feature detectors have been introduced over the past decades, and they can be roughly classified into second-order partial derivative- and

123

International Journal of Computer Vision

region segmentation-based detectors. Second-order partial derivative-based methods are based on the Laplacian scale selection and/or Hessian matrix calculation for affine invari-ant. While segmentation-based methods prefer to detect blob features by segmenting the morphological regions first, then estimate the affine information with ellipse fitting. Compared with corner features, blob features are more useful for visual applications with high precision requirement, because more image cues are utilized for feature identification and repre-sentation, thus enabling the blob features to be more accurate and robust to image transformation.

**2.3.1 Second-Order Partial Derivative-Based Detectors**

In methods based on *second-order partial derivatives*, the Laplacian of Gaussian (LoG) (Lindeberg [1998](#page49)) is applied based on scale space theory. Here, the Laplace operator is first used for edge detection in accordance with the zero crossings in the second-order differential of an image, and the Gaussian convolution filtering is then applied as a preprocessing to reduce noise.

LoG can detect the local extremum point and the area with normalized response arising from the circular symme-try of the Gaussian kernel. Different standard deviations of the Gaussian function can detect the scale-invariant blobs in different scales by searching the extremum in the multi-scale space as the final stable blob feature. The difference of Gaussians (DoG) (Lowe et al. [1999](#page50); Lowe [2004](#page50)) filter can be used to approximate the LoG filter, and greatly speeds up the computations. Another classical blob feature detec-tion strategy is based on the determinant of Hessian (DoH) (Mikolajczyk and Schmid [2001](#page51), [2004](#page51)). This is more affine invariant because the eigenvalue and eigenvector of the sec-ond matrix can be applied to estimate and correct the affine region.

Interest point detection by using DoG, DoH, and both has been widely utilized in recent visual applications. The famous SIFT (Lowe et al. [1999](#page50); Lowe [2004](#page50)) extracts key-point as the local extrema in a DoG pyramid, filtered using the Hessian matrix of the local intensity values (the according description part will be reviewed in the next section). Miko-lajczyk et al. combined the Harris and Hessian detectors with the Laplacian and Hessian matrices for scale and affine fea-ture detection (Mikolajczyk and Schmid [2001](#page51), [2004](#page51)), i.e., the Harris/Hessian-Laplacian/affine. SURF (Bay et al. [2006](#page44)) accelerates the SIFT by approximating the Hessian matrix-based detector using Haar wavelet calculation, together with an integral image strategy, thus simplifying the construction of a second-order differential template.

Several SIFT- and SURF-based improvements, have been successively proposed for better property in subsequent applications. Such improvements include a fully affine invari-ant SIFT detector (ASIFT) (Morel and Yu [2009](#page51)), a center-

surround extremum (Agrawal et al. [2008](#page44)) strategy feature detector with the Laplace calculation approximated by the proposed bilateral filtering to enhance the efficiency, and the efficient approximation of DoH with piecewise triangle filters in DARTs (Marimon et al. [2010](#page50)). In addition, a cosine-modulated Gaussian filter is utilized in the SIFT-ER detector (Mainali et al. [2013](#page50)) to obtain high feature detectability with minimum scale-space localization errors, in which the fil-terbank system has a highly accurate filter approximation without any image sub/upsampling. An edge foci-based blob detector (Zitnick and Ramnath [2011](#page57)) has also been intro-duced for the matching task. In this detector, the edge foci is defined as the point in an image that is roughly equidistant from the closest edge with orientations perpendicular to this point.

Unlike the circle-like Gaussian response function, a non-linear partial differential equation is applied in KAZA detector for blob feature searching with nonlinear diffu-sion filtering (Alcantarilla et al. [2012](#page44)). An accelerated version called AKAZA (Alcantarilla and Solutions [2011](#page44)) is implemented by embedding the fast explicit diffusion in a pyramidal framework to dramatically speedup feature detec-tion in nonlinear scale spaces. However, it still suffers from high computation complexity. Another method is WADE (Salti et al. [2013](#page53)), which implements nonlinear feature detec-tion by a wave propagation function.

**2.3.2 Segmentation-Based Detectors**

The segmentation-based blob detectors begin with an irreg-ular region segmentation based on constant pixel intensity or zero gradient. One of the most famous region segmentation-based blob feature is maximally stable extremal region (MSER) (Matas et al. [2004](#page51)). It extracts regions that remain stable under a large range of intensity thresholding values. This approach does not need extra processing for scale esti-mation, and is robust to large viewpoint changes. The term “maximally stable” describes the threshold selection process, given that every extremal region is a connected component of a watershed image by thresholding. An extension to MSER was introduced in Kimmel et al. ([2011](#page48)) to exploit shape struc-ture cues. Other improvements are based on the watershed regions of principal curvature images (Deng et al. [2007](#page46); Fer-raz and Binefa [2012](#page47)) or considered color information for a higher discrimination (Forssén [2007](#page47)).

Similar to MSER, other segmentation-based features, such as intensity- and edge-based regions (Tuytelaars and Van Gool [2004](#page54)), are also used for affine covariant region detection. However, feature detection of this type is of less use for feature matching, and it is gradually developed toward saliency detection and segmentation in computer vision. Spe-cific method investigation and comprehensive reviews can be found in Mikolajczyk et al. ([2005](#page51)) and Li et al. ([2015](#page50)).

123

International Journal of Computer Vision

**2.4 Learnable Features**

Over the recent years, data-driven learning-based methods have achieved significant progress in general visual pattern recognition tasks, and have also been applied to image feature detection. This pipeline can be roughly classified into the using of classical learning and deep learning.

**2.4.1 Classical Learning-Based Detectors**

Early from the past decade, classical learning-based meth-ods, such as decision tree, support vector machine (SVM), and other classifiers by opposition to Deep Learning, have already been used in handcrafted keypoint detection (Tra-jkovi´c and Hedley [1998](#page54); Strecha et al. [2009](#page53); Hartmann et al. [2014](#page48); Richardson and Olson [2013](#page52)). FAST (Trajkovi´c and Hedley [1998](#page54)) detector was the first attempt to use tradi-tional learning for reliable and matchable point identification, and similar strategies have been applied in many subsequent improvements (Mair et al. [2010](#page50); Rublee et al. [2011](#page53)). Strecha et al. ([2009](#page53)) trained the Wald-Boost classifier to learn key-points with high repeatability on pre-aligned training sets.

More recently, Hartmann et al. ([2014](#page48)) showed that it can be learnt from a structure-from-motion (SfM) pipeline to predict which candidate points are matchable, thus signifi-cantly reducing the number of interest points without losing excessive true matches. Meanwhile, Richardson and Olson ([2013](#page52)) reported that hand-designed detectors can be learned by random sampling in the space of convolutional filters and tried to find the optimal filter using a learning strategy over frequency-domain constraints. However, classical learning has only been used for reliable feature selection through clas-sifier learning, rather than the extraction of interest features directly from raw images until the emergence of deep learn-ing.

**2.4.2 Deep Learning-Based Detectors**

Inspired by the handcrafted feature detectors, a general solu-tion for CNN-based detection is to construct response maps to search the interest points in a supervised (Yi et al. [2016](#page56); Verdie et al. [2015](#page54); Zhang et al. [2017b](#page56)), self-supervised (Zhang and Rusinkiewicz [2018](#page56); DeTone et al. [2018](#page46)), or unsupervised manner (Lenc and Vedaldi [2016](#page49); Savinov et al. [2017](#page53); Ono et al. [2018](#page51); Georgakis et al. [2018](#page47); Barroso-Laguna et al. [2019](#page44)). The task is often converted into a regression problem that can be trained in a differentiable way under the transformation and imaging condition invariance con-straints. Supervised methods have shown the benefits of using anchors (e.g., obtained from SIFT method) to guide their training, but the performance could be largely restricted by the method of anchor construction, because the anchor itself is intrinsically difficult to reasonably define and may pre-

vent the network from proposing new keypoints in case no anchor exists in the proximity (Barroso-Laguna et al. [2019](#page44)). Self-supervised and unsupervised methods train detectors without any human annotations, and only the geometric con-straints between two images are required for optimization guidance; a simple human aid is sometimes asked for pre-training (DeTone et al. [2018](#page46)). In addition, many methods integrate feature detection into the entire matching pipeline by jointly training with feature description and matching (Yi et al. [2016](#page56); DeTone et al. [2018](#page46); Ono et al. [2018](#page51); Shen et al. [2019](#page53); Dusmanu et al. [2019](#page46); Choy et al. [2016](#page46); Rocco et al. [2018](#page52); Dusmanu et al. [2019](#page46); Revaud et al. [2019](#page52)), which can enhance the final matching performance and optimize the entire procedure in an end-to-end manner.

For instance, TILDE (Verdie et al. [2015](#page54)) trains multiple piecewise linear regression models to detect repeatable key-points under drastic imaging changes of weather and lighting conditions. First, it identifies good keypoint candidates in multiple training images taken from the same viewpoints using DoG for training set collection, and then trains a gen-eral regressor to predict a score map, whose maxima after non-maximum suppression (NMS) can then be regarded as the desired interest points.

DetNet (Lenc and Vedaldi [2016](#page49)) is the first fully general formulation for learning local covariant features; it casts the detection task as a regression problem and then derives a covariance constraint to automatically learn stable anchors for local feature detection under geometric transformations. Meanwhile, Quad-net (Savinov et al. [2017](#page53)) realizes keypoint detection under transformation-invariant quantile ranking with a single real-valued response function, enabling it to learn the detector completely from scratch by optimizing for a repeatable ranking. A similar detector in Zhang and Rusinkiewicz ([2018](#page56)) combines this “ranking” loss with a “peakedness” loss and produces a more repeatable detector.

Zhang et al. ([2017b](#page56)) proposed TCDET detector by defin-ing a novel formulation based on the new concepts of “standard patch” and “canonical feature” to place equal focus on discriminativeness and covariant constraint. The proposed detector can detect discriminative and repeatable features under diverse image transformations. Key.Net (Barroso-Laguna et al. [2019](#page44)) combines handcrafted and learned CNN filters within a shallow multiscale architecture and proposes a light/efficient trainable detector. The handcrafted filters pro-vide anchor structures for localizing, scoring, and ranking repeatable features that are fed to learned filters. CNN is used to represent the scale space by detecting keypoints at differ-ent levels; the loss function is defined to detect robust feature points from different scales and maximize the repeatability score. The affine region-based interest point is also learned using CNNs in Mishkin et al. ([2017](#page51), [2018](#page51)).

The methods of integrating a detector into a matching pipeline are similar to those solely designed for detection

123

International Journal of Computer Vision

reviewed above. The main difference may lie in the way of training, and the core challenge is to make the entire pro-cess differentiable. For example, Yi et al. ([2016](#page56)) attempted to train a detector, an orientation estimator, and a descriptor jointly based on inputting four patches. Their proposed LIFT can be regarded as a trainable version of SIFT and requires supervision from the SfM system for determining the fea-ture anchor. The training procedure is conducted individually from descriptor to detector and can use the learned results to guide the detector training, thus promoting detectability. Unlike LIFT, SuperPoint (DeTone et al. [2018](#page46)) introduces a fully convolutional model by inputting full-sized images and jointly computing pixel-level interest point locations and associated descriptors in one forward pass; a synthetic dataset is constructed for pseudo-ground truth generation and pre-training, and the homography adaption module enables it to achieve self-supervised training while promoting detection repeatability.

LF-Net (Ono et al. [2018](#page51)) confines the end-to-end pipeline to one branch to optimize the entire procedure in a dif-ferentiable way; it also uses a fully convolutional network operating on full-sized images to generate a rich feature score map, which can then be used to extract keypoint locations and the feature attributes, such as scale and ori-entation; simultaneously, it performs a differentiable form of NMS, namely, *so f t ar gmax*, for subpixel location and increasing the accuracy and saliency of keypoint. Similar to LF-Net, RF-Net (Shen et al. [2019](#page53)) selects high-response pixels as keypoints on multiscales, but the response maps are constructed by receptive feature maps. Bhowmik et al. ([2020](#page44)) indicated that increased accuracy for these low-level matching scores does not necessarily translate to better per-formance in high-level vision tasks, thus they embedded the feature detector in a complete vision pipeline, where the learnable parameters are trained in an end-to-end man-ner. The authors overcome the discrete nature of keypoint selection and descriptor matching using principles from rein-forcement learning. Luo et al. ([2020](#page50)) proposed ASFeat to explore local shape information of feature points and enhance the accuracy of points detection, by jointly learning local feature detectors and descriptors. Another detection-related learning-based method is to estimate the orientation (Moo Yi et al. [2016](#page51)), while the spatial transformation network (STN) (Jaderberg et al. [2015](#page48)) could also be a great reference in deep learning-based detectors for rotation invariance (Yi et al. [2016](#page56); Ono et al. [2018](#page51)).

Unlike local feature descriptors, there is little review on salient feature detectors, particularly for the recent CNN-based techniques. To our best knowledge, the most recent survey (Lenc and Vedaldi [2014](#page49)) focuses on local feature detection. It introduces the basic idea of several well-known methods from handcrafted detectors to accelerated and learned ones.

**2.5 3-D Feature Detectors**

Dedicated on 3-D keypoint detectors, Tombari et al. ([2013](#page54)) provided an excellent survey on the state-of-the-art meth-ods and a detailed evaluation of their performances. In brief, the existing methods were divided into two categories, *fixed-scale detectors* and *adaptive-scale detectors*. In bothcategories, keypoints are selected as local extrema of a pre-defined saliency measurement. The difference lies in the involvement of the scale characteristic, which defines the support for the subsequent description stage. The fixed-scale detectors tend to search keypoints at a specific scale level, which is given as prior information. The adaptive-scale detectors either extend the scale concept for 2-D images by adopting a scale space defined on the surface or implement the traditional scale-space analysis by embedding 3-D data onto a 2-D plane.

**2.5.1 Fixed-Scale Detectors**

Chen and Bhanu ([2007](#page45)) introduced the local surface patch (LSP) method. The saliency of a point in LSP is measured by its shape index (Dorai and Jain [1997](#page46)), as defined by the principal curvatures at the point. Zhong ([2009](#page56)) introduced the intrinsic shape signature (ISS) method, in which saliency is derived from the eigenvalue decomposition of the scat-ter matrix of the support region. In this approach, the ratio of eigenvalues is used to prune some points, and the final saliency is determined by the eigenvector. In this way, points with large variations along each principal direction are iden-tified. Analogous to ISS, Mian et al. ([2010](#page51)) also utilized the scatter matrix to prune nondistinctive points but with a differ-ent curvature-based saliency measurement. Sun et al. ([2009](#page54)) presented the heat kernel signature (HKS) method, based on the properties of the heat diffusion process on a shape. In this method, the saliency measurement is defined by the restriction of the heat kernel to the temporal domain. The heat kernel is uniquely determined by the underlying man-ifold, which makes HKS a compact characterization of the shape.

**2.5.2 Adaptive-Scale Detectors**

It is desirable to adaptively fit with the scale in detection. For this purpose, Unnikrishnan and Hebert ([2008](#page54)) proposed a Laplace-Beltrami scale space by computing the designed function on the increasing support around each point. This function is defined by a novel operator that reflects the local mean curvature of the underlying shape and provides the saliency information. Zaharescu et al. ([2009](#page56)) presented the MeshDoG method, which is analogous to the DoG operator in the 2-D case (Lowe [2004](#page50)); nonetheless, the operator is computed on a scalar function defined on the manifold. The

123

International Journal of Computer Vision

output of the DoG operator represents the saliency for key-points detection. Castellani et al. ([2008](#page45)) also built scale space using the DoG operator but directly on the 3-D mesh. Mian et al. ([2010](#page51)) proposed an automatic scale selection technique for extracting scale invariant features. The scale space is built by increasing the support size, and automatic scale selection at each keypoint is performed by using NMS along scale. The disadvantage of sensitivity to scale of HKS was addressed by Bronstein and Kokkinos ([2010](#page45)), who used Fourier transform magnitude to extract a scale-invariant quantity from the HKS without the need to perform scale selection. Sipiran and Bus-tos ([2011](#page53)) extended the well-known Harris operator ([1988](#page47)) into 3-D data with an adaptive-scale determination technique. Readers are referred to Tombari et al. ([2013](#page54)) for further dis-cussion on other adaptive-scale detectors. Salti et al. ([2015](#page53)) devised a learning-based 3-D keypoint detector, whereby the keypoint detection problem was cast as a binary classifica-tion problem, to determine whose support can be correctly matched by a predefined 3-D descriptor.

**2.6 Summary**

The basic idea of feature detectors is to distinguish the interest feature from others through the response value, thus leading to the solutions of two problems: (i) how to define discrimi-nant patterns in an image, and (ii) how to repeatedly detect the salient feature under different image conditions and image qualities (Zhang et al. [2017b](#page56)). Along with the development of these detectors, the main improvements and common strate-gies are related to four aspects, i.e., feature response type and improvements on efficiency, robustness, and accuracy, which lead to an increase in the matchability of detected features and the improved performance of their subsequent applica-tions.

For traditional methods, using more image cues can result in better robustness and repeatability, but usually requires more computational cost. In addition to using low-order fea-ture detectors, several strategies, such as approximate and pre-compute, are designed to largely speed up the computa-tion and maintain the matchability. To ensure the robustness, scale and affine information estimation is usually required when searching stable features. While for accuracy enhance-ment, a local extremal searching for subpixel accuracy and NMS strategy in pixel and scale space to avoid features locally gathered, are two popular choices in traditional pipelines.

As for learning-based detectors, repeatable and salient keypoints can be extracted based on high-level cues cap-tured by CNNs, except for intensity, gradient, or second-order derivative. While the efficiency would largely depend on the network structure, and early deep learning methods are often time-consuming. Methods proposed recently, such as SuperPoint and Key.Net, have already achieved good imple-

mentation in real time while maintaining state-of-the-art performance. Multiscale sampling or changed receptive field would make these deep learning-based detectors invariant to scale, where the scale or rotation information is directly estimated in networks. They can achieve promising results, because the deep learning techniques can easily distinguish the same structures, despite the fact that images suffer from apparent variance and geometrical transformation. The accu-racy can be optimized directly in the loss function of the learning-based methods, and the differentiable form of NMS is often used for subpixel accuracy location and repeatability enhancement.

**3 Feature Description**

Once discriminative interest points are detected from raw images, a local patch descriptor is required to be coupled for each feature in order to establish feature correspondence correctly and efficiently across two or more images. In other words, the feature descriptors are commonly used to trans-form the original local information around the interest point into a stable and discriminative form, usually as a high-dimensional vector, so that two corresponding features are as close as possible in the descriptor space, and two non-corresponding features are as far as possible.

**3.1 Overview of Feature Descriptors**

The processing procedure of feature description can be divided into three steps: local low-level feature extrac-tion, spatial pooling, and feature normalization (Lowe [2004](#page50); Rublee et al. [2011](#page53); Brown et al. [2010](#page45)). First, the low-level information of a local image region has to be extracted. This information consists of pixel intensity and gradient or is obtained from a series of steerable filters. Subsequently, the local patch is divided into several parts and the local informa-tion is pooled in each part, then concatenate them by using pooling methods, such as rectangular gridding (Lowe [2004](#page50)), polar gridding (Mikolajczyk and Schmid [2005](#page51)), Gaussian sampling (Tola et al. [2010](#page54)), and others (Rublee et al. [2011](#page53)); the joint feature representation is transformed into a more dis-criminative one that may preserve significant information in a simplified form for better matching performance. Finally, a descriptor is obtained from the normalized results of the pooled local information, which aims to map the aggregated results into a long vector of either floating-point or binary values for easily evaluating the similarity between image fea-tures.

Similar to feature detectors, existing descriptors are pro-posed and improved to become highly robust, efficient, and discriminant for addressing image matching problems. Esti-mating a good size and orientation for a cropped image

123

International Journal of Computer Vision

patch is core problems in the task of feature description and matching. By correctly identifying the size and orien-tation, the matching methods can be robust and invariant to global and/or local deformations, such as rotation and scal-ing. The original intention of feature description is focused on discrimination enhancement compared with direct simi-larity measurement using raw image information. Numerous well-designed descriptors can improve the discrimination and matching performance, by using pooling parameter optimization, sampling rule design, or the use of machine learning and deep learning techniques.

Feature description has drawn increasing attention. Descrip-tors can be regarded as distinguishable and robust representa-tions for given images and are widely used not only in image matching but also in image coding for image retrieval, face recognition, and other tasks that are based on image similar-ity measurements. However, direct similarity measurements for two image patches using raw image information will be regarded as an area-based image matching method, which will be reviewed in the next section. As for image patch-based feature descriptors, we will review the traditional ones, i.e., floating and binary descriptors, in terms of their data types. A new subsection will be added for the recent data-driven methods, including classical machine learning- and emerg-ing deep learning-based methods. We will comprehensively review handcrafted and learning-based feature description methods and show the connections among these methods to provide useful instructions for the readers toward their further research, especially for developing better description approaches using deep learning/CNN techniques. In addi-tion, we will also review the 3-D feature descriptors, where features are typically obtained from point data without any image pixel information but with spatial position relation-ships (e.g., 3-D point cloud registration).

**3.2 Handcrafted Feature Descriptors**

Handcrafted feature descriptors often depend on expert pri-ori knowledge, which are still widely used in many visual applications. Following the construction procedure of a tra-ditional local descriptor, the first step is to extract low-level information, which can be briefly classified into image gradi-ent and intensity. Subsequently, the commonly used pooling and normalizing strategies, such as statistic and comparison, are applied to generate long and simple vectors for discrim-inative description with respect to the data type (float or binary). Therefore, handcrafted descriptors mostly rely on the knowledge of their authors, and description strategies can be classified into gradient statistic-, local binary pat-tern statistic-, local intensity comparison- and local intensity order statistic-based methods.

**3.2.1 Gradient Statistic-Based Descriptors**

Gradient statistic methods are often used to form float type descriptors such as the histogram of oriented gradients (HOG) (Dalal and Triggs [2005](#page46)) as introduced in SIFT (Lowe et al. [1999](#page50); Lowe [2004](#page50)) and its improvement versions (Bay et al. [2006](#page44); Morel and Yu [2009](#page51); Dong and Soatto [2015](#page46); Tola et al. [2010](#page54)), and they are still widely used in several modern visual tasks. In SIFT, feature scale and orientation are respec-tively determined by DoG computation and the largest bin in a histogram of gradient orientation from a local circular region around the detected keypoint, thus achieving scale and rotation invariance. In the description stage, the local region of detected feature is first rectangularly divided into 4 × 4 non-overlapping grids based on the normalized scale and rotation, then a histogram of gradient orientation with 8 bins is conducted in each cell and embedded into a 128-dimensional float vector as the SIFT descriptor.

Another representative descriptor, namely, SURF (Bay et al. [2006](#page44)), can accelerate the SIFT operator by using the responses of Haar wavelets to approximate gradient com-putation; integral images are also applied to avoid repeated computation in Haar wavelet responses, enabling more effi-cient computation than SIFT. Other improvements based on these two typically focus on discrimination, efficiency, robustness, and coping with specific image data or tasks. For instance, CSIFT (Abdel-Hakim and Farag [2006](#page43)) uses additional color information to enhance the discrimination, and ASIFT (Morel and Yu [2009](#page51)) simulates all image views obtainable by varying the two camera axis orientation param-eters for fully affine invariance. Mikolajczyk and Schmid ([2005](#page51)) use a polar division and histogram statistics of gradi-ent orientations. SIFT-rank (Toews and Wells [2009](#page54)) has been proposed to investigate ordinal image description based on off-the-shelf SIFT for invariant feature correspondence. A Weber’s law-based method (WLD) (Chen et al. [2009](#page45)) has been studied to compute a histogram by encoding differen-tial excitations and orientations at certain locations.

Arandjelovi´c and Zisserman ([2012](#page44)) used a square root (Hellinger) kernel instead of the standard Euclidean dis-tance measurement to transform the original SIFT space to the RootSIFT space and yielded superior performance without increasing processing or storage requirements. Dong and Soatto ([2015](#page46)) modified SIFT by pooling the gradi-ent orientation across different domain sizes and proposed DSP-SIFT descriptor. Another efficient dense descriptor for wide-baseline stereo based on SIFT, namely, DAISY (Tola et al. [2010](#page54)), uses a log-polar grid arrangement and Gaussian pooling strategy to approximate the histograms of gradient orientations. Inspired by DAISY, DARTs (Marimon et al. [2010](#page50)) can efficiently compute scale space and reuse it for descriptors, thus resulting in high efficiency. Several handcrafted float-type descriptors have also been proposed

123

International Journal of Computer Vision

recently and shown promising performance; for example, the pattern of local gravitational force local descriptor (Bhat-tacharjee and Roy [2019](#page44)) is inspired from the law of universal gravitation and can be regarded as a combination of force magnitude and angle.

for fast computing and matching with low memory cost while remaining robust to scale, rotation, and noise. Handcrafted binary descriptors and classical machine learning techniques are also widely studied and these shall be introduced in the learning-based subsection.

**3.2.2 Local Binary Pattern Statistic-Based Descriptors** **3.2.4 Local Intensity Order Statistic-Based Descriptors**

Different from SIFT-like approaches, several intensity statistic-based methods, which are inspired by the local binary pattern (LBP) (Ojala et al. [2002](#page51)), have been proposed in the past decades. LBP has properties that favor its usage in inter-est region description, such as tolerance against illumination change and computational simplicity. The drawbacks are that the operator produces a rather long histogram and is insignificantly robust in flat image areas. Center-symmetric LBP (CS-LBP) (Heikkilä et al. [2009](#page48)) (using SVM for clas-sifier training) is a modified version of LBP combining the strengths of SIFT and LBP to address the flat area problem. Specifically, CS-LBP uses a SIFT-like grid and replaces the gradient information with an LBP-based feature. To address the noise, center-symmetric local ternary pattern (CS-LTP) (Gupta et al. [2010](#page47)) suggests the use of a histogram of rel-ative orders in patch and a histogram of LBP codes, such as histogram of relative intensities. The two CS-based meth-ods are designed to be more robust to Gaussian noise than previously considered descriptors. RLBP (Chen et al. [2013](#page45)) improves the robustness of LBP by changing the coding bit; a completed modeling of the LBP operator and an associ-ated completed LBP scheme (Guo et al. [2010](#page47)) have been developed for texture classification. LBP-like methods are widely used in texture representation and face recognition community, and additional details can be found in the review literature (Huang et al. [2011](#page48)).

**3.2.3 Local Intensity Comparison-Based Descriptors**

Another form of descriptors is based on the comparison of local intensities, which is also called binary descriptors and the core challenge is the selection rule for comparison. Because of their limited distinctiveness, these methods are mostly limited to short-baseline matching. Calonder et al. ([2010](#page45)) proposed the BRIEF descriptor built by concatena-tion of the results of a binary test of intensities for several random point pairs in image patch. Rublee et al. ([2011](#page53)) pro-posed rotated BRIEF combined with oriented FAST corners and selected robust binary tests using an machine learning strategy in their ORB algorithm to alleviate the limitations in rotation and scale change. Leutenegger et al. ([2011](#page49)) devel-oped the BRISK method using a concentric circle sampling strategy with increasing radius. Inspired by the retina struc-ture, Alahi et al. ([2012](#page44)) proposed the FREAK descriptor by comparing image intensities over a retinal sampling pattern

Thus far, many methods have been devised using orders of pixel values rather than raw intensities, achieving more promising performance (Tang et al. [2009](#page54); Toews and Wells [2009](#page54)). Pooling by intensity orders is invariant to rotation and monotonic intensity changes and also encodes ordi-nal information into descriptor; the intensity order-pooling scheme may enable the descriptors to be rotation-invariant without estimation of a reference orientation as SIFT, which appears as a major error source for most existing methods. To solve this problem, Tang et al. proposed the ordinal spa-tial intensity distribution (Tang et al. [2009](#page54)) method, which normalizes captured texture information and structure infor-mation using an ordinal and spatial intensity histogram; the proposed method is invariant to any monotonically increas-ing brightness changes.

Fan et al. ([2011](#page47)) pooled local features based on their gra-dient and intensity orders in multiple support regions and proposed the multi-support region order-based gradient his-togram and the multi-support region rotation and intensity monotonic invariant descriptor methods. A similar strategy was used in LIOP (Wang et al. [2011](#page55), [2015](#page55)), to encode the local ordinal information of each pixel. In that work, the over-all ordinal information was used to divide the local patch into subregions, which were used to accumulate LIOP. LIOP was further improved into OIOP/MIOP (Wang et al. [2015](#page55)), which can then encode overall ordinal information for noise and distortion robustness. They also proposed a learning-based quantization to improve its distinctiveness.

**3.3 Learning-Based Feature Descriptors**

Handcrafted descriptors, as reviewed above, require exper-tise to design and may disregard useful patterns hidden in the data. This requirement has prompted the investigations on learning-based descriptors, which have recently become dominantly popular due to their data-driven property and promising performance. In the following, we will discuss a group of classical learning-based descriptors introduced before the deep learning era.

**3.3.1 Classical Learning-Based Descriptors**

The learning-based descriptors can be traced back to PCA-SIFT (Ke et al. [2004](#page48)), in which principal component analysis (PCA) is used to form a robust and compact descriptor by

123

International Journal of Computer Vision

reducing the dimensionality of a vector made of the local image gradients. Cai et al. ([2010](#page45)) investigated the use of linear discriminant projections to reduce dimensionality and improve the discriminability of local descriptors. Brown et al. ([2010](#page45)) introduced a learning framework with a set of building blocks for constructing descriptors by using Powell mini-mization and linear discriminant analysis (LDA) technique to find the optimal parameters. Simonyan et al. ([2014](#page53)) pre-sented a novel formulation to represent the spatial pooling and dimensionality reduction in descriptor learning as con-vex optimization problems based on Brown’s work (Brown et al. [2010](#page45)). Meanwhile, Trzcinski et al. ([2012](#page54), [2014](#page54)) applied the boosting trick to learn boosted, complex non-linear local visual feature representations from multiple gradient-based weak learners.

Apart from the above-mentioned float-valued descrip-tors, binary descriptors are also of great interest in classical descriptor learning due to their beneficial properties, such as low storage requirements and high matching speed. A nat-ural way to obtain binary descriptors is to learn it from the provided float-valued descriptors. This task is convention-ally achieved by the hashing methods, thus suggesting that compact representations of high-dimensional data should be learned while maintaining their similarity in the new space. Locality sensitive hashing (LSH) (Gionis et al. [1999](#page47)) is arguably a popular unsupervised hashing method. This method generates embeddings via random projections and has been used for many large-scale search tasks. Some vari-ants of LSH include kernelized LSH (Kulis and Grauman [2009](#page49)), spectral hashing (Weiss et al. [2009](#page55)), semantic hashing (Salakhutdinov and Hinton [2009](#page53)) and p-stable distribution-based LSH (Datar et al. [2004](#page46)). These variants are unsuper-vised by design.

Supervised hashing methods have also been extensively investigated, where different machine learning strategies have been proposed to learn feature spaces tailored to specific tasks. In this case, a plethora of methods have been proposed (Kulis and Darrell [2009](#page49); Wang et al. [2010](#page55); Strecha et al. [2012](#page53); Liu et al. [2012a](#page50); Norouzi and Blei [2011](#page51); Gong et al. [2013](#page47); Shakhnarovich [2005](#page53)), among which image matching is considered an important experimental validation task. For example, the LDA technique is utilized in Strecha et al. ([2012](#page53)) to aid hashing. Semi-supervised sequential learning algorithms are proposed in Liu et al. ([2012a](#page50)) and Wang et al. ([2010](#page55)) to find discriminative projections. Minimal loss hash-ing (Norouzi and Blei [2011](#page51)) provided a new formulation to learn binary hash functions on the basis of structural SVMs with latent variables. Gong et al. ([2012](#page47)) proposed searching a rotation of zero-centered data to minimize the quantization error of mapping the descriptor to the vertices of a zero-centered binary hypercube.

Trzcinski and Lepetit ([2012](#page54)) and Trzcinski et al. ([2017](#page54)) reported that a straightforward way of developing binary

descriptors is to directly learn representations from image patches. In Trzcinski and Lepetit ([2012](#page54)), they proposed to project image patches to a discriminant subspace by using a linear combination of a few simple filters and then threshold their coordinates for creating the compact binary descrip-tor. The success of descriptors (e.g., SIFT) during image matching indicates that non-linear filters, such as gradient response, are more suitable than linear ones. Trzcinski et al. ([2017](#page54)) proposed to learn a hash function of the same form as an AdaBoost strong classifier, i.e. the sign of a linear com-bination of nonlinear weak learners, for each descriptor bit. This work is more general and powerful than Trzcinski and Lepetit ([2012](#page54)), which is based on simple thresholded lin-ear projections. Trzcinski et al. ([2017](#page54)) proposed to generate binary descriptors that are independently adapted per patch. This objective is achieved by inter- and intra-class online optimization for descriptors.

**3.3.2 Deep Learning-Based Descriptors**

Descriptors using deep techniques are usually formulated as a supervised learning problem. The objective is to learn a rep-resentation that can enable the two matched features to be as close as possible while the unmatched ones are far apart in the measuring space (Schonberger et al. [2017](#page53)). Descrip-tor learning is often conducted with cropped local patches centered on the detected keypoints; thus, it is also known as patch matching. In general, existing methods consist of two forms, namely, metric learning (Weinberger and Saul [2009](#page55); Zagoruyko and Komodakis [2015](#page56); Han et al. [2015](#page47); Kedem et al. [2012](#page48); Wang et al. [2017](#page55); Weinberger and Saul [2009](#page55)) and descriptor learning (Simo-Serra et al. [2015](#page53); Balntas et al. [2016a](#page44), [2017](#page44); Zhang et al. [2017c](#page56); Mishchuk et al. [2017](#page51); Wei et al. [2018](#page55); He et al. [2018](#page48); Tian et al. [2019](#page54); Luo et al. [2019](#page50)), according to the output of deep learning-based descriptors. These two forms are often jointly trained. Specifically, metric learning methods often learn a discriminative metric for simi-larity measurement with raw patches or generated descriptors as inputs. By contrast, descriptor learning tends to generate the descriptor representation from raw images or patches. Such a process requires a measurement method, such as L2 distance or trained metric network, for similarity evaluation. In contrast with single metric learning, the use of CNNs to generate description vectors is more flexible and may save time by avoiding repeated computation when a large number of candidate patches are available for correspondence search. Deep learning has achieved satisfying performance in feature description due to its strong ability in information extraction and representation.

Descriptors with deep learning techniques can be regarded as an extension of those based on classical learning (Schon-berger et al. [2017](#page53)). For instance, the Siamese structure in Chopra et al. ([2005](#page46)) and the commonly used loss func-

123

International Journal of Computer Vision

tions, such as hinge, Siamese, triplet, ranking, and contrastive losses, have been borrowed and modified in recent deep methods. Specifically, Zagoruyko and Komodakis ([2015](#page56)) proposed their DeepCompare and demonstrated the mech-anism by which to directly learn from raw image pixels with a general patch similarity function. In such scenario, various Siamese-type CNN models are applied to encode the sim-ilarity function. These models are then trained to identify the positive and negative image patch pairs. The attempted different network structures include Siamese with shared or unshared weights and central-surround form. MatchNet (Han et al. [2015](#page47)) is proposed to simultaneously learn the descriptor and metric. Such a technique is implemented by cascading a Siamese-like description network and fully con-volutional decision network. The task is converted into a classification problem under a cross-entropy loss. DeepDesc (Simo-Serra et al. [2015](#page53)) uses CNNs to learn discriminant patch representations together with L2 distance measuring. In particular, it trains a Siamese network with pairs of posi-tive and negative patches by minimizing the pairwise hinge loss, and the proposed hard negative mining strategy has alleviated the unbalanced positive and negative samples. Consequently, the description performance is siginificantly enhanced. Wang et al. ([2014](#page55)) proposed a novel deep rank-ing model to learn fine-grained image similarity. The model employs a triplet-based hinge loss and ranking function to characterize fine-grained image similarity relationships. A multiscale neural network architecture is utilized to capture the global visual properties and image semantics.

Kumar et al. ([2016](#page49)) first used the global loss to enlarge the distance margin between positive and negative patch pairs. It is implemented through triplet and Siamese networks trained with a combination of triplet and global losses. TFeat (Balntas et al. [2016b](#page44)) proposes to utilize triplets of training samples for CNN-based patch description and matching. It is implemented with shallow convolutional networks and fast hard negative mining strategy. In L2Net (Tian et al. [2017](#page54)), Tian et al. applied a progressive sampling strategy to optimize the relative distance-based loss function in the Euclidean space. The authors of that work considered the intermediate feature map and compactness of descriptor to achieve bet-ter performance. HardNet (Mishchuk et al. [2017](#page51)) achieves better improvement than L2Net by using a simple hinge triplet loss with the “hardest-within-batch” mining. PN-Net (Balntas et al. [2016a](#page44)) uses ideas introduced in the field of distance metric learning and online boosting by simultane-ously training with positive and negative constraints. The proposed SoftPN loss function exhibits faster convergence and lower error than hinge loss or SoftMax ratio (Wang et al. [2014](#page55); Zagoruyko and Komodakis [2015](#page56)). Zhang et al. ([2017c](#page56)) trained their networks by using their proposed global orthog-onal regularization together with triplet loss for encouraging

the descriptor to be sufficiently “spread out”. It was carried out to fully utilize the descriptor space.

Descriptor learning based on average precision attention (He et al. [2018](#page48)), introduces a general-purpose learning to rank formulation. This approach is defined to a constraint wherein the true matches should be ranked above all false path matches and is optimized on the basis of the binary and real-value local feature descriptors. BinGAN (Zieba et al. [2018](#page57)) proposes a regularization method for genera-tive adversarial networks (Goodfellow et al. [2014](#page47)) to learn discriminative yet compact binary representations of image patches. In comparison, other methods focused on binary descriptor learning are proposed in Erin Liong et al. ([2015](#page46)), Lin et al. ([2016a](#page49)) and Duan et al. ([2017](#page46)). Except for loss function, network structure, regularization and hard nega-tive mining, Wei et al. ([2018](#page55)) learned a discriminative deep descriptor by using kernelized subspace pooling. Tian et al. ([2019](#page54)) used second-order similarity in their SOSNet. In Con-textDesc, a more recent method, Luo et al. ([2019](#page50)) combined the local patch similarity constraint with the spatial geomet-rical constraint of interest point to train their networks, which largely improves the matching performance.[1](#page13)

As mentioned in the CNN-based detectors, an increas-ing number of end-to-end learning methods integrate the feature description together with the detectors into the com-plete matching pipeline. These methods are similar to those that have been singly designed for the description reviewed above. The main difference may lie on the way of training and the design of the entire network structure. The core challenge is to make the whole process differentiable and trainable. For example, LIFT (Yi et al. [2016](#page56)) attempts to simultaneously implement keypoint detection, orientation estimation, and feature description, by end-to-end CNN networks.

SuperPoint (DeTone et al. [2018](#page46)) proposes a self-supervised framework for training interest point detectors and descrip-tors for multiple view geometrical problems. The fully con-volutional model operates on full-sized images and jointly computes pixel-level interest point locations and associated descriptors, which is in contrast with path-based networks. LF-Net (Ono et al. [2018](#page51)) devises a two-branch setup and cre-ates virtual target responses iteratively to allow training from scratch without handcrafted priors. This technique realizes feature map generation, scale-invariant keypoint detection using top K selection and NMS, orientation estimation, and descriptor extraction. In LF-Net, the target function includes image level loss (satisfying additional constraints among image pairs, depth map, and essential matrix), patch-wise loss (learning keypoints that are good for matching and involves the orientation and scale component geometric con-sistency), and triplet loss for descriptor learning.

1. <https://image-matching-workshop.github.io/leaderboard/>.

123

International Journal of Computer Vision

Subsequently, RF-Net (Shen et al. [2019](#page53)) creates an end-to-end trainable matching framework that is modified from the LF-Net structure. First, the constructed receptive feature maps lead to effective keypoint detection. Second, a general loss function term, that is, neighbor mask, facilitates training patch selection to enhance the stability in descriptor train-ing. D2-Net (Dusmanu et al. [2019](#page46)) uses a single CNN to play a dual role: simultaneously achieving a dense feature descriptor and a feature detector. In Bhowmik et al. ([2020](#page44)), a keypoint selection and descriptor matching are optimized under high-level vision tasks by using principles from rein-forcement learning. In addition, Li et al. ([2020](#page49)) introduced dual-resolution correspondence networks to obtain pixel-wise correspondences in coarse-to-fine manner by extracting different resolution feature maps.

Except for feature matching for the same target or scene, semantic matching for images that are captured from sim-ilar targets/scenes has also been studied using CNNs and distinct promotion has been achieved. The semantic match-ing problem may pose a challenge for handcrafted methods due to the required understanding of semantic similarity. To this end, UCN (Choy et al. [2016](#page46)) uses deep metric learning to directly learn a feature space that preserves either geo-metric or semantic similarity. The use of such an approach also helps generate dense and accurate correspondences for either geometric or semantic correspondence tasks. Specif-ically, UCN implements a fully convolutional architecture with a correspondence contrastive loss for fast training and testing, and proposes a convolutional spatial transformer for local patch normalization. NCN (Rocco et al. [2018](#page52)) devel-ops an end-to-end trainable CNN architecture based on the classic idea of disambiguating feature matching by using semi-local constraints to find reliable dense correspondences between a pair of images. This framework identifies sets of spatially consistent matches by analyzing the neighbor-ing consensus patterns for a global geometric model. The model can be efficiently trained via weak supervision with-out any manual annotations of point correspondences. This type of framework can be applied for both category-level and instance-level matching tasks, and other similar methods are presented in Han et al. ([2017](#page47)), Plötz and Roth ([2018](#page52)), Chen et al. ([2018](#page45)), Laskar and Kannala ([2018](#page49)), Kim et al. ([2018](#page48), [2020](#page48)), Ufer and Ommer ([2017](#page54)) and Wang et al. ([2018](#page55)).

**3.4 3-D Feature Descriptors**

Extensive studies on 3-D feature descriptors have been con-ducted. As previously mentioned, many researchers have turned their attention to deep learning paradigm due to its revolutionary success in numerous different areas. This fact motivates us to categorize modern descriptors into two groups, i.e. handcrafted and learning-based ones. Guo et al. ([2016](#page47)) presented a comprehensive performance evaluation

of conventional handcrafted 3-D feature descriptors, while the learning-based methods are left out. In the following sec-tion, we provide a brief introduction of the state-of-the-art handcrafted descriptors and the learning-based ones.

**3.4.1 Handcrafted 3-D Descriptors**

Guo et al. ([2016](#page47)) divided the handcrafted descriptors into *spatial distribution histogram*- and *geometric attribute his-togram*-based descriptors, with the former representing thelocal feature by histograms that encode spatial distribu-tions of the points in the support region. In general, the local reference frame/axis is constructed for each keypoint. Accordingly, the 3-D support region is partitioned into bins to form a histogram. The values of each bin are calcu-lated by accumulating the spatial distribution measurements. Some representative work include spin image (Johnson and Hebert [1999](#page48)), 3-D shape context (Frome et al. [2004](#page47)), unique shape context (Tombari et al. [2010a](#page54)), rotational projection statistics (Guo et al. [2013](#page47)) and tri-spin-image (Guo et al. [2015](#page47)). The spatial distribution histogram descriptors repre-sent the local features by generating histograms from the statistics of geometric attributes (e.g., normals, curvatures) in the support region. These histograms include local sur-face patch (Chen and Bhanu [2007](#page45)), THRIFT (Flint et al. [2007](#page47)), point feature histogram (Rusu et al. [2008](#page53)), fast point feature histogram (Rusu et al. [2009](#page53)) and signature of his-togram of orientations (Tombari et al. [2010b](#page54)). Apart from the geometric attribute and spatial distribution histogram-based descriptors, Zaharescu et al. ([2009](#page56)) introduced the Mesh-HoG descriptor, which is analogous to SIFT (Lowe [2004](#page50)), and uses gradient information to generate a histogram.

The spectral descriptors, such as global point signature (Rustamov [2007](#page53)), HKS (Sun et al. [2009](#page54)) and wave kernel signature (WKS) (Aubry et al. [2011](#page44)), also make up an impor-tant category in this area. The descriptors are obtained from the spectral decomposition of the Laplace-Beltrami operator associated with the shape. The Global Point Signature (Rus-tamov [2007](#page53)) utilizes the eigenvalues and eigenfunctions of the Laplace–Beltrami operator on the shape to represent the local feature of points. The HKS (Sun et al. [2009](#page54)) and WKS (Aubry et al. [2011](#page44)) are based on the heat diffusion process and the temporal evolution of quantum mechanical particles on the shape, respectively.

**3.4.2 Learning-Based 3D Descriptors**

Efforts have also been devoted to generalizing spectral descriptors by using different learning schemes. Litman and Bronstein ([2014](#page49)) generalized the spectral descriptors to a generic family and proposed to learn from examples for obtaining optimized descriptors for a specific task. The learn-ing scheme resembles the spirit of Wiener filter in signal

123

International Journal of Computer Vision

processing. Rodolà et al. ([2014](#page52)) proposed a learning method that enables the wave kernel descriptor to recognize a broader class of deformations from the example set by using the random forest classifier. Windheuser et al. ([2014](#page55)) proposed a metric learning method to improve the representation of the spectral descriptors. Modern deep learning techniques have also been successfully applied. Masci et al. ([2015](#page50)) pro-posed the first attempt and introduced a generalization of the CNN paradigm to non-Euclidean manifolds for shape corre-spondences. Subsequently, Boscaini et al. proposed to learn descriptors by spectral convolutional networks (Boscaini et al. [2015](#page45)), and anisotropic CNNs (Boscaini et al. [2016](#page45)). Monti et al. ([2017](#page51)) proposed a unified framework for gener-alizing CNN architectures to non-Euclidean domains (graphs and manifolds). Xie et al. ([2016](#page55)) constructed a deep metric network to form a binary spectral shape descriptor for shape characterization. The input is based on the eigenvalue decom-position of the Laplace-Beltrami operator.

In the spatial domain, the differences of various deep learning methods often lie in the representation of the con-sumed data. Wei et al. ([2016](#page55)) trained a deep CNN on the depth map representation of shapes to find correspondences. Zeng et al. ([2017](#page56)) proposed to use a 3D deep CNN for learning a local volumetric patch descriptor. This descriptor consumes a voxel grid of truncated distance function val-ues of the local region. Elbaz et al. ([2017](#page46)) proposed a deep neural network auto-encoder to address the 3D matching problem. The authors used a random sphere cover set algo-rithm to detect feature points and project each local region into a depth map as input to the neural network for producing descriptors. Khoury et al. ([2017](#page48)) parameterized the input by using spherical histograms centered at each point and uti-lized fully connected networks to generate low-dimensional descriptors. Georgakis et al. ([2018](#page47)) recently employed a Siamese architecture network that processes depth maps. Zhou et al. ([2018](#page57)) proposed to learn from the images of multiple views for the description of 3D keypoints. Wang et al. ([2018b](#page55)) parameterized the multiscale localized neigh-borhoods of a keypoint into regular 2D grids as the input of a triplet-architecture CNN. Deng et al. ([2018](#page46)) first presented an order-free network on the basis of PointNet (Qi et al. [2017a](#page52)). This network can consume raw point clouds to exploit the full sparsity in the 3D matching task.

**3.5 Summary**

As previously mentioned, the image patch descriptor is des-ignated to enable accurate and effective correspondence establishment between detected feature points. The objective is to transform the original image information into a discrim-inative and stable representation that makes the two matched features as close as possible, while the unmatched ones are far apart. To this end, the descriptors should be easy to compute

with low computation and storage request. These descriptors should also maintain their discriminative and invariant fea-tures against serious deformations and imaging conditions. In the following section, we provide a comprehensive analysis of the handcrafted descriptors and introduce the mechanism by which the learning-based methods can partly address these challenges and achieve promising performance.

Following the construction procedure of traditional local descriptors, the first step is to extract the low-level infor-mation, which can be briefly classified into image gradient and intensity. Specifically, the gradient information can be regarded as a higher order image cue than raw intensity. The pooling strategy together with a histogram or statistic manner is often required to form a float descriptor. Thus, this strategy is more invariant to geometrical transformations (perhaps the pool and statistic strategy make it more independent to pixel position and geometrical variety). Nevertheless, it requires additional computation in gradient calculation and statistics as well as the distance measure of float-type data. LBP-based methods typically have high discriminative ability and good robustness to illumination change and image contrast, which are frequently used in texture representation and face recog-nition.

In contrast with the gradient and/or statistic-based meth-ods, the simple comparison strategy on image intensity would sacrifice great discrimination and robustness. A classical machine learning technique is often designed to identify sub-stantial useful bits. These types of methods are typically in need of the reference orientation estimation to achieve rota-tion invariant, which appears to be a major error source for most existing methods. However, the use of intensity order is intrinsically invariant to rotation and intensity changes without any geometrical estimation. It can achieve promising performance due to the combination of the use of intensity order and statistical strategy.

Learning-based methods have largely avoided the require-ment of manual experience and knowledge priori. They automatically optimize and obtain the optimal parameters and directly construct the wanted descriptor. Traditional learning methods aim to enable the generated descriptors superior in terms of efficiency, low storage, and discrimina-tion. However, the used image cues, such as intensity and gradient, are still with low order, and they highly rely on the framework in handcrafted methods. Nevertheless, the target function, training skills, and datasets that appeared at that time are significant and useful for designing better learning-based methods. Thus, the emergence of deep learning has further advanced this procedure in traditional learning.

Several skills can help improve the discriminability and robustness of deep descriptors. On the one hand, the central-surround and triplet (even more) structure may provide substantial significant information to learn. The hard nega-tive sample mining strategy would make the structure focused

123

International Journal of Computer Vision

on hard samples (may result in overfitting as well) and thus can achieve better matching performance. More reliable loss functions should also be designed according to the basic and intrinsic properties of description task. For instance, recently designed triplet, ranking, contrastive, and global losses, are superior than early simple hinge and cross-entropy losses. On the other hand, valid and comprehensive ground truth datasets are also required for better performance in match-ing and generalization ability. Training a descriptor together with detectors into the complete matching pipeline through an end-to-end manner has also drawn great attention at present. This can jointly optimize the detector and descriptor, thus can achieve encouraging performance, and the unsu-pervised training in it can perform without the need of any labeled ground truth patch data. The current descriptors can achieve significant matching performance across image pairs of appearance variances, such as illumination and day-night, by using deep techniques. However, these descriptors still suffer from serious geometrical deformation, such as large rotation or low-overlapped image pair. The low generaliza-tion ability for new types of data is also another limitation.

The overall performance of descriptor also depends on the appropriate detector. Different combinations of detectors and descriptors may result in varied matching performance. For this reason, the descriptors should be chosen according to a specific task and the type of image data. The advanced descriptors using deep learning have shown great potential.

**4 Matching Methods**

The matching task aims to establish the correct image pixel or point correspondences between two images with or with-out using the feature detection and/or description. This task has played a significant role for the entire image matching pipeline. Different definitions of matching task are intro-duced for specific applications and scenarios and may show their own strengths.

**4.1 Overview of Matching Methods**

Over the past decades in the image matching community, existing methods can be roughly classified into two cat-egories, saying *area-based* and *feature-based* (Zitova and Flusser [2003](#page57); Litjens et al. [2017](#page49)). Area-based methods typically refer to dense matching, also known as image regis-tration, which usually do not detect features. In feature-based methods, when the feature points and their local descriptors are extracted from the image pairs, the image matching task could be converted into matching them in indirect and direct ways, which correspond to the use and non-use of the local image descriptors.

Direct feature matching aims to establish the corre-spondences from two given feature sets by directly using the spatial geometrical relations and optimization meth-ods, which can be roughly classified into *graph matching* and *point set registration*. In comparison, indirect feature matching methods typically casts the matching task into a two-stage problem. Such task commonly starts with estab-lishing preliminary correspondences through the similarity of descriptors with the distance judging from the measur-ing space. Thereafter, the false matches are removed from the putative match sets by using extra local and/or global geometrical constraints. Dense matching from sparse feature correspondences often requires a post-process of transform model estimation, followed by image resampling and inter-polation (warping).

We will separate the learning-based methods from area-and feature-based methods and introduce them in a new sub-section. From the aspect of input data, learning from images and point data are the two main forms in learning-based matching. These methods can achieve better performance for some scenarios compared to the traditional ones. The matching task in 3-D cases is also briefly introduced in this section.

**4.2 Area-Based Matching**

Area-based methods aim for image registration and estab-lish dense pixel correspondences by directly using the pixel intensity of the entire image. A similarity metric together with an optimization method is in need for geometrical transformation estimation and common area alignment by minimizing the overall dissimilarity between the target and warped moving images. Consequently, several manual simi-larity metrics are frequently used, including correlation-like, domain transformation, and mutual information (MI) meth-ods. The optimization methods and transform models are also required to perform the final registration task (Zitova and Flusser [2003](#page57)).

In the image registration community, correlation-like methods, which are regarded as a classical representative in area-based methods, correspond two images by maximizing the similarities of two sliding windows (Zitova and Flusser [2003](#page57); Li et al. [2015](#page49)). For example, the maximum correlation of wavelet features has been developed for automatic regis-tration (Le Moigne et al. [2002](#page49)). However, this type of method may greatly suffer from the serious image deformations (can only be successfully applied when slight rotation and scaling are presented), windows containing a smooth area without any prominent details, and huge computational burden.

Domain transformed methods tend to align two images on the basis of converting the original images into another domain, such as phase correlation based on Fourier shift the-orem (Reddy and Chatterji [1996](#page52); Liu et al. [2005](#page50); Chen et al.

123

International Journal of Computer Vision

[1994](#page45); Takita et al. [2003](#page54); Foroosh et al. [2002](#page47)), and Walsh transform-based methods (Lazaridis and Petrou [2006](#page49); Pan et al. [2008](#page52)). Such methods are robust against the correlated and frequency-dependent noise and non-uniform, time vary-ing illumination disturbances. Nevertheless, these methods have some limitations in case of image pairs with signifi-cantly different spectral contents and small overlap area.

Based on information theory, the MI, such as non-rigid image registration using MI together with B-splines (Klein et al. [2007](#page48)) and conditional MI (Loeckx et al. [2009](#page50)), is a mea-surement of statistical dependency between two images and works with the entire image (Maes et al. [1997](#page50)). Thus, MI is particularly suitable for the registration of multi-modalities (Chen et al. [2003a](#page45), [b](#page45); Johnson et al. [2001](#page48)). Recently, Cao et al. ([2020](#page45)) proposed a structure consistency boosting trans-form to enhance the structural similarity in multi-spectral and multi-modal image registration problem, thus avoiding spectral information distortion. However, the MI exhibits difficulty in determining the global maximum of the entire searching space, inevitably reducing its robustness. More-over, optimization methods (e.g., continuous optimization, discrete optimization, and their hybrid form) and transfor-mation models (e.g., rigid, affine, thin plate spline (TPS), elastic body, and diffusion models) are considered suffi-ciently mature. Please refer to Zitova and Flusser ([2003](#page57)), Dawn et al. ([2010](#page46)), Sotiras et al. ([2013](#page53)) and Ferrante and Paragios ([2017](#page47)) for representative literature and further details.

The area-based methods are acceptable for medical or remote sensing image registration, which many feature-based methods are not workable anymore because the images often contain less textural details and large variance of image appearance due to the different imaging sensors. However, the area-based methods may greatly suffer from the serious geometrical transformations and local deformations. While deep learning has proven its efficacy, in which the early ones are usually employed as a direct extension of the classical registration framework, and later ones use a reinforcement learning paradigm to iteratively estimate the transformation, even directly estimate the deformative field in an end-to-end manner. The area-based matching with learning strategies will be reviewed in the part of learning-based matching.

**4.3 Graph Matching Methods**

Given the feature points extracted from an image, we can construct a graph by associating each feature point to a node and specifying edges. This procedure naturally pro-vides convenience to investigate the intrinsic structure of image data, especially for the matching problem. By this definition, graph matching (GM) refers to the establishment of node-to-node correspondences between two or multiple graphs. For its importance and fundamental challenge, GM

has been a long-standing research area over decades and is still of great interest to researchers. From the problem setting perspective, GM can be divided into two categories, namely, exact and inexact matching. Exact matching methods con-sider GM to be a special case of the graph or subgraph isomorphism problem. It aims to find the bijection of two binary (sub)graphs; consequently, all edges are strictly pre-served *babai2018groups,cook2006mining,levi1973note*). In fact, this requirement is too strict for real-world tasks like computer vision. Hence researchers often resort to inexact matching with weighted attributes on nodes and edges. Such an approach enjoys good flexibility and utility in practice. Therefore, we primarily concentrate on the review of inexact matching methods in this survey.

To some extent, GM possesses a simple yet general for-mulation of the feature matching problem, which encodes the geometrical cues into the node affinities (first-order rela-tions) and edge affinities (second-order relations) to deduce the true correspondences between two graphs. Aside from the geometrical cues, the high-level information of feature points can also be incorporated in GM (e.g. descriptor similarities as node affinities). This information only serves as a supplemen-tary one and is not necessarily required. In the general and recent form, GM can be formulated as a Quadratic Assign-ment Problem (QAP) (Loiola et al. [2007](#page50)). Although different forms exist in the literature, the main body of research has focused on the Lawler’s QAP (Lawler [1963](#page49)). Given two graphs *G*1 = *(V*1*,* *E*1*)* and *G*2 = *(V*2*,* *E*2*)*, where |*V*1 | = *n*1, |*V*2| = *n*2, each node *vi* ∈ *V*1 or *v* *j* ∈ *V*2 represents a feature point, and each edge *ei* ∈ *E*1 or *e* *j* ∈ *E*2 is defined over a pair of nodes. Without loss of generality we assume *n*1 ≥ *n* 2, Lawler’s QAP formulation of GM then can be written as:

max *J* *(***X***)* = vec*(***X***)* **K**vec*(***X***),*

(1)

*s.t.* **X**∈{0*,* 1}*n*1×*n*2 *,* **X1***n*2≤**1***n*1 *,* **X 1***n*1=**1***n*2 *,*

where **X** denotes the permutation matrix, i.e. **X***i j* = 1 indi-cates that node *vi* ∈ *V*1 corresponds to node *v* *j* ∈ *V*2 and **X***i j*=0 otherwise, vec*(***X***)*denotes the column-wise vector-ization of **X**, and **1***n*1 and **1***n*2 respectively denote the column vectors of all ones, **K** denotes the affinity matrix, whose diagonal and non-diagonal entries encode the first-order and second-order edge affinities between the two graphs. No universal approach can be utilized to construct the affinity matrix; however, a simple strategy is to use the similarities of feature descriptors [e.g. Shape Context (Belongie et al. [2001](#page44))] and differences of edge length to determine node and edge affinities.

The Koopmans–Beckmann’s QAP is another popular for-mulation. The form is different from Lawler’s QAP as expressed as:

|  |  |
| --- | --- |
| *J (***X***)* =tr*(***K** *p* **X***)* +tr*(***A**1**XA**2**X** *),* | (2) |

123

International Journal of Computer Vision

where **A**1 and **A**2 are the weighted adjacency matrices of the two graphs, respectively, and **K** *p* is the node affinity matrix. In Zhou and De la Torre ([2015](#page57)), the relation between Koopmans–Beckmann’s and Lawler’s QAP has been inves-tigated, which reveals that Koopmans–Beckmann’s QAP can be regarded as a special case of Lawler’s.

The GM problem is translated into finding the optimal one-to-one correspondences **X** that maximizes the overall affinity score *J* *(***X***)*. As a combinatorial QAP problem in general, GM is known to be NP-hard. Most methods relax the stringent constraints and provide approximate solutions in an affordable over head. In this regard, many relaxation strategies are introduced in the literature, thereby leading to a variety of GM solvers. In the following, we briefly review the influential ones through the development course of GM.

**4.3.1 Spectral Relaxations**

The first group of methods follow a strategy of spectral relax-ation. Leordeanu and Hebert ([2005](#page49)) proposed to replace the one-to-one mapping constraint and the binary constraint by constraining vec*(***X***)* 22 = 1. In this case, the solution **X** can be obtained by solving an eigenvector problem. Each element in **X** is interpreted as the association of one corre-spondence with the optimal cluster (true correspondences). A discretization strategy is used to enforce the mapping con-straints. The idea was later improved by Cour et al. ([2007](#page46)), who explicitly considered enforcing the one-to-one map-ping constraint to achieve tighter relaxation. This method can also be solved in closed forms as an eigenvector prob-lem. Liu and Yan ([2010](#page50)) proposed to detect multiple visual patterns by using a *l*1-norm-based spectral relaxation tech-nique, i.e. constraining vec*(***X***)* 1 = 1. The solution can be efficiently obtained by replicator equation from evolution-ary game theory. Jiang et al. ([2014](#page48)) presented a non-negative matrix factorization technique, which extends the constraint as vec*(***X***)* *p* = 1*,* *p* ∈ [1*,* 2]. Meanwhile, Egozi et al. ([2012](#page46)) presented a fairly different approach. In their work, they provided a probabilistic interpretation of spectral match-ing schemes and derived a novel probabilistic matching scheme wherein the affinity matrix is also updated in the iteration process. With Koopmans–Beckmann’s QAP for-mulation, the spectral methods (Umeyama [1988](#page54); Scott and Longuet-Higgins [1991](#page53); Shapiro and Brady [1992](#page53); Caelli and Kosinov [2004](#page45)) relax **X** to be orthogonal, i.e. **X X** = **I**. This expression can be solved in a closed form as an eigenvalue problem. These methods possess the merit of efficiency due to the loose relaxation. However, the accuracy is not advan-taged in general.

**4.3.2 Convex Relaxations**

Many studies have turned to investigating convex relaxations of the original problem to obtain theoretical advantages for solving the non-convex QAP issue. Strong convex relax-ations can be obtained by lifting methods that add auxiliary variables representing quadratic monomials in the original variables. This enables the addition of additional convex con-straints on the lifted variables. Semi-definite programming (SDP) is a general tool for combinatorial problems and has been applied to solving GM (Schellewald and Schnörr [2005](#page53); Torr [2003](#page54); Zhao et al. [1998](#page56); Kezurer et al. [2015](#page48)). The SDP relaxation is quite tight and allows finding a strong approxi-mation in polynomial time. However, the high computational cost prohibits its scalability. Some other lifting methods with linear programming relaxations have also been developed (Almohamad and Duffuaa [1993](#page44); Adams and Johnson [1994](#page44)). The dual problem of the LP relaxations are recently exten-sively considered to solve GM (Swoboda et al. [2017](#page54); Chen and Koltun [2015](#page45); Swoboda et al. [2017](#page54); Torresani et al. [2012](#page54); Zhang et al. [2016](#page56)), which has a strong link with the MAP inference algorithms.

**4.3.3 Convex-to-Concave Relaxations**

One useful strategy is to utilize the path-following technique. This approach gradually achieves a convex-to-concave pro-cedure of the original problem to finally find a good solution with the constraints satisfied. The computational complex-ity is also much lower than those of the lifting methods. Zaslavskiy et al. ([2009](#page56)) adopted this strategy for GM prob-lem with Koopmans–Beckmann’s QAP formulation, which is extended by to directed graphs (Liu et al. [2012b](#page50)) and par-tial matching (Liu and Qiao [2014](#page50)). Zhou and De la Torre ([2015](#page57)) presented a unified framework of GM based on the factorization of affinity matrix based on Lawler’s QAP. Such a framework effectively reduces the computational complex-ity and reveals the relation between Koopmans–Beckmann’s and Lawler’s QAPs. The (advanced) doubly stochastic (DS) relaxation methods improve upon these approaches by iden-tifying tighter formulations (Fogel et al. [2013](#page47); Dym et al. [2017](#page46); Bernard et al. [2018](#page44)), where the tightness of spectral, SDP, and DS relaxation is discussed and theoretically veri-fied.

123

International Journal of Computer Vision

**4.3.4 Continuous Relaxations**

A large volume of GM methods has focused on devising accu-rate or efficient algorithms to solve the QAP approximately, albeit with no global optimality guarantee. In most cases, **X** is simply relaxed to be continuous, as a DS matrix. Gold and Rangarajan ([1996](#page47)) proposed a graduated assignment algo-rithm, which performs gradient ascent on the relaxed problem under an annealing schedule. The convergence of this method has been revisited and improved by Tian et al. ([2012](#page54)) with a soft constrained mechanism. van Wyk and van Wyk ([2004](#page54)) proposed to enforce the one-to-one mapping constraint by successively projecting onto the convex set of the desired integer constraints. Leordeanu et al. ([2009](#page49)) proposed an effi-cient algorithm that optimizes in the (quasi) discrete domain via solving a sequence of linear assignment problems. Many famous optimization techniques, such as ADMM (Lê-Huu and Paragios [2017](#page49)), tabu search (Adamczewski et al. [2015](#page44)) and multiplicative update algorithm (Jiang et al. [2017a](#page48)), have also been tested. Recent studies also include Jiang et al. ([2017b](#page48)) and Yu et al. ([2018](#page56)), which introduce new schemes to asymptotically approximate the original QAP, and Maron and Lipman ([2018](#page50)), which presents a new (probably) concave relaxation technique. Yu et al. ([2020b](#page56)) introduced a determi-nant regularization technique together with gradient-based optimization to relax this problem into continuous domain.

**4.3.5 Multi-graph Matching**

In contrast to the classic two-graph matching setting, jointly matching a batch of graphs with consistent correspondences, i.e. multi-graph matching, has recently drawn increasing attention due to its methodological advantage and potential to incorporate cross-graph information. Arguably, one cen-tral issue of multi-graph matching lies in the enforcement of cycle-consistency for a feasible solution. In general, this concept refers to the fact that the bijection correspondence between two graphs shall be consistent with a derived one through an intermediate graph. Put it more concretely, for any pair of graphs *Ga* and *Gb* with their node correspon-dence matrix **X***ab*, let *Gc* be an intermediate graph, the cycle consistency constraint is enforced: **X***ac***X***cb* = **X***ab*, where **X***ac*and **X***cb*are the matching solutions of*Ga*and*Gc*and

1. *c* and *Gb*, respectively.

Existing multi-graph matching methods can be roughly grouped into three lines of works. For the methods falling into the first group, the multi-graph matching problem is solved by an iterative procedure for computing a number of two-graph matching tasks (Yan et al. [2013](#page55), [2014](#page55), [2015a](#page55), [b](#page56); Jiang et al. [2020b](#page48)). In each iteration, a two-graph matching solution is computed to locally maximize the affinity score, which can leverage off-the-shelf pairwise matching solvers, such as in Jiang et al. ([2020b](#page48)), both offline batch mode and

online setting are considered to explore the concept of cycle-consistency over pairwise matching. Another body of work takes the initial (noisy) pairwise matching result as input, and aims to recover a globally consistent pairwise matching set (Kim et al. [2012](#page48); Pachauri et al. [2013](#page52); Huang and Guibas [2013](#page48); Chen et al. [2014](#page45); Zhou et al. [2015](#page57); Wang et al. [2018](#page55); Hu et al. [2018](#page48)). In these methods, matching over all graphs is jointly and equally considered to form a bulk matrix that includes all pairwise matchings. The intrinsic structure of this matrix induced by the matching problem, such as cycle-consistency, is investigated. The last group utilizes clustering or low rank recovery techniques to solve multi-graph match-ing, which provides a new perspective in the feature space for the problem (Zeng et al. [2012](#page56); Yan et al. [2015c](#page55), [2016a](#page55); Tron et al. [2017](#page54)). More recently, the multi-graph matching problem has been considered in the optimization framework with a theoretically well-grounded convex relaxation (Swo-boda et al. [2019](#page54)), or with projected power iterations to search for a feasible solution (Bernard et al. [2019](#page44)).

**4.3.6 Other Paradigms**

Although the QAP formulation is prevalent in GM, the way of formulation is not unique. Numerous methods deal with GM from different perspectives or paradigms and also form an important category in this field.

Cho et al. ([2010](#page45)) provided a random walk view of GM and devised a technique to obtain solution by simulating random walks on the association graph. Lee et al. ([2010](#page49)) and Suh et al. ([2012](#page54)) introduced Monte Carlo methods to improve the matching robustness. Cho and Lee ([2012](#page45)) further devised a progressive GM method, which combines pro-gression of graphs with matching of graphs to reduce the computational complexity. Wang et al. ([2018a](#page55)) proposed to use a functional representation of graphs and conduct match-ing by minimizing the discrepancy between the original and the transformed graphs. Subsequently, in order to suppress the matching of outliers, Wang et al. ([2020](#page55)) assigned zero-valued vectors to the potential outliers in the obtained optimal correspondence matrix. The affinity matrix plays a key role in the GM problem. However, the handcrafted **K** is vulnerable to scale and rotation differences. To this end, unsupervised (Leordeanu et al. [2012](#page49)) and supervised (Caetano et al. [2009](#page45)) methods are devised to learn **K**. Zanfir and Sminchisescu ([2018](#page56)) recently addressed this issue with an end-to-end deep learning scheme. Wang et al. ([2020](#page55)) introduced a fully train-able framework for graph matching. In this framework, they utilized a graph network block module and simultaneously considered the learning of node/edge affinities and the solv-ing of combinatorial optimization.

123

123

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | International Journal of Computer Vision |  |
|  | | | |  |  |
| The extension of GM to a high-order formulation is a | | | | (ICP) algorithm is a popular method (Besl and McKay [1992](#page44)). |  |
| natural way to improve the robustness by mostly exploring | | | | ICP iteratively alternates between hard assignments of cor- |  |
| the geometrical cues. This leads to a tensor-based objective | | | | respondences for the closest points in two point sets and |  |
| (Lee et al. [2011](#page49)) also called hypergraph matching: | | | | the closed-form rigid transformation estimation until con- |  |
| *JH (***X***)* =**H**⊗1**x**⊗2**x** *. . .* ⊗*m* **x***,* | | |  | vergence. The ICP algorithm is widely used as baselines due |  |
| (3) | to its simplicity and low computational complexity. How- |  |
| where *m* is the order of affinities, H denotes the *m*-order tensor | | | | ever, a good initialization is required because ICP is prone |  |
| to be trapped into local optima. Numerous studies, such as |  |
| encoding the affinities between hyperedges in the graphs, ⊗*k* | | | | EM-ICP (Granger and [Pennec 2002](#page47)), LM-ICP (Fitzgibbon |  |
| is the tensor product, and **x** = vec*(***X***)*. Representative studies | | | | [2003](#page47)), and TriICP (Chetverikov et al. [2005](#page45)), in the research |  |
| on hypergraph matching include Zass and Shashua ([2008](#page56)), | | | | field of PSR have been proposed to improve ICP. The reader |  |
| Chertok and Keller ([2010](#page45)), Lee et al. ([2011](#page49)), Chang and | | | | is referred to a recent survey (Pomerleau et al. [2013](#page52)) for |  |
| Kimia ([2011](#page45)), Duchenne et al. ([2011](#page46)) and Yan et al. ([2015d](#page55)). | | | | a detailed discussion of ICP’s variants. The robust point |  |
| **4.4 Point Set Registration Methods** | | |  | matching (RPM) algorithm (Gold et al. [1998](#page47)) are proposed |  |
|  | to overcome the ICP limitations; the soft assignment and |  |
| Point set registration (PSR) aims to estimate the spatial trans- | | | | deterministic annealing strategy are adopted, and the rigid |  |
| transformation model is generalized to a non-rigid one by |  |
| formation that optimally aligns two point sets. In feature | | | | using the thin-plate spline [TPS-RPM (Chui and Rangarajan |  |
| matching, different formulations are adopted in PSR and GM. | | | | [2003](#page46))]. |  |
| For two point sets, GM methods determine the alignment | | | |  |  |
| via maximizing the overall affinity score of unary corre- | | | | **4.4.2 EM-Based Methods** |  |
| spondence and pairwise correspondences. By contrast, PSR | | | |  |  |
| methods determine the underlying global transformation. | | | | RPM is also a representative of the EM-like PSR methods, |  |
| Given the two point sets {**x***i* }*in*=11 and {**y***i* }*in*=2 | | | 1, the general | which form an important category in this field. The EM-like |  |
| conventional objective can be expressed as | | |  | methods formulate PSR as an optimization problem of either |  |
| min *J* *(***P***,* *θ* *)* = | | *pi j* **y** *j* − *T (***x***i , θ )* 22+ *g(***P***)* | | a weighted squared loss function or the log-likelihood max- |  |
| imization of Gaussian mixture models (GMMs), and local |  |
|  |  | *i , j* | (4) | optimum is searched through EM or EM-like algorithms. The |  |
| *s.t. θ* ∈ | **P** ∈{0*,*1}*n*1×*n*2*,* **P1***n*2≤ **1***n*1*,* **P** | | **1***n*1≤**1***n*2 *,* | posterior probability of each correspondence is computed in |  |
|  |  |  |  | the E-step, and the transformation is refined in the M-step. |  |
| where *θ* denotes the parameters of the predefined transfor- | | | | Sofka et al. ([2007](#page53)) investigated the modeling of uncertainty |  |
| mation. The regularization term *g(***P***)* avoids trivial solutions, | | | | in the registration process and presented a covariance driven |  |
| such as **P** = **0**. Compared to GM, this model only repre- | | | | correspondence method in an EM-like framework. Myro- |  |
| sents the general principles, but does not necessarily cover | | | | nenko and Song ([2010](#page51)) proposed the well-known coherent |  |
| all the algorithms for PSR. For example, a probabilistic inter- | | | | point drift (CPD) method in which a probabilistic framework |  |
| pretation or a density-based objective can be used, and the | | | | is established on the basis of GMM; here, the EM algo- |  |
| constraints for **P** may be only partially imposed during opti- | | | | rithm is utilized for maximum likelihood estimation of the |  |
| mization, which all differ from the above formulation. | | | | parameters. Horaud et al. ([2011](#page48)) developed an expectation |  |
| PSR poses a stronger assumption on the data, that is, | | | | conditional maximization-based probabilistic method, which |  |
| the existence of a global transformation between point sets, | | | | allows the use of anisotropic covariance for the mixture |  |
| which is the key feature that differentiates it from GM. | | | | model components and improves over isotropic covariance |  |
| Although the generality is restricted, this assumption leads to | | | | case. Ma et al. ([2016b](#page51)) and Zhang et al. ([2017a](#page56)) exploited the |  |
| low computational complexity because of the few parameters | | | | unification of local feature and global feature in the GMM- |  |
| needed for global transformation models. A sophisticated | | | | based probabilistic framework. Lawin et al. ([2018](#page49)) presented |  |
| transformation model is developed from rigid to non-rigid | | | | a density adaptive PSR method via modeling the underlying |  |
| ones in order to enhance the generalization ability. Various | | | | structure of the scene as a latent probability distribution. |  |
| schemes are also proposed to improve robustness against | | | |  |  |
| degradations, such as noise, outliers, and missing points. | | | | **4.4.3 Density-Based Methods** |  |
| **4.4.1 ICP and Its Variants** | | |  | Density-based methods introduce generative models to the |  |
|  |  |  |  | PSR problem, in which no explicit point correspondence |  |
| PSR has been an important research topic for the last few | | | | is established. Each point set is represented by a density |  |
| decades in computer vision, and the iterative closest point | | | | function, such as GMM. Registration is achieved by the mini- |  |

International Journal of Computer Vision

mization of a statistical discrepancy measure between the two density functions. Tsin and Kanade ([2004](#page54)) were the first to propose such a method and used kernel density functions to model the point sets, and the discrepancy measure is defined as kernel correlation. Meanwhile, Glaunes et al. ([2004](#page47)) rep-resented the point sets by using relaxed Dirac delta functions. They then determined the optimal diffeomorphic transforma-tion that minimizes the distance of the two distributions. Jian and Vemuri ([2011](#page48)) extended this approach by using GMM-based representation and minimizing the L2 error between the densities. The authors also provided a unified framework of density-based PSR. Many popular methods, including Myronenko and Song ([2010](#page51)) and Tsin and Kanade ([2004](#page54)) can be regarded as special cases in theory. Campbell and Peters-son ([2015](#page45)) proposed to use a support vector parameterized GMM for adaptive data representation. This approach can improve the robustness of density-based methods to noise, outliers, and occlusions. Recently, Liao et al. ([2020](#page49)) utilized fuzzy clusters to represent a scanned point set, then registed two point sets by minimizing a fuzzy weighted sum of dis-tances between their fuzzy cluster centers.

**4.4.4 Optimization-Based Methods**

A group of optimization-based methods have been proposed as globally optimal solutions to alleviate the local optimum issue. These methods generally search in a limited transfor-mation space for timing saving, such as rotation, translation, and scaling. Stochastic optimization techniques, including genetic algorithms (Silva et al. [2005](#page53); Robertson and Fisher [2002](#page52)), particle swarm optimization (Li et al. [2009](#page49)), parti-cle filtering (Sandhu et al. [2010](#page53)) and simulated annealing schemes (Papazov and Burschka [2011](#page52); Blais and Levine [1995](#page45)), are widely used, but no convergence is guaranteed. Meanwhile, Branch and bound (BnB) is a well-established optimization technique that can efficiently search the glob-ally optimal solution in the transformation space and form the theoretical basis of many optimization-based methods, including Li and Hartley ([2007](#page49)), Parra Bustos et al. ([2014](#page52)), Campbell and Petersson ([2016](#page45)), Yang et al. ([2016](#page56)) and Liu et al. ([2018b](#page50)). In addition to these methods, Maron et al. ([2016](#page50)) introduced a semidefinite programming (SDP) relaxation-based method, in which a global solution is guar-anteed for isometric shape matching. Lian et al. ([2017](#page49)) formulated PSR as a concave QAP by eliminating the rigid transformation variables, and BnB is utilized to achieve a globally optimal solution. Yao et al. ([2020](#page56)) presented a formulation for robust non-rigid PSR based on a globally smooth robust estimator for data fitting and regularization, which is optimized by majorization-minimization algorithm to reduce each iteration in solving a simple least-squares problem. Another method in Iglesias et al. ([2020](#page48)) presents a study of global optimality conditions for PSR with missing

data. This method applies Lagrangian duality to generate a candidate solution for the primal problem thus enables it to obtain the corresponding dual variable in a closed form.

**4.4.5 Miscellaneous Methods**

Apart from the commonly used rigid model or non-rigid transformation model based on TPS (Chui and Rangara-jan [2003](#page46)) or Gaussian radial basis functions (Myronenko and Song [2010](#page51)), additional complex deformations are also considered in the literature. These models include simple articulated extensions, such as Horaud et al. ([2011](#page48)) and Gao and Tedrake ([2019](#page47)). A smooth locally affine model is intro-duced as the transformation model and developed under the ICP framework in non-rigid ICP (Amberg et al. [2007](#page44)), which is also adopted in Li et al. ([2008](#page49)). However, this model should be used in conjunction with sparse hand selected feature correspondences as it allows many degrees of freedom. A different linear skinning model, which does not require user’s involvement in the registration process, has been proposed and applied in another work (Chang and Zwicker [2009](#page45)).

Another line of PSR methods introduce shape descriptors into the registration process. Local shape descriptors, such as spin images (Johnson and Hebert [1999](#page48)), shape contexts (Belongie et al. [2001](#page44)), integral volume (Gelfand et al. [2005](#page47)) and point feature histograms (Rusu et al. [2009](#page53)) are gener-ated. Sparse feature correspondences are established by a similarity constraint of descriptors. Subsequently, the under-lying rigid transformation can be estimated using random sampling consensus (RANSAC) (Fischler and Bolles [1981](#page47)) or BnB search (Bazin et al. [2012](#page44)). Ma et al. ([2013b](#page50)) proposed a robust algorithm based on the *L*2 *E* estimator in a non-rigid case.

Some new schemes for PSR based on different observa-tions have emerged. Golyanik et al. ([2016](#page47)) modeled point set as particles with gravity as attractive force, and registration is accomplished by solving the differential equations of New-tonian mechanics. Ma et al. ([2015a](#page51)) and Wang et al. ([2016](#page55)) proposed the use of context-aware Gaussian fields to address the PSR problem. Vongkulbhisal et al. ([2017](#page54), [2018](#page55)) proposed the discriminative optimization method. This approach learns the search direction from training data to guide optimization without the need of defining cost functions. Danelljan et al. ([2016](#page46)) and Park et al. ([2017](#page52)) considered the color informa-tion of point sets, whereas Evangelidis and Horaud ([2018](#page46)) and Giraldo et al. ([2017](#page47)) addressed the problem of joint reg-istration of multiple point sets.

**4.5 Descriptor Matching with Mismatch Removal**

Descriptor matching followed by mismatch removal, also called indirect image matching, casts the matching task into a two-stage problem. This method commonly starts with estab-

123

International Journal of Computer Vision

lishing preliminary correspondences through the similarity of local image descriptors with the distance judging from the measuring space. Several common strategies, including fixed threshold (FT), nearest neighbor (NN) also called brute force matching, mutual NN (MNN), and NN distance ratio (NNDR), are available for the construction of putative match sets. Thereafter, the false matches are removed from the puta-tive match sets by using extra local and/or global geometrical constraints. We briefly divide the mismatch removal meth-ods into resampling-based, non-parametric model-based, and relaxed methods. In the following sections, we will introduce these methods in detail and provide comprehensive analysis.

**4.5.1 Putative Match Set Construction**

Suppose that we have detected and extracted *M* and *N* local features to be matched from the considering two images *I*1 and *I*2. The descriptor matching stage operates by computing the pairwise distance matrix with *M* × *N* entries and then selecting the potential true matches through the aforemen-tioned rule.

The FT strategy considers the matches with their distances below a fixed threshold. However, this strategy can be sen-sitive and may incur numerous one-to-many matchings in contrast to the one-to-one correspondence nature. This situ-ation results in poor performance in feature matching task. The NN strategy can effectively deal with the data sensitivity problem and recall more potential true matches. Such a strat-egy has been applied in various descriptor matching methods, but it cannot avoid the one-to-many cases. In mutual NN descriptor matching, each feature in *I*1, looks for its NN in *I*2(and vice versa), and the feature pairs that are mutual NNbecome candidate matches in the putative match set. This type of strategy can obtain high ratio of correct matches but may sacrifice many other true correspondences. The NNDR considered that the distance difference between first and sec-ond NN is significant. Hence, the use of the distance ratio with a predefined threshold would obtain robust and promising matching performance while not sacrifice many true matches. However, NNDR relies on the stable distance distribution of these descriptors even though the method is widely used and well performed in SIFT-like descriptor matching. In fact, NNDR is no longer applicable for descriptors of other types, such as binary or some learning based descriptors (Rublee et al. [2011](#page53); Ono et al. [2018](#page51)).

The optimal choice of these methods for descriptor match-ing should rely on the property of descriptor and the specific application. For example, the MNN is stricter than others with high inlier ratio but may sacrifice many other potential true matches. By contrast, NN and NNDR tend to be more general in feature matching task with relatively better performance. Mikolajczyk and Schmid ([2005](#page51)) proposed a simple test about these candidate match selection strategies. Although various

approaches are available for putative feature correspondence construction the use of only local appearance information and simple similarity-based putative match selection strate-gies, will unavoidably result in a large number of incorrect matches, particularly when images undergo serious non-rigid deformation, extreme viewpoint changes, low quality, and/or repeated contents. Therefore, a robust, accurate, and efficient mismatch elimination method is urgently required in the sec-ond stage to preserve as many true matches as possible while keeping the mismatch to a minimum by using additional geo-metrical constraints.

**4.5.2 Resampling-Based Methods**

Resampling technique is (arguably) a prevalent paradigm and is represented by the classic RANSAC algorithm (Fischler and Bolles [1981](#page47)). Basically, the two images are assumed to be coupled by a certain parametric geometric relation, such as projective transformation or epipolar geometry. The RANSAC algorithm then follows a hypothesize-and-verify strategy: repeatedly sample a minimal subset from the data, e.g. four correspondences for projective transformation and seven correspondences for fundamental, estimate a model as hypothesis, and verify the quality by the number of consis-tent inliers. Finally, the correspondences consistent with the optimal model are recognized as inliers.

Various methods have been proposed to improve the per-formance of RANSAC. In MLESAC (Torr and Zisserman [1998](#page54), [2000](#page54)), the model quality is verified by a maximum likelihood process, which albeit under certain assumptions, can improve the results and is less sensitive to the pre-defined threshold. The idea of modifying the verification stage is not only utilized but also further extended in many following studies due to the simple implementation. The modification of sampling strategy has also been considered in quite a few studies due to the appealing result of efficiency enhancement. In essence, diverse prior information is incor-porated to increase the probability of selecting an all-inlier sample subset. Specifically, the inliers are assumed to be spa-tially coherent in NAPSAC (Nasuto and Craddock [2002](#page51)), or exist with some groupings in GroupSAC (Ni et al. [2009](#page51)). PROSAC (Chum and Matas [2005](#page46)) exploits a priori predicted inlier probability, and EVSAC (Fragoso et al. [2013](#page47)) uses an estimate of confidence with extreme value theory of the correspondences. Another seminal work is the locally opti-mized RANSAC (LO-RANSAC) (Chum et al. [2003](#page46)), with the key observation that taking minimal subsets can amplify the underlying noise and yield hypotheses that are far from the ground truth. This problem is addressed by introducing a local optimization procedure when arriving at the *so-far-the-best* model. In the original paper, local optimization isimplemented as an iterated least squares fitting process with a shrinking inlier-outlier threshold inside an inner RANSAC.

123

International Journal of Computer Vision

This has a large-than-minimal sampling and is applied only to the inliers of the current model. The computational cost issue of LO-RANSAC is addressed in Lebeda et al. ([2012](#page49)), where several implementation improvements are suggested. The local optimization step is augmented with a graph-cut tech-nique in Barath and Matas ([2018](#page44)). Many improving strategies for RANSAC are integrated in USAC (Raguram et al. [2012](#page52)).

More recently, Barath et al. ([2019b](#page44)) applyed *σ* -consensus in their MAGSAC, to eliminate the need of a user-defined threshold by marginalizing over a range of noise scales. Whereafter, observing that nearby points are more likely to originate from the same geometric model, Barath et al. ([2019a](#page44)) extracted the local structure for global sampling and parameter model estimation by drawing samples from grad-ually growing neighborhoods. Based on above two methods, they introduced MAGSAC++ (Barath et al. [2020](#page44)) with a new scoring function. This method avoids requiring the inlier-outlier decision, in which a novel marginalization procedure formulated as an M-estimation is solved by an iteratively re-weighted least squares procedure, and the progressive growing sampling strategy in Barath et al. ([2019a](#page44)) is also applied for RANSAC-like robust estimation.

Some fundamental shortcomings are exhibited by the resampling methods despite their efficacy in wide appli-cations of computer vision. For example, the theoretically required runtime exponentially grows with the increase of outlier rate. The minimal subset sampling strategy only applies to parametric models and fails to handle image pairs undergoing complex transformations, such as non-rigid ones. This situation motivates researchers to develop new algo-rithms divorced from the resampling paradigm.

**4.5.3 Non-parametric Model-Based Methods**

A group of non-parametric model-based methods have been proposed. Instead of simple parametric models, non-parametric models address more general priors in matching, e.g. motion coherence, and can deal with degenerated sce-narios. These methods are distinguished by different defor-mation functions to model the transformation and different means to cope with gross outliers. Pilet et al. ([2008](#page52)) proposed the use triangulated 2-D mesh to model the deformation using a tailored robust estimator for eliminating the detrimental effect of outliers. The idea of robust estimators is also lever-aged in Gay-Bellile et al. ([2008](#page47)), with Huber estimator, and Ma et al. ([2015](#page50)), with *L*2 *E* estimator, despite of their dif-ferent modeling of deformation. A fairly different method is proposed in Li and Hu ([2010](#page49)), in which the Support Vec-tor Regression technique is employed to robustly estimate a *correspondence function* and reject mismatches.

The seminal work vector field consensus (VFC) (Ma et al. [2013a](#page51), [2014](#page51)) introduces a new framework for non-rigid matching. The deformation function is restricted within

the reproducing kernel Hilbert space in association with Tikhonov regularization to enforce the smoothness con-straint. The estimation is conducted in a Bayesian model, where the outliers are explicitly considered for robustness. The VFC algorithm, and its variants (Ma et al. [2015b](#page51), [2017a](#page50), [2019b](#page51)) have been proven effective.

**4.5.4 Relaxed Methods**

The recent trend has been towards developing relaxed meth-ods for matching, where the geometric constraint is made less strict to accommodate even complex scenarios, such as motion discontinuities arising from image pairs of wide baselines or with objects undergoing independent motions. Certain GM methods (Leordeanu and Hebert [2005](#page49); Liu and Yan [2010](#page50)) are available for such requirements and use quadratic models that incorporate pairwise geometric rela-tions of correspondences to find the potentially correct ones. However, the results are often coarse.

Lipman et al. ([2014](#page49)) considered deformations that are piecewise affine; they then formulated feature matching into a constrained optimization problem that seeks for such a deformation consistent with the most correspondences and exerts a bounded distortion. Lin et al. ([2014](#page49), [2017](#page49)) pro-posed to identify true matches with likelihood functions estimated using nonlinear regression technique in a specially designed domain of correspondence, where motion coher-ence is imposed, while discontinuities are also allowed. This concept corresponds to enforcing a local motion coherence constraint. Ma et al. ([2018a](#page50), [2019d](#page51)) presented a locality pre-serving approach for matching, whereby a global distortion model for matching is relaxed to focus on the locality of each correspondence in exchange for generality and effi-ciency. The derived criterion has been proven able to rapidly and accurately filter erroneous matches. A similar method appeared in Bian et al. ([2017](#page44)) wherein a simple criterion based on local supporting matches to reject outliers is intro-duced. Jiang et al. ([2020a](#page48)) casted feature matching as a spatial clustering problem with outliers to adaptively cluster the putative matches into several motion consistent clusters together with an outlier/mismatch cluster. Another method in Lee et al. ([2020](#page49)) formulates the feature matching prob-lem as a Markov random field that uses both local descriptor distance and relative geometric similarities to enhance the robustness and accuracy.

**4.6 Learning for Matching**

Apart from detectors or descriptors, learning-based matching methods are commonly used to substitute traditional meth-ods in information extraction and representation or model regression. The matching step by learning can be roughly classified into image-based and point-based learning. Based

123

International Journal of Computer Vision

on the traditional methods, the former aims to cope with three typical tasks, namely image registration (Wu et al. [2015a](#page55)), stereo matching (Poursaeed et al. [2018](#page52)) and camera local-ization or transformation estimation (Poursaeed et al. [2018](#page52); Erlik Nowruzi et al. [2017](#page46); Yin and Shi [2018](#page56)). Such a method can directly realize task-based learning without attempting to detect any salient image structure (e.g. interest points) in advance. By contrast, point-based learning prefers conduct-ing on the extracted point sets; such methods are commonly used for point data processing, such as classification, seg-mentation (Qi et al. [2017a](#page52), [b](#page52)) and registration (Simonovsky et al. [2016](#page53); Liao et al. [2017](#page49)). Researchers have also used these for correct match selection and geometrical transfor-mation model estimation from putative match sets (Moo Yi et al. [2018](#page51); Ma et al. [2019a](#page50); Zhao et al. [2019](#page56); Ranftl and Koltun [2018](#page52); Poursaeed et al. [2018](#page52)).

**4.6.1 Learning from Images**

Matching methods of image-based learning often use CNNs for image-level latent information extraction and similar-ity measurement, as well as geometrical relation estimation. Therefore, the patch-based learning (Sect. [3.3](#page11): learning-based feature descriptors) is frequently used as an extension of area-based image registration and stereo matching. This is because traditional similarity measurements in a sliding window can be easily replaced with a deep manner, i.e., deep descriptors. However, the success achieved by researchers in using deep learning in spatial transformation networks (STN) (Jaderberg et al. [2015](#page48)) and optical flow estimation (FlowNet) (Dosovitskiy et al. [2015](#page46)) has aroused a wave of studies on directly estimating the geometrical transformation or non-parametric deformation field with deep learning techniques, even achieving an end-to-end trainable framework.

*Image registration.* For area-based image registration,early deep learning is generally used as a direct extension of the classical registration framework, and later use the reinforcement learning paradigm to iteratively estimate the transformation, even directly estimate the deformative field or displacement field for the registration task. The most intu-itional approach is to use deep learning networks to estimate the similarity measurement for the target image pair in order to drive an iterative optimization procedure. In this way, the classical measure metrics, such as the correlation-like and MI methods, etc., can be substituted with more supe-rior deep metrics. For instance, Wu et al. ([2015a](#page55)) achieved deformable image registration by using the convolutional stacked auto-encoder (CAE) to discover compact and highly discriminative features from the observed image patch data for similarity metrics learning. Similarly, to obtain better similarity measure, Simonovsky et al. ([2016](#page53)) used a deep network trained from a few aligned image pairs. In addi-tion, a fast, deformable image registration method called

Quicksilver (Yang et al. [2017b](#page56)) has been devised by the patch-wise prediction of a deformation model directly using image appearance, whereby a deep encoder-decoder network is used for predicting the large deformation diffeomorphic model. Inspired by deep convolution, Revaud et al. ([2016](#page52)) introduced a dense matching algorithm based on a hierarchi-cal correlation architecture. This method can handle complex non-rigid deformations and repetitive textured regions. Arar et al. ([2020](#page44)) introduced an unsupervised multi-modal image registration technique based on an image-to-image transla-tion network with geometric preserving constraints.

Different from metric learning, a trained agent is used for image registration with a reinforcement learning paradigm, and typically for estimating a rigid transformation model or a deformable field. Liao et al. ([2017](#page49)) first used the rein-forcement learning for rigid image registration, in which an artificial agent and a greedy supervised approach coupled with attention-driven hierarchical strategy are used to realize the “strategy learning” process and find the best sequence of motion actions to yield image alignment. An artificial agent, which explores the parametric space of a statistical deforma-tion model by training from a large number of synthetically deformed image pairs, is also trained in Krebs et al. ([2017](#page48)) to cope with deformable registration problem and the diffi-culty in extracting reliable ground-truth deformable fields of real data. Instead of using a single agent, Miao et al. ([2018](#page51)) proposed a multi-agent reinforcement learning paradigm for medical image registration in which the auto-attention mech-anism is used for receptive multiple image regions. However, the reinforcement learning is often used to predict iterative updates of the regression procedure and still consumes large computation in the iterative process.

To reduce the run time and avoid explicitly defining a dissimilarity metric, end-to-end registration in one shot has received increasing attention. Sokooti et al. ([2017](#page53)) first designed deep regression networks to directly learn a dis-placement vector field from a pair of input images. Another method in de Vos et al. ([2017](#page46)) similarly trained a deep network to regress and output the parameters of spatial trans-formation, which can then generate the displacement field to warp the moving image to the target image. However, a similarity metric between image pairs is still required to achieve unsupervised optimization. More recently, a deep learning framework has been introduced in de Vos et al. ([2019](#page46)) for unsupervised affine and deformable image reg-istration. The trained networks can be used to register pairs of unseen images in one shot. Similar methods regarding deep networks as a regressor can directly learn the parame-ter transform model from image pairs, such as Fundamental (Poursaeed et al. [2018](#page52)), Homography (DeTone et al. [2016](#page46)), and non-rigid deformation (Rocco et al. [2017](#page52)).

Many other end-to-end image level learning-based reg-istration methods are presented. Chen et al. ([2019](#page45)) pro-

123

International Journal of Computer Vision

posed end-to-end trainable deep networks to directly predict the dense displacement field for image alignment. Wang and Zhang ([2020](#page55)) introduced DeepFLASH for efficient deformable medical image registration, which is imple-mented in a low dimensional bandlimited space thus dra-matically reduces the computational and memory request. To simultaneously enhance the topology preservation and smoothness of the transformation model, Mok and Chung ([2020](#page51)) proposed an efficient unsupervised symmetric image registration method which maximizes the similarity between images within the space of diffeomorphic maps and estimates both forward and inverse transformations simultaneously. In Truong et al. ([2020](#page54)), the authors introduced a universal net-work for geometric matching, optical flow estimation and semantic corresponding, which can achieve both high accu-racy and robustness by investigating the combined use of global and local correlation layers. See more details in the registration-specific reviews (Ferrante and Paragios [2017](#page47); Haskins et al. [2020](#page48)).

*Stereo matching.* Over the past years, analogous to regis-tration, numerous studies in stereo matching have focused on accurately computing the matching cost by using deep con-volutional techniques and refining the disparity map (Zbontar and LeCun [2015](#page56); Luo et al. [2016](#page50); Zbontar and LeCun [2016](#page56); Shaked and Wolf [2017](#page53)). In addition to the deep descriptors, such as DeepCompare (Zagoruyko and Komodakis [2015](#page56)) and MatchNet (Han et al. [2015](#page47)), etc., Zbontar and LeCun ([2015](#page56)) introduced a deep Siamese network to compute the matching cost, which is trained to predict the similarity between image patches. They further proposed a series of CNNs (Zbontar and LeCun [2016](#page56)) for the binary classifi-cation of pairwise matching and applied these in disparity estimation. Similar to converting the computation of match-ing costs into a multi-label classification problem, Luo et al. ([2016](#page50)) proposed an efficient Siamese network for fast stereo matching. In addition, Shaked and Wolf ([2017](#page53)) improved the performance by computing the matching cost with the proposed constant highway networks and the disparity esti-mation with reflective confidence learning.

The end-to-end deep manner for this matching task has drawn increasing attention in recent years. For instance, Mayer et al. ([2016](#page51)) trained an end-to-end CNN in their Disp-Net to obtain a fine disparity map, which is extended by Pang et al. ([2017](#page52)) with a two-stage CNN called cascade residual learning (CRL). More recently, a spatial pyramid pooling module together with a 3-D convolutional strat-egy has been introduced in Chang and Chen ([2018](#page45)). This approach can exploit global context information to enhance stereo matching. Inspired from CycleGAN (Zhu et al. [2017](#page57)) and to deal with domain gap, Liu et al. ([2020](#page50)) proposed an end-to-end training framework to translate all synthetic stereo images into realistic ones simultaneously maintain epipolar constraints. This method is implemented through

a jointly optimizing between domain translation and stereo matching. Another method in Yang et al. ([2020](#page55)) learns the wavelet coefficients of the disparity rather than the disparity itself, which can learn global context information from low frequency submodule and details from others. Moreover, the guided strategy (Zhang et al. [2019a](#page56); Poggi et al. [2019](#page52)) is also utilized for stereo matching.

Stereo matching with deep convolutional techniques has been dominated for their top performance in public bench-marks[2](#page25). However, the use of CNNs in stereo matching community is limited by the input image pairs, which are generally captured from the binocular camera with a narrow baseline and epipolar rectification. Nevertheless, the network structure, basic ideas, and some tricks or strategies in these learning-based stereo matching may have a strong reference for general image matching tasks.

**4.6.2 Learning from Points**

Learning from points is not as popular as those in images for feature extraction, representation and similarity mea-surements. Point-based learning, particularly for feature matching, has only been introduced in recent years. This is because using CNNs on point data is more difficult than on raw images due to the unordered structure and dispersed nature of sparse points. Moreover, operating and extract-ing the spatial relationships, such as neighboring elements, relative positions, length, and angle information, among multi-points using deep convolutional techniques are chal-lenging. However, using deep learning techniques to solve points-based tasks has received increasing considerations. These techniques can be roughly divided into parameter fit-ting (Brachmann et al. [2017](#page45); Ranftl and Koltun [2018](#page52)) and point classification and/or segmentation (Qi et al. [2017a](#page52), [b](#page52); Moo Yi et al. [2018](#page51); Ma et al. [2019a](#page50); Zhao et al. [2019](#page56)). The for-mer is inspired by the classical RANSAC algorithm and aims to estimate the transformation model, such as fundamen-tal matrix (Ranftl and Koltun [2018](#page52)) and epipolar geometry (Brachmann and Rother [2019](#page45)), by means of a data-driven optimization strategy with CNNs. However, the latter tends to train a classifier to identify the true matches from putative match set. Generally, parameter fitting and point classifica-tion are trained jointly for performance enhancement.

For trainable fundamental matrix estimation, Brachmann et al. ([2017](#page45)) proposed a differentiable RANSAC, termed as DSAC, which is based on reinforcement learning in an end-to-end manner. They replaced the deterministic hypothesis selection by probabilistic selection to decrease the expected loss and optimize the learnable parameters. Subsequently, Ranftl and Koltun ([2018](#page52)) presented a trainable method for

1. [http://www.cvlibs.net/datasets/kitti/eval\_scene\_flow.php?](http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo) [benchmark=stereo](http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo)

123

International Journal of Computer Vision

fundamental matrix estimation from noise, which is casted as a series of weighted homogeneous least-squares prob-lem, where the robust weights are estimated with deep networks. Similar to DSAC, using learning techniques to improve re-sampling strategy is also introduced in Brach-mann and Rother ([2019](#page45)) and Kluger et al. ([2020](#page48)). Brachmann and Rother ([2019](#page45)) proposed NG-RANSAC, a robust estima-tor using learned guidance of hypothesis sampling. It uses the inlier count itself as training objective to facilitate self-supervised learning of NG-RANSAC, and can incorporate non-differentiable task loss functions and non-differentiable minimal solvers. While CONSAC (Kluger et al. [2020](#page48)) is introduced as a robust estimator for multiple parametric model fitting. It uses neural network to sequentially update the conditional sampling probabilities for the hypothesis selection.

Learning-based mismatch removal methods have been developed in recent years. Moo Yi et al. ([2018](#page51)) first attempted to introduce a learning-based technique termed as learning to find good correspondences (LFGC), which aims to train a net-work from a set of sparse putative matches together with the image intrinsics under the rigid geometrical transformation constraints, and to label the test correspondences as inliers or outliers and output the camera motion simultaneously. How-ever, the LFGC may sacrifice many true correspondences to estimate the motion parameters, failing to handle general matching problems, such as deformable and non-rigid image matching. To this end, Ma et al. ([2019a](#page50)) proposed a gen-eral framework to learn a two-class classifier for mismatch removal called LMR, which uses a few images, and hand-crafted geometrical representation for training. Their method showed promising matching performance with linearithmic time complexity. More recently, Zhang et al. ([2019b](#page56)) focused on the geometrical recovery based on their order-aware net-works (OAN) and have achieved promising performance on pose estimation. Sarlin et al. ([2020](#page53)) proposed Super-Glue, to match two sets of local features by jointly finding correspondences and rejecting non-matchable points. This method is implemented with graph neural networks (Scarselli et al. [2009](#page53)) for differentiable transport problem optimiza-tion. Similar graph neural network pipeline has been adopted by an emerging research branch namely deep graph match-ing (Wang et al. [2019](#page55); Yu et al. [2020a](#page56); Fey et al. [2020](#page47)), where cross-graph convolution (Wang et al. [2019](#page55)), channel-independent embedding (Yu et al. [2020a](#page56)) and Spline-based convolution (Fey et al. [2020](#page47)) are proposed and adopted for supervised graph correspondence learning.

Even though applying CNNs onto point data is difficult, the latest techniques have shown great potential for matrix estimation and point data classification with deep regressor and classifier, particularly for the challenging data or scenar-ios. Moreover, the multi-layer perception methods in natural language processing and the graph convolutional techniques

may serve as great references for addressing these dispersed and unstructured point data in the matching task.

**4.7 Matching in 3-D Cases**

Similar to its 2-D counterpart, 3-D matching methods often involve two steps, i.e., namely, keypoint detection and local feature description, and a sparse correspondence set can then be established by calculating the similarities between descriptors. Although most methods use local fea-ture descriptors, which are designed to be robust to noise and deformations to establish correspondences between 3-D instances, a variety of classical and recent works fall into another category. We refer the readers to the recent surveys (Biasotti et al. [2016](#page44); Van Kaick et al. [2011](#page54)) in the shape matching area given that a detailed review of the literature is beyond the scope for this paper.

The embedding methods aim to parametrize the com-plex matching problem with less degrees of freedom for tractability by exploiting some natural assumptions (e.g., approximate isometry). A traditional approach is proposed by Elad and Kimmel ([2003](#page46)) to match shapes by embedding them in an intermediate Euclidean space. In this approach, the geodesic distances are approximated by Euclidean ones, and the original non-rigid registration problem is reduced to rigid registration in the intermediate space. Notably, another work developed conformal mapping approaches that also use embedding space (Lipman and Funkhouser [2009](#page49); Kim et al. [2011](#page48); Zeng et al. [2010](#page56)).

A more direct approach is to find a point-wise match-ing between (subsets of) points on shapes by minimizing the structure distortion. This formulation was developed by Bronstein et al. ([2006](#page45)), who introduced a highly non-convex and non-differentiable objective and generalized multidimensional scaling technique for optimization. Some researchers have also attempted to mitigate the prohibitively high computational complexity issue (Sahillioglu and Yemez [2011](#page53); Tevs et al. [2011](#page54)) while considering the quadratic assignment formulation (Rodola et al. [2012](#page52), [2013](#page52); Chen and Koltun [2015](#page45); Wang et al. [2011](#page55)) in graph matching.

The family of methods based on the functional map frame-work was first developed by Ovsjanikov et al. ([2012](#page52)). Instead of point-to-point matching in Euclidean space, these meth-ods represent the correspondences using the functional map between two manifolds, which can be characterized by linear operators. The functional map can be encoded in a compact form by using the eigenbases of the Laplace-Beltrami oper-ator. Most natural constraints on the map, such as landmark correspondences and operator commutativity, become lin-ear in this formulation, leading to an efficient solution. This approach was adopted and extended in many follow-up works (Aflalo et al. [2016](#page44); Kovnatsky et al. [2015](#page48); Pokrass et al. [2013](#page52); Rodolà et al. [2017](#page52); Litany et al. [2017](#page49)).

123

International Journal of Computer Vision

Point set learning in 3-D cases for registration is also a hot topic. Yew et al. ([2020](#page56)) proposed RPM-Net for rigid point cloud registration, in which it desensitizes initializa-tion and improves convergence performance with learned fusion features. Gojcic et al. ([2020](#page47)) introduced an end-to-end multiview point cloud registration framework by directly learning to register all views of a scene in a globally consistent manner. Pais et al. ([2020](#page52)) introduced a learning architecture for 3D point registration, namely 3DRegNet. This method can identify true point correspondences from a set of puta-tive matches, and regress the motion parameters to align the scans into a common reference frame. Choy et al. ([2020](#page46)) used high-dimensional convolutional networks to detect lin-ear subspaces in high-dimensional spaces, then applied it for 3D registration under rigid motions and image correspon-dence estimation.

**4.8 Summary**

Given a pair of images of similar object/scene and with/without the feature detection and/or description, the matching tasks have been extended into several different forms, such as image registration, stereo matching, feature matching, graph matching, and point set registration. These different matching definitions are generally introduced for specific applications, with their own strengths presented.

Traditional image registration and stereo achieve dense matching by means of patch-wise similarity measuring together with optimization strategy to search the overall opti-mal solution. However, they are conducted on image pairs of high overlapping area (slight geometrical deformation) and binocular camera, and these may require large computational burden and the limited handcrafted measuring metrics.

The introduction of deep learning has promoted registra-tion accuracy and disparity estimation due to advancements in network design and loss definition, as well as abundant training samples. However, we also find that using deep learning for these matching tasks is usually performed on image pairs undergoing slight geometrical deformation such as medical image registration and binocular stereo match-ing. Applying them for more complex scenarios, such as wide baseline images stereo or image registration with seri-ous geometric deformations, still remains open.

Feature-based matching can effectively address the limita-tions in large viewpoint, wide baseline, and serious non-rigid image matching problems. Among those proposed in the literature, the most popular strategy is to construct the puta-tive matches based on descriptor distance, followed by a robust estimator such as RANSAC. However, a large num-ber of mismatches in the putative match sets may negatively affect the performance in subsequent visual task and also require considerable time for model estimation. Therefore, the mismatch removal method is required and integrated to

preserve as many true matches as possible while maintain-ing the mismatch to a minimum level using extra geometrical constraints. Specifically, the resampling-based method, such as RANSAC, can estimate the latent parameter model and simultaneously remove the outliers. However, their theoreti-cally required runtime grows exponentially with the increase in outlier rate, and they cannot process the image pairs that undergo more complex non-rigid transformations. The non-parametric model-based methods can handle the non-rigid image matching problem by using high-dimensional non-parametric model, but it is still challenging in defining the objective function and finding the optimal solution in a more complex solution space. Different from the global constraints in the resampling and non-parameter model-based methods, the relaxed mismatch removal methods are commonly con-ducted on a local coherent assumption of potential inliers. Thus, much simpler but efficient rules are designed to fil-ter out the outliers while maintaining the inliers within an extremely short time. However, methods of this type are lim-ited due to their parameter sensitivity; moreover, they are prone to preserve evident outliers, thereby affecting the accu-racy of subsequent pose estimation and image registration.

In addition, the image patch-based descriptor may not be workable due to the matching request in less-texture images, shape, semantic images, and the raw points directly captured from specific device. Therefore, for performing the matching task of these situations, the graph matching and point regis-tration methods are more suitable. The graph structure among neighboring points and the overall corresponding matrix are applied to optimize and find the optimal solution. However, these pure point-based methods are limited by restrictions in their computation burden and outlier sensitivity. Therefore, designing appropriate problem formulation and constraint conditions, and proposing more efficient optimization meth-ods, are still open problems in image matching community and require further research attention.

Analogously to image-based learning, increasing studies have used deep learning in feature-based matching commu-nity. The latest techniques have shown great potential for matrix estimation (e.g. fundamental matrix) and point data classification (such as mismatch removal) with deep regres-sor and classifier, particularly for handling challenging data or scenarios. However, conducting convolutional networks on point data is not as easy as on raw images due to the unordered structure and dispersed nature of these sparse points. Nevertheless, recent studies have shown the feasibil-ity of using the graph convolutional strategy and multi-layer perception methods, together with specific normalization on such point data. In addition to rigid transformation parameter estimation, matching on point data with non-rigid and even serious deformation by using deep convolutional techniques may be a more challenging and significant problem.

123

International Journal of Computer Vision

**5 Matching-Based Applications**

Image matching is a fundamental problem in computer vision and is considered a critical prerequisite in a wide range of applications. In this section, we briefly review several repre-sentative applications.

**5.1 Structure-from-Motion**

Structure-from-motion (SfM) involves recovering the 3-D structure of a stationary scene from a series of images, which are obtained from different viewpoints by estimat-ing the camera motions corresponding to these images. SfM involves three main stages, namely, (i) feature matching across images, (ii) camera pose estimation, and (iii) recovery of the 3-D structure using the estimated motion and features. Its efficacy largely depends on the admissible set of feature matches.

In modern SfM systems (Schonberger and Frahm [2016](#page53); Wu [2018](#page55); Sweeney et al. [2015](#page54)), the feature matching pipeline is widely adopted across images, i.e., feature detection, description, and nearest-neighbor matching, to provide ini-tial correspondences. The initial correspondences contain a number of outliers. Thus, geometric verification is required, which is tackled via the estimation fundamental matrix using RANSAC (Fischler and Bolles [1981](#page47)). This can potentially be addressed by mismatch removal methods.

Meanwhile, to enhance the SfM task, researchers have focused on performing robust feature matching, i.e., thus establishing rich and accurate correspondences. Evidently, advanced descriptors can greatly affect this task (Fan et al. [2019](#page46)). Moreover, Shah et al. ([2015](#page53)) proposed a geometry-aware approach, which initially uses a small sample of features to estimate the epipolar geometry between the images and leverages it for the guided matching of the remaining features. Lin et al. ([2016b](#page49)) utilized RANSAC to guide the training of match consistency curves for differ-entiating true and false matches. Their approach traces the common problems of wide-baselines and repeated structures for reconstructing modern cities. These correspondences are also the prerequisites for camera pose estimation, and the effective substitution of commonly used RANSAC for this task has also been investigated (Moo Yi et al. [2018](#page51)), with a pre-stage of identifying good correspondences.

**5.2 Simultaneous Localization and Mapping**

Acquiring maps of the environment is a fundamental task for autonomous mobile robots, thereby forming the basis of many different higher-level tasks, such as navigation and localization. The problem of simultaneous localization and mapping (SLAM) (Davison et al. [2007](#page46); Mur-Artal et al. [2015](#page51);

Sturm et al. [2012](#page53)) has received intensive attention over the decades.

In common SLAM systems, feature matching is needed to establish correspondences between frames, which then serve as the input for estimating the relative camera pose and local-ization. Similar to SfM, the full-fledged feature matching pipeline is used in most SLAM systems. Typically, in Endres et al. ([2012](#page46)), Endres et al. introduced a SLAM system that incorporates feature matching to establish spatial relations from the sensor data in the front-end. The well-known SIFT (Lowe [2004](#page50)), SURF (Bay et al. [2008](#page44)), and ORB (Rublee et al. [2011](#page53)) algorithms are optionally used to detect and describe features, and RANSAC (Fischler and Bolles [1981](#page47)) is subse-quently used for robust matching.

An evaluation of different feature detectors and descrip-tors can be found in Gil et al. ([2010](#page47)). Recently, Lowry and Andreasson ([2018](#page50)) proposed a spatial verification method for visual localization, which is robust in the presence of a high proportion of outliers. For a SLAM system that percepts 3-D range scans, the point set registration methods (e.g. ICP) (Nüchter et al. [2007](#page51)) are also used for scan matching and localizing the robot.

Loop closure detection–another core module in SLAM application–refers to accurately asserting that an agent has returned to a previously visited location. It is crucial to reduce the drift of the estimated trajectory caused by accumula-tive error. A group of appearance-based approaches have been developed to use image similarities to identify previ-ously visited places. Feature matching results are naturally applicable to measure the similarity of two scenes and have been the bases of many state-of-the-art methods. For exam-ple, Liu and Zhang ([2012](#page50)) performed feature matching with SIFT between the current image and each previously vis-ited image, after which they determined the closed loop on the basis of the number of accurate matches in the results. Zhang et al. ([2011](#page56)) used directed matching of raw features extracted from images for detecting loop-closure events. To achieve loop closure detection, Wu et al. ([2014](#page55)) used LSH as the basic technique by matching the binary visual features in the current view of a robot with the visual features in the robot appearance map. Liu et al. ([2015a](#page50)) developed a consen-sus constraint to prune outliers and verified the superiority of their methods for loop closure detection.

**5.3 Visual Homing**

Visual homing aims to navigate a robot from an arbitrary starting position to a goal or home position based solely on visual information. This is often accomplished by esti-mating a homing vector/direction (pointing from the current position to the home position) from two panoramic images, which are captured respectively at the current position and the home position. Conventionally, feature matching serves

123

International Journal of Computer Vision

as the building block of correspondence methods in visual homing research (Möller et al. [2010](#page51)). In this category, the homing vector can be determined by transforming the cor-respondences into motion flows (Ma et al. [2018b](#page50); Churchill and Vardy [2013](#page46); Liu et al. [2013](#page50); Zhao and Ma [2017](#page56)).

Ramisa et al. ([2011](#page52)) combined the average landmark vec-tor with invariant feature points automatically detected in panoramic images to achieve autonomous visual homing. However, the feature matches are solely determined by the similarity of the descriptors in the method, thus leading to a number of mismatches. The presence of outliers has been ver-ified to be the reason of performance degradation for visual homing (Schroeter and Newman [2008](#page53)). In order to resolve the degradation caused by mismatches, Liu et al. ([2013](#page50)) used a RANSAC-like method to remove mismatches. Meanwhile, Zhao and Ma ([2017](#page56)) proposed a visual homing method by simultaneously mismatch removal and robust interpolation of sparse motion flows under a smoothness prior. Ma et al. ([2018b](#page50)) also proposed a guided locality preserving matching method to handle extremely large proportions of outliers and improve the visual homing robustness.

**5.4 Image Registration and Stitching**

Image registration is the process of aligning two or more images of the same scene obtained from different view-points, at different times, or from different sensors (Zitova and Flusser [2003](#page57)). In the past decades, feature-based methods in which the key requirement is feature matching have gained increasing attention due to its robustness and efficiency. Once the correspondence is established, image registration is reduced to estimate the transformation model (e.g., rigid, affine, or projective). Finally, the source image is transformed by means of the mapping functions, which rely on some interpolation technique (e.g., bilinear and nearest neighbor). A large number of works have been proposed for feature matching and image registration. Ma et al. ([2015b](#page51)) pro-posed a Bayesian formulation for rigid and non-rigid feature matching and image registration. To further exploit the geo-metrical cues, the locally linear transforming constraint is incorporated. They also recently proposed a guided local-ity preserving matching method (Ma et al. [2018a](#page50)). Their proposed method can significantly reduce the computational complexity and is able to deal with a more complex transfor-mation model. For non-rigid image registration, Pilet et al. ([2008](#page52)) and Gay-Bellile et al. ([2008](#page47)) proposed solutions, where robust matching techniques are insensitive to outliers. Some efforts (Paul and Pati [2016](#page52); Ma et al. [2017b](#page51); Yang et al. [2017a](#page56)) also attempted to modify feature detectors and descriptors to improve the registration process.

The problem of multi-modal image registration is more complicated due to the high variability of appearance caused by different modalities, which frequently arise in medical

image and multi-sensor image analysis. For example, Chen et al. ([2010](#page45)) developed the partial intensity invariant feature descriptor (PIIFD) to register retinal images, whereas Wang et al. ([2015](#page55)) extended PIIFD in a more robust registration framework with SURF detector (Bay et al. [2008](#page44)) and a single Gaussian point matching model. On the basis of the charac-teristics of multi-modal images, Liu et al. ([2018a](#page50)) proposed an affine and contrast invariant descriptor for IR and visible image registration. Du et al. ([2018](#page46)) also proposed an IR and visible image registration method based on scale-invariant PIIFD feature and locality preserving matching. Ye et al. ([2017](#page56)) proposed a novel feature descriptor based on the struc-tural properties of images for multi-modal registration. A detailed discussion of feature matching-based, multi-modal registration techniques of the medical image analysis area, which are categorized as geometric methods, can be found in Sotiras et al. ([2013](#page53)).

Meanwhile, image stitching or image mosaic involves obtaining a wider field-of-view of a scene from a sequence of partial views (Ghosh and Kaabouch [2016](#page47)). Compared to image registration, image stitching deals with low overlap-ping images and requires accurate alignment in the pixel-level to avoid visual discontinuities. Feature-based stitching methods are popular in this area because of their invariance properties and efficiency. For example, in order to iden-tify geometrically consistent feature matches and achieve accurate homography estimation, Brown and Lowe ([2007](#page45)) proposed the use of the SIFT (Lowe [2004](#page50)) feature matching and the RANSAC (Fischler and Bolles [1981](#page47)) algorithm. Lin et al. ([2011](#page49)) used SIFT (Lowe [2004](#page50)) to pre-compute matches and then jointly estimating the matching and the smoothly varying affine fields for better stitching performance. Inter-ested readers can refer to the comprehensive survey (Ghosh and Kaabouch [2016](#page47); Bonny and Uddin [2016](#page45)) for an overview of more feature-based image mosaic and stitching methods.

**5.5 Image Fusion**

To generate a more conducive image to subsequent appli-cations, image fusion is adopted to combine the meaningful information from images acquired by different sensors or under different shooting settings (Pohl and Van Genderen [1998](#page52)), wherein the source images have been accurately aligned in advance. The very premise of image fusion is to register source images using feature matching methods, and the accuracy of registration directly affects the fusion quality. Liu et al. ([2017](#page50)) used the CNN to jointly generate the activ-ity level measurement and fusion rules for multi-focus image fusion. Meanwhile, Ma et al. ([2019c](#page51)) proposed an end-to-end model for infrared and visible image fusion, which generates images with a dominant infrared intensity and an additional visible gradient under the framework of generative adversar-ial networks. Subsequently, they introduced a detail loss and

123

International Journal of Computer Vision

a target edge-enhancement loss to further enrich the texture details (Ma et al. [2020](#page50)).

A group of methods aim to fuse images based on the local features, among which the dense SIFT is the most popu-lar. Liu et al. ([2015b](#page50)) proposed the fusion of multi-focus images with dense scale invariant feature transform, wherein the local feature descriptors are used not only as the activity level measurement, but also to match the mis-registered pix-els between multiple source images to improve the quality of the fusion results. Similarly, Hayat and Imran ([2019](#page48)) pro-posed a ghost-free multi-exposure image fusion technique using the dense SIFT descriptor with a guided filter, which can produce high-quality images using ordinary cameras. In addition, Chen et al. ([2015](#page45)) and Ma et al. ([2016a](#page50)) introduced a method that can perform image registration and image fusion simultaneously, thus fulfilling image fusion on unaligned image pairs.

**5.6 Image Retrieval, Object Recognition and Tracking**

Feature matching can be used to measure similarity between images, thereby enabling a series of high-level applications, including image retrieval (Zhou et al. [2017](#page56)), object recogni-tion, and tracking. The goal of image retrieval is to retrieve all images that exhibit similar scenes for a given query image. In local feature-based image retrieval, the image similarity is intrinsically determined by the feature matches between images. Thus, the image similarity score can be obtained by aggregating votes from the matched features. In Zhou et al. ([2011](#page57)), the relevance score is simply determined by the number of feature matches across two images. In Jégou et al. ([2010](#page48)), the scoring function is defined as a cumulation of the squared term frequency inverse document frequency weights on shared visual words, which is essentially a bag of features of inner products.

Moreover, geometric context verification, a common tech-nique for refining initial image retrieval result, is directly related to feature matching. By incorporating the geometri-cal information, geometric context verification technique can be used to address the false match problem caused by the ambiguity of local descriptor and the quantization loss. For image retrieval, a large group of methods estimate the trans-formation model in an explicit approach to verify the tentative matches. For example, Philbin et al. ([2007](#page52)) used a RANSAC-like method to find the inlier correspondences, whereas Avrithis and Tolias ([2014](#page44)) developed a simple spatial match-ing model inspired by Hough voting in the transformation space. Another line of works address geometric context ver-ification without explicitly handling a transformation model. For example, Sivic and Zisserman ([2003](#page53)) utilized the con-sistency of spatial context in local feature groups to verify the tentative correspondences. Zhou et al. ([2010](#page57)) proposed

the spatial coding method, whereby the valid visual word matches are identified by verifying the global relative posi-tion consistency.

With the function of measuring similarity, feature match-ing also plays an important role in object recognition and tracking. For example, Lowe et al. ([1999](#page50)) used SIFT features to match sample images and new images. In their proposed method, the potential model pose is identified through a Hough transform hash table and then through a least-squares fit to achieve a final estimate of model parameters. The pres-ence of the object is strongly evident if at least three keys agree on the model parameters with low residuals. Modern attempts for object recognition also include some specifically handcrafted features (Dalal and Triggs [2005](#page46); Hinterstoisser et al. [2012](#page48)) and, more recently, deep learning approaches (Wohlhart and Lepetit [2015](#page55)).

Tracking basically refers to estimating the trajectory of an object over images. Feature matching across images is the basis of feature-based tracking, and a variety of algo-rithms for these tasks have been proposed in the literature. The feature matching pipeline is adopted in most visual tracking systems, except that the matching is constrained to those of the known features that are predicted to lie close to the encountered position. The readers are referred to a comprehensive evaluation of different feature detectors and descriptors for tracking by Gauglitz et al. ([2011](#page47)), and the recently presented benchmark (Wu et al. [2015b](#page55)), which cov-ers a review of modern object tracking methods as well as the role played by feature representation methods.

**6 Experiments**

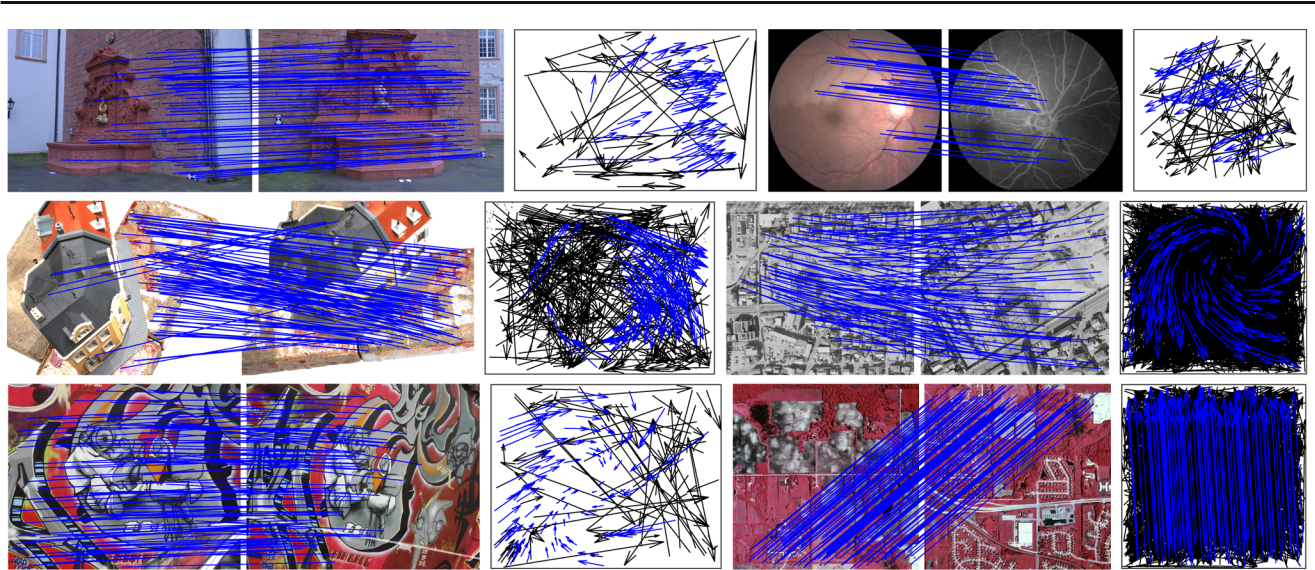
Diverse methods for image matching have been proposed, particularly when the deep learning techniques are becoming increasingly popular. However, the question of which method would be suitable for specific applications under different scenarios and requirements still remains. We are encouraged to conduct more comprehensive and objective comparative analysis of these classical and state-of-the-art techniques.

**6.1 Overview of Existing Reviews**

To evaluate the existing matching methods at an early time, the classical image registration survey (Zitova and Flusser [2003](#page57)) provided several definitions for evaluation of regis-tration accuracy including localization error, matching error, and alignment or registration error. In 2005, Mikolajczyk et al. evaluated affine region detectors (Mikolajczyk et al. [2005](#page51)) and local descriptors (Mikolajczyk and Schmid [2005](#page51)) against changes of viewpoint, scale, illumination, blur, and image compression on their own proposed VGG (a.k.a. Oxford) datasets. They also presented a comprehensive compari-

123

International Journal of Computer Vision



**Fig. 2** Examples of the five datasets. The ground truth is given usingcolored correspondences. The head and tail of each arrow in the motion field correspond to the positions of feature points in two images (blue

* true positive, black = true negative). For visibility, in the image pairs, at most 100 randomly selected matches are presented, and the true neg-atives are not shown (Color figure online)

son on *repeatability* and *accuracy* for detectors and *recall*, 1 − *pr eci si on* for descriptors. Subsequently, Strecha et al. ([2008](#page53)) published a dense 3-D dataset for wide-baseline stereo and 3-D geometrical and camera pose evaluation.

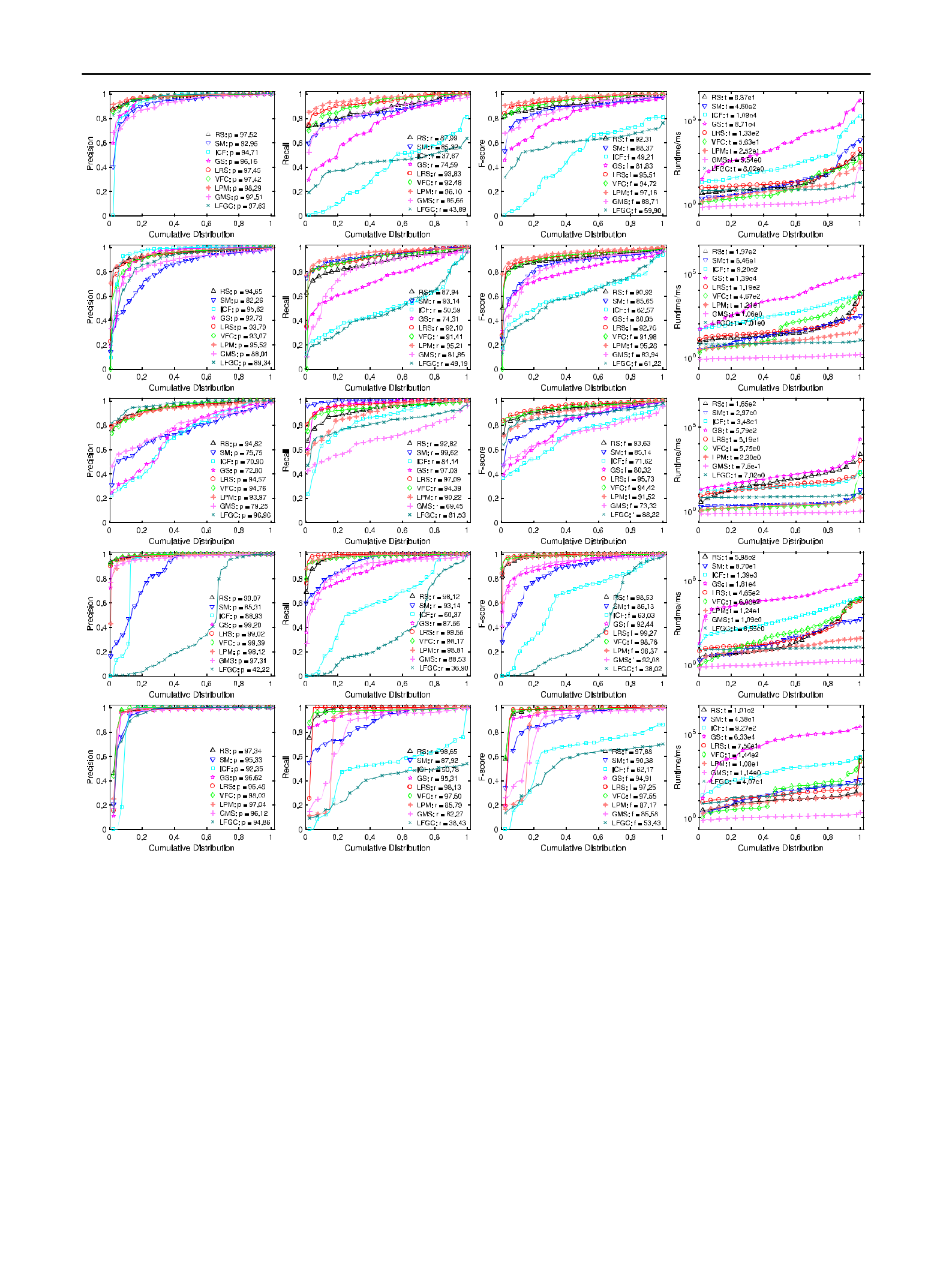
In addition, Aanæs et al. ([2012](#page43)) evaluated some repre-sentative detectors using a large dataset of known camera positions, controlled illumination, and 3-D models, namely, DTU. At the same time, Heinly et al. ([2012](#page48)) compared the traditional float and binary feature operators in 2012 and eval-uated their matching performance with the inter-combination of existing detectors and descriptors on the public and their own datasets. The evaluation was conducted on more system-atic performance metrics consisting of *putative match ratio*, *precision*, *matching score*, *recall*, and *entropy*. Similarly,using inter-combination strategy, Mukherjee et al. ([2015](#page51)) provided a comparative experimental analysis for selecting appropriate combination of various detectors and descrip-tors in order to solve the problems of image matching using different image data.

More recently, inspired by emerging deep learning tech-niques, Balntas et al. ([2017](#page44)) reported that existing defective datasets and evaluation metrics may lead to unreliable com-parative results. Thus, they proposed and publicized a large benchmark for handcrafted and learned local image descrip-tors called Hpathes. They also comprehensively evaluated the performance of widely used handcrafted descriptors and recent deep ones with extensive experiments on patch recognition, patch verification, image matching, and patch retrieval. Schonberger et al. ([2017](#page53)) conducted an experimen-tal evaluation of learned local features, including classical machine learning based variants of SIFT and recent CNN-

based techniques, in which they considered that finding additional true matches between similar images does not necessarily improve performance when matching images under extreme viewpoint or illumination changes. Mitra et al. ([2018](#page51)) provided a PhotoSynth (PS) dataset for training local image descriptors. Komorowski et al. ([2018](#page48)) provided a sta-bility evaluation for handcrafted and learning-based interest point detectors on ApolloScape street dataset (Huang et al. [2018](#page48)). A comprehensive comparison of local image feature detectors based on both classical and CNN techniques is con-ducted on public datasets (Lenc and Vedaldi [2014](#page49)). That work proposed a modified repeatability for detection eval-uation, which is more robust to feature scale variety. Jin et al. ([2020](#page48)) introduced a benchmark for local features and robust estimation algorithms, focusing on the accuracy of the reconstructed camera pose as their practical evaluation. In addition, Bellavia and Colombo ([2020](#page44)) provided a compre-hensive analysis and evaluation about the descriptor design based on SIFT.

From the above mentioned, we can know that several com-prehensive and thorough evaluation of feature detectors and descriptors can be found in Komorowski et al. ([2018](#page48)), Lenc and Vedaldi ([2014](#page49)), Heinly et al. ([2012](#page48)) and Schonberger et al. ([2017](#page53)). However, in order to evaluate the local feature methods, many studies compared the matching performances on a 3-D reconstruction task, including the works of Fan et al. ([2019](#page46)) and Schonberger et al. ([2017](#page53)). In the 3-D case, Tombari et al. ([2013](#page54)) presented a thorough evaluation of sev-eral state-of-the-art 3-D keypoint detectors, and Guo et al. ([2016](#page47)) compared ten popular local feature descriptors in the contexts of 3-D object recognition, 3-D shape retrieval, and

123

International Journal of Computer Vision

**Fig. 3** Quantitative performance of the state-of-the-art mismatchremoval algorithms on the introduced five datasets. The statistics of precision, recall, F-score and runtime are reported for each dataset, and the average values are given in the legend. From top to bottom, the statis-tics of DAISY, DTU, Retina, RemoteSensing and VGG. The results are

presented in cumulative distribution, a point on the curve with coordi-nate (x, y) denotes that there are *(*100 ∗ *x)* percents of image pairs which have the performance value (i.e., precision, recall, F-score or runtime) no more than y

3-D modeling. Several matching related applications, such as image retrieval (Zheng et al. [2018](#page56)) and visual localization (Piasco et al. [2018](#page52)), have also been evaluated recently. We refer the readers to these works for a detailed discussion of their performance. For mismatch removal, point set regis-tration, graph matching, and the application performance of

pose estimation and loop-closure detection, we will present both quantitative and qualitative comparisons.

**6.2 Results on Mismatch Removal**

We conduct experiments on five image matching datasets with ground truth. Our primary aim is to evaluate different

123

International Journal of Computer Vision

mismatch removal methods. The features of each image are assumed to be detected and described, and the open source VLFeat toolbox is used to determine the putative correspon-dence using SIFT (Lowe [2004](#page50)). The details of the adopted datasets are described as follows, and some representative image matching examples from the used datasets are illus-trated in Fig. [2](#page31). The ground truth of each dataset is checked by the provided geometrical transform matrix, such as homo-graph, or provided in the manner that each match is manually labeled as true or false. The experiments of this part are per-formed on a desktop with 3.4 GHz Intel Core i5-7500 CPU, 8GB memory.

*DAISY* (Tola et al.[2010](#page54)):The dataset consists of widebaseline image pairs with ground truth depth maps, includ-ing two short image sequences and several individual image pairs. We match each two images in one sequence and all the individual pairs are used, which creates in total 47 image pairs for evaluation. This dataset is a challenging one due to the large number of matches, which is up to 8000. The average numbers of matches and inlier rate are 1191.6 and 77*.*99%, respectively.

*DTU* (Aanæs et al.[2016](#page43)):The dataset is originally desig-nated for multiple-view stereo evaluation, which involves a number of different scenes with a wide range of objects. The ground truth camera positions and internal camera param-eters have high accuracy. Two scenes are selected for this dataset (i.e., Frustum and House), after which we create 130 image pairs for evaluation. These scenes generally have large viewpoint changes in the scenes. The average numbers of matches and inlier rate are 729*.*3 and 58*.*83%, respectively.

*Retina* (Ma et al.[2019d](#page51))It consists of 70 retinal imagepairs with non-rigid transformation. Due to different modal-ities between images, ambiguous putative matches are gen-erated, resulting in a small number of correct matches and a low inlier ratio. The average numbers of matches and inlier rate are 158*.*4 and 41*.*56%, respectively.

*RemoteSensing* (Ma et al.[2019d](#page51))There are 161 remotesensing image pairs including color-infrared, SAR, and panchromatic photographs. The feature matching task for such image pairs typically arises in image-based position-ing as well as navigating and change detection. The average numbers of matches and inlier rate are 767*.*6 and 68*.*50%, respectively.

*VGG* (Mikolajczyk and Schmid[2005](#page51))It contains 40image pairs either of planar scenes or captured by a cam-era in a fixed position during acquisition. Hence, the image transformation can be precisely described by homography. The ground truth homographies are included in the dataset.

These abovementioned datasets are collected and avail-able at. [3](#page33) In addition, a small UAV image registration dataset (SUIRD) is also provided for image registration or match-

1. <https://github.com/StaRainJ/Imgae_matching_Datasets>.

ing research. This dataset includes 60 pairs of low-altitude remote sensing images captured by small UAV and their groundtruth. The image pairs contain viewpoint changes in horizontal, vertical, their mixture and extreme patterns which produce problems of low overlap, image distortion and severe outliers.[4](#page33) Throughout the experiments, we use three evalua-tion metrics: precision, recall, and F-score. Given the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the precision is obtained by:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *P* = | *T P* | | | | | (5) |  |
|  |  |  | *.* | |  |
| *T P* | + | *F P* |  |
|  |  |  |  |  |  |  |
| The recall is given as follows: | | | | | |  |  |
| *R* = | *T P* | | | | | (6) |  |
|  |  |  | | *.* |  |
| *T P* | + | *F N* | |  |
|  |  |  |  |  |  |  |

The F-score, as a summary statistic of precision and recall, is obtained as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *F* | = |  | 2×*P*×*R* | | | *.* | (7) |  |
|  | | |  |
|  |  | *P* | + | *R* | |  |  |
|  |  |  |  |  |  |  |  |

The mismatch removal methods include: RANSAC (Fis-chler and Bolles [1981](#page47)) (abbreviated as RS), SM (Leordeanu and Hebert [2005](#page49)), ICF (Li and Hu [2010](#page49)), GS (Liu and Yan [2010](#page50)), LO-RANSAC (Lebeda et al. [2012](#page49)) (abbreviated as LRS), VFC (Ma et al. [2014](#page51)), LPM (Ma et al. [2019d](#page51)), GMS (Bian et al. [2017](#page44)), and LFGC (Moo Yi et al. [2018](#page51)).

Figure [3](#page32) shows the performance on the five datasets evaluated by precision, recall, F-score, and runtime with cumulative distribution. In addition, the average values of each statistic is summarized in Table [1](#page34) for a more straightfor-ward comparison. The graph matching methods, SM and GS, have shown relatively weak performances given the graphi-cal model, albeit with strong generality, only excavates the shallow pairwise geometric constraints. Random sampling methods, RS and LRS, hold the key assumption that the image pairs are related by parametric models. This assump-tion seems to work well in the datasets; however, their time costs are not favorable. The non-parametric interpolation method VFC is relatively robust and outperforms ICF. How-ever, its computational cost is higher than that of some other strong competitors, e.g., LPM. LPM is simple to implement. It utilizes a more relaxed geometric constraint, yet it achieves surprisingly excellent performance and becomes the best per-former considering the time cost. Compared with GMS, it obtains much better performance with only a slight increase in runtime. The recent trend has suggested a deep learning paradigm for differentiating mismatches, e.g. LFGC. LFGC has proven to be much more effective than the traditional

1. <https://github.com/yyangynu/SUIRD/tree/master/SUIRD_v2.2>.

123

|  |
| --- |
| 123 |

**Table 1** Quantitative performance of the state-of-the-art mismatch removal algorithms on the introduced five datasets

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alg. |  | RS (Fischler and | SM (Leordeanu | ICF (Li and Hu | GS (Liu and Yan | LRS (Lebeda | VFC (Ma et al. | LPM (Ma et al. | GMS (Bian et al. | LFGC (Moo Yi |
|  |  | Bolles 1981) | and Hebert | 2010) | 2010) | et al. 2012) | 2014) | 2019d) | 2017) | et al. 2018) |
|  |  |  | 2005) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| DAISY | **P** | 97.52 | 92.95 | 94.71 | 96.16 | 97.45 | 97.42 | 98.29 | 92.51 | 97.63 |
|  | **R** | 87.99 | 85.32 | 37.67 | 74.59 | 93.83 | 92.48 | 96.10 | 85.65 | 43.89 |
|  | **F** | 92.31 | 88.37 | 49.21 | 81.83 | 95.51 | 94.72 | 97.16 | 88.71 | 59.90 |
|  | **T** | 8.37e1 | 4.60e2 | 1.09e4 | 8.71e4 | 1.33e2 | 5.63e1 | 2.52e1 | 5.54e0 | 8.02e0 |
| DTU | **P** | 94.65 | 82.26 | 95.62 | 92.73 | 93.79 | 93.07 | 95.52 | 88.01 | 89.34 |
|  | **R** | 87.94 | 93.14 | 50.59 | 74.31 | 92.10 | 91.41 | 95.21 | 81.85 | 49.19 |
|  | **F** | 90.92 | 85.65 | 62.57 | 80.05 | 92.76 | 91.98 | 95.28 | 83.94 | 61.22 |
|  | **T** | 1.92e2 | 5.46e1 | 9.20e2 | 1.39e4 | 1.19e2 | 4.67e2 | 1.21e1 | 1.06e0 | 7.01e0 |
| Retina **P** | | 94.82 | 75.75 | 70.90 | 72.80 | 94.57 | 94.76 | 93.97 | 79.25 | 96.96 |
|  | **R** | 92.82 | 99.62 | 84.14 | 97.03 | 97.09 | 94.39 | 90.22 | 69.45 | 81.53 |
|  | **F** | 93.63 | 85.14 | 71.62 | 80.32 | 95.73 | 94.42 | 91.52 | 73.32 | 88.22 |
|  | **T** | 1.65e2 | 2.97e0 | 3.48e1 | 5.79e2 | 5.19e1 | 5.75e0 | 2.30e0 | 7.5e-1 | 7.02e0 |
| RS | **P** | 99.07 | 85.31 | 88.93 | 99.20 | 99.02 | 99.39 | 98.12 | 97.31 | 42.22 |
|  | **R** | 98.12 | 93.14 | 60.37 | 87.56 | 99.55 | 98.17 | 98.81 | 88.53 | 36.90 |
|  | **F** | 98.53 | 86.13 | 63.03 | 92.44 | 99.27 | 98.76 | 98.37 | 92.08 | 38.02 |
|  | **T** | 5.98e2 | 8.70e1 | 1.39e3 | 1.81e4 | 4.65e2 | 6.08e2 | 1.24e1 | 1.09e0 | 8.55e0 |
| VGG | **P** | 97.34 | 95.33 | 92.35 | 96.62 | 96.48 | 98.03 | 97.04 | 96.12 | 94.86 |
|  | **R** | 98.65 | 87.92 | 50.78 | 95.31 | 98.13 | 97.50 | 85.79 | 82.27 | 38.43 |
|  | **F** | 97.88 | 90.38 | 62.17 | 94.91 | 97.25 | 97.55 | 87.17 | 85.58 | 53.43 |
|  | **T** | 1.01e2 | 4.38e1 | 9.27e2 | 6.33e4 | 7.56e1 | 1.44e2 | 1.08e1 | 1.14e0 | 4.07e1 |

The average statistics of precision (**P**), recall (**R**), F-score (**F**) in percentage and runtime (**T**) in milliseconds with scientific notation are reported for each dataset. The *RemoteSensing* dataset is abbreviated as RS

|  |
| --- |
| International Journal of Computer Vision |

International Journal of Computer Vision



**Fig. 4** 2-D shape contours used in our experiments, from left to right, *fish*, *whale*, *fu*, *beijing*

**Fig. 5** The*bunny*and*wolf*pattern of 3-D point cloud used in ourexperiments

methods. However, in our case, it has a restricted performance with low recall and high accuracy, resulting in the failure in RemoteSensing. This finding indicates that the learning methods are data-dependent with limited generality.

**6.3 Results on Point Set Registration**

The experiments for point set registration consist of two parts: non-rigid registration with 2-D shape contour data and rigid registration with 3-D point cloud data. In the 2-D case, six representative methods, namely, TPS-RPM (Chui and Rangarajan [2003](#page46)), GMM (Jian and Vemuri [2011](#page48)), CPD (Myronenko and Song [2010](#page51)), *L*2 *E* (Ma et al. [2013b](#page50)), PR-GLS (Ma et al. [2015a](#page51)), and APM (Lian et al. [2017](#page49)) are evaluated. In the 3-D case, the rigid versions of GMM and CPD as well as ICP (Besl and McKay [1992](#page44)) and GoICP (Yang et al. [2016](#page56)) are evaluated. The experiments of this part are performed on a desktop with 3.4 GHz Intel Core i5-7500 CPU, 8GB memory.

The point data are normalized as inputs, thus allowing the use of a fixed threshold to evaluate the registration per-formance. Specifically, a point is accurately aligned if its

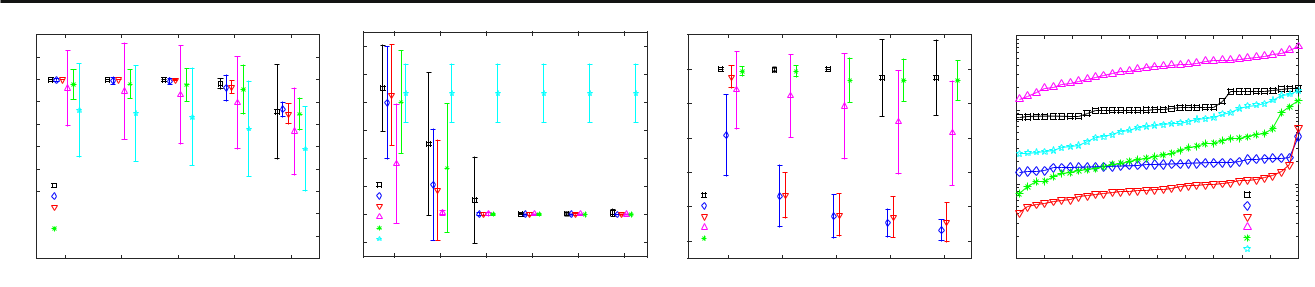
distance to the ground truth corresponding point is below a given threshold. Thus, we can define the accuracy of reg-istration as the percentage of accurately aligned points. In our experiment, the threshold is empirically set to 0*.*1. Four patterns are collected to evaluate the non-rigid 2-D registra-tion results, as shown in Fig. [4](#page35). We also create five deformed shapes for each pattern as the data to be registered, generating a total of 20 instances. We also conduct noise, outlier, and rotation experiments on these instances. For the 3-D case, as shown in Fig. [5](#page35), two patterns are used, and we exert random rotation to create 20 instances for each pattern. Noise and outlier experiments are also conducted on these 40 instances.

The results of non-rigid 2-D registration are presented in Fig. [6](#page36). The outlier experiments of APM are excluded due to its prohibitive runtime with the increase in data points. The experimental setting is relatively challenging, and the weaknesses of each method have emerged. For instance, TPS-RPM is generally robust to outliers, but it can be degraded in the case of severe noises. CPD and GMM have similar performances and are sensitive to outliers. *L*2 *E* and PR-GLS utilize the information of shape context descriptor to guide the registration, but their performances are unstable. APM can only deal with affine deformation, thus leading to its inferior performance. However, compared to other meth-ods that are only locally convergent and fail to handle violent rotations, APM is invariant to rotation owing to its global optimality.

The results of rigid 3-D point cloud registration are pre-sented in Fig. [7](#page36). In our random rotation settings, the locally convergent methods, i.e., GMM, CPD, and ICP, fail to accu-rately register the point clouds. In this regard, the globally optimal method, GoICP, outperforms them by a large margin.

123

International Journal of Computer Vision



|  |
| --- |
| Average Accuracy |

1.2

1.1

1

0.9

0.8

0.7

0.6

TPSRPM

0.5

GMM

CPD

0.4  L2E

PR-GLS

0.3  APM

0.2

0.01 0.03 0.05 0.07 0.09

Noise Level

|  |
| --- |
| Average Accuracy |

1.2

1

0.8

0.6

0.4

0.2

0

-0.2

TPSRPM

GMM

CPD

L2E

PR-GLS

APM

30 60 90 120 150 180

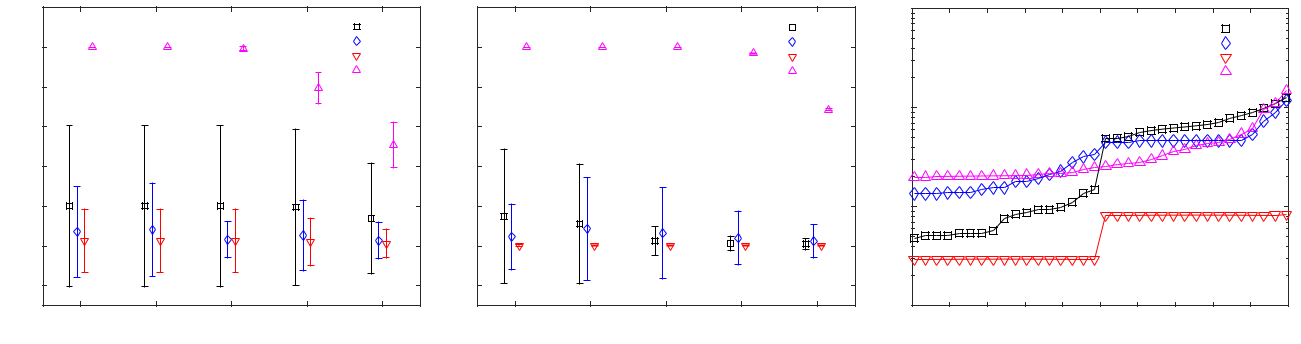
Degree of Rotation

|  |
| --- |
| Average Accuracy |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.2 |  |  |  |  |  | 101 |  |  |  |  |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.8 |  |  |  |  |  | 100 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.6 |  |  |  |  |  | Time/s |  |  |  |  |  |  |  |  |  |  |  |
| 0.4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | TPSRPM |  |  |  |  | 10-1 |  |  |  |  |  |  |  |  | TPSRPM |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.2 | GMM |  |  |  |  |  |  |  |  |  |  |  |  |  | GMM |  |  |
|  | CPD |  |  |  |  |  |  |  |  |  |  |  |  |  | CPD |  |  |
|  | L2E |  |  |  |  |  |  |  |  |  |  |  |  |  | L2E |  |  |
| 0 | PR-GLS |  |  |  |  |  |  |  |  |  |  |  |  |  | PR-GLS |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | APM |  |  |
|  |  |  |  |  |  | 10-2 |  |  |  |  |  |  |  |  |  |  |
|  | 0.2 | 0.4 | 0.6 | 0.8 | 1 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |  |
|  | 0 |  |

Outlier Ratio Cumulative Distribution

**Fig. 6** Quantitative evaluation of non-rigid 2-D shape contour registration



|  |
| --- |
| Average Accuracy |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.2 |  |  |  |  |  | 1.2 |  |  |  |  |  | 103 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | GMM |  |  |  |  |  | GMM |  |  |  |  |  |  |  |  |  |  | GMM |  |  |
| 1 |  |  |  | CPD |  | 1 |  |  |  | CPD |  |  |  |  |  |  |  |  |  |  | CPD |  |  |
|  |  |  | ICP |  |  |  |  | ICP |  |  |  |  |  |  |  |  |  |  | ICP |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | GoICP |  |  |  |  |  | GoICP |  |  |  |  |  |  |  |  |  |  | GoICP |  |  |
| 0.8 |  |  |  |  |  | 0.8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.2 |  |  |  |  | AverageAccuracy | 0.2 |  |  |  |  | Time/s | 102 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | 101 |  |  |  |  |  |  |  |  |  |  |  |
| 0.6 |  |  |  |  |  | 0.6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.4 |  |  |  |  |  | 0.4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0 |  |  |  |  |  | 0 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| -0.2 |  |  |  |  |  | -0.2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.01 | 0.03 | 0.05 | 0.07 | 0.09 |  | 0.2 | 0.4 | 0.6 | 0.8 | 1 |  | 100 | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 180 | 200 |  |
|  |  | 0 |  |

Noise Level Outlier Ratio Cumulative Distribution

**Fig. 7** Quantitative evaluation of rigid 3-D point cloud registration

**6.4 Results on Graph Matching**

Graph matching represents an alternative means to establish correspondences between two feature sets. Here, we evalu-ate seven state-of-the-art methods in the literature, namely, SM (Leordeanu and Hebert [2005](#page49)), SMAC (Cour et al. [2007](#page46)), IPFP (Leordeanu et al. [2009](#page49)), RRWM (Cho et al. [2010](#page45)), TM (Duchenne et al. [2011](#page46)), GNCCP (Liu and Qiao [2014](#page50)), and FGM (Zhou and De la Torre [2015](#page57)) on several extensively used and publicly available datasets. These datasets include the CMU house sequence (Cho et al. [2010](#page45); Zhou and De la Torre [2015](#page57)), the car and motorbike dataset (Zhou and De la Torre [2015](#page57); Leordeanu et al. [2012](#page49)), and the Chinese char-acter dataset (Liu and Qiao [2014](#page50); Zhang et al. [2016](#page56)). The experiments of this part are performed on a desktop with 3.4 GHz Intel Core i5-7500 CPU, 8GB memory.

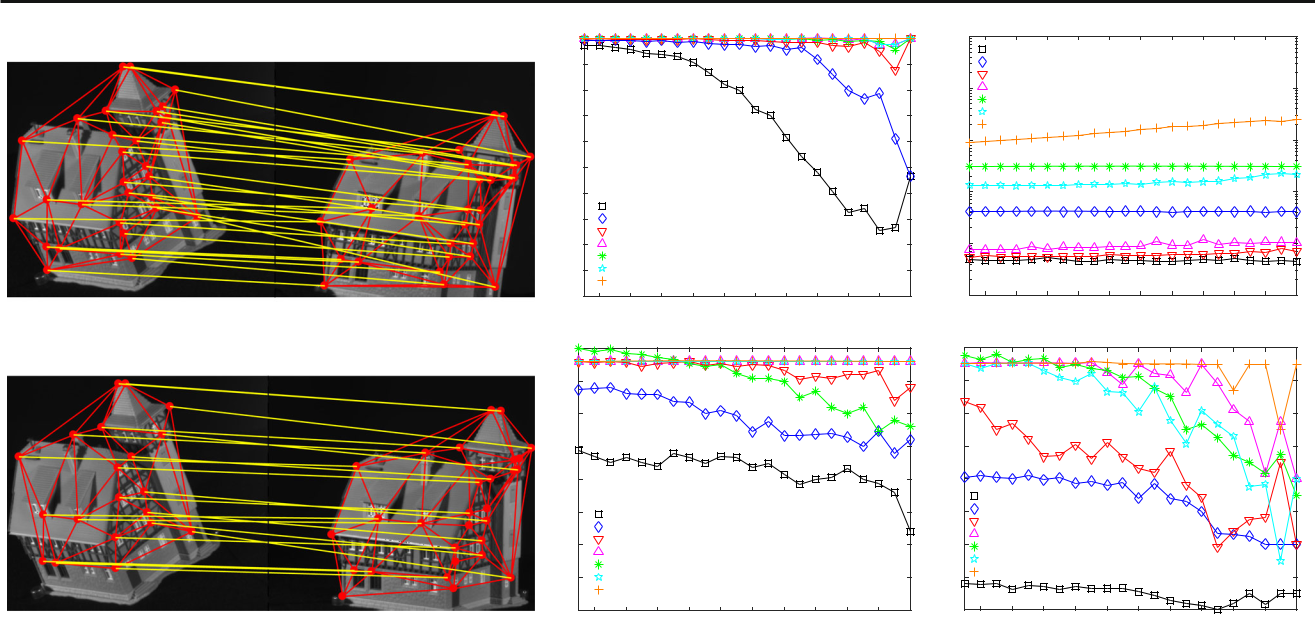
The CMU *house* sequence consists of 111 images of a toy house captured from different viewpoints. Each image has 30 manually marked landmark points with known corre-spondences. We match all images spaced by 5*,* 10*, . . . ,* 110 frames and compute the average performance per separation gap. The large gaps indicate more challenging scenes due to the increasing perspective changes. We build the graph using Delaunay triangulation and construct the affinity matrix sim-ply by the edge distance as in Zhou and De la Torre ([2015](#page57)), except for TM, which has high order. Different from the original equal-size 30-node to 30-node matching, we remove some nodes and conduct unequal-size matching experiments with the corresponding settings of 25 versus 30 and 20 versus

30 on this dataset to test the robustness of these algorithms, as presented in Fig. [8](#page37). The figure shows that in the equal-size matching, most GM methods can achieve near-optimal performance, except for the spectral relaxed baselines. For unequal-size matching, the performance gap has emerged. In summary, FGM achieves the best performance with the highest time cost, and RRWM is the most balanced algo-rithm, which is only inferior to FGM in accuracy but is much more efficient.

The *car* and *motorbike* dataset consists of 30 pairs of car images and 20 pairs of motorbike images obtained from the PASCAL challenges (Everingham et al. [2010](#page46)). Each pair contains 30–60 ground-truth correspondences. We consider the most general graph wherein the edge is directed and the edge feature is asymmetrical. Similarly, the graph is built with Delaunay triangulation, and the affinity matrix is con-structed as in Zhou and De la Torre ([2015](#page57)) except for TM. To test the robustness to outliers, 2 ∼ 20 outliers are ran-domly selected from the background. As shown in Fig. [9](#page37), the path following algorithms, i.e., GNCCP and FGM, outper-form all other methods, except for TM with the highest time cost. The RRWM remains competitive with high accuracy and low runtime. The higher-order TM has achieved remark-able performance in this experiment with consistent optimal performance. Moreover, its runtime is reasonable due to the adopted random sampling strategy used for constructing the three-order affinity matrix. The direct comparison of pairwise and higher-order graph matching methods can be unfair, but

123

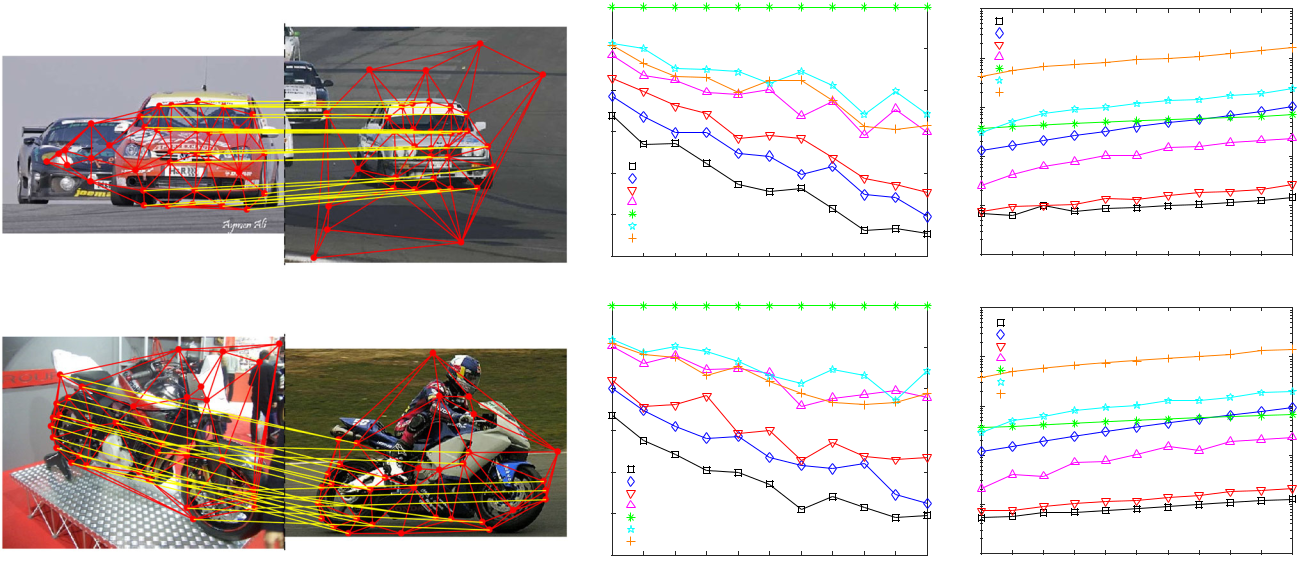
International Journal of Computer Vision



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 |  |  |  |  |  |  |  |  |  |  | 102 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | SM |  |  |  |  |  |  |  |  |  |  |
|  | 0.95 |  |  |  |  |  |  |  |  |  |  |  |  | SMAC |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | IPFP |  |  |  |  |  |  |  |  |  |  |
|  | 0.9 |  |  |  |  |  |  |  |  |  |  | 101 |  | RRWM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  | TM |  |  |  |  |  |  |  |  |  |  |
|  | 0.85 |  |  |  |  |  |  |  |  |  |  |  |  | GNCCP |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | FGM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Accuracy | 0.8 |  |  |  |  |  |  |  |  |  | Time/s | 100 |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.75 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.7 |  |  |  |  |  |  |  |  |  | 10-1 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | SM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.65 | SMAC |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | IPFP |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.6 | RRWM |  |  |  |  |  |  |  |  |  | 10-2 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | TM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.55 | GNCCP |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | FGM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.5 |  |  |  |  |  |  |  |  |  |  | 10-3 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 |  |
|  | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 |  |  |
|  |  |  |  |  | Baseline | |  |  |  |  |  |  |  |  |  |  | Baseline | |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |  |  |  |  | 1 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.9 |  |  |  |  |  |  |  |  |  |  | 0.9 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.8 |  |  |  |  |  |  |  |  |  |  | 0.8 |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.7 |  |  |  |  |  |  |  |  |  |  | 0.7 |  |  |  |  |  |  |  |  |  |  |  |  |
| Accuracy | 0.6 |  |  |  |  |  |  |  |  |  | Accuracy | 0.6 |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | SM | |  |  |  |  |  |  |  |  |  |  |
|  | 0.5 | SM |  |  |  |  |  |  |  |  |  | 0.5 | SMAC | |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | IPFP | |  |  |  |  |  |  |  |  |  |  |
|  |  | SMAC |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | RRWM | |  |  |  |  |  |  |  |  |  |  |
|  |  | IPFP |  |  |  |  |  |  |  |  |  | 0.4 |  |  |  |  |  |  |  |  |  |  |
|  | 0.4 |  |  |  |  |  |  |  |  |  | TM | |  |  |  |  |  |  |  |  |  |  |
|  | RRWM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | GNCCP | |  |  |  |  |  |  |  |  |  |  |
|  |  | TM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | FGM | |  |  |  |  |  |  |  |  |  |  |
|  | 0.3 | GNCCP |  |  |  |  |  |  |  |  |  | 0.3 |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | FGM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.2 |  |  |  |  |  |  |  |  |  |  | 0.2 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 |  |
|  | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 |  |  |
|  |  |  |  |  | Baseline | |  |  |  |  |  |  |  |  |  |  | Baseline | |  |  |  |  |  |  |

**Fig. 8** Quantitative evaluation on the CMU*house*dataset. Top row(from left to right): illustration of equal-size matching with ground-truth correspondence, 30 versus 30 matching results and its runtime

statistics. Bottom row (from left two right): example of unequal-size matching with ground-truth correspondence, 25 versus 30 matching results and 20 versus 30 matching results



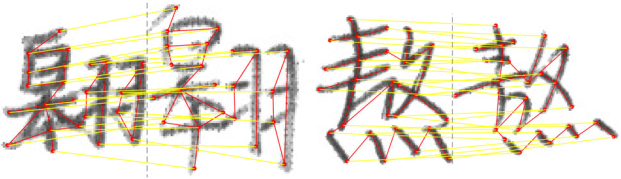
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 |  |  |  |  |  |  |  |  |  |  | 102 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | SM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | SMAC |  |  |  |  |  |  |  |  |  |  |
|  | 0.9 |  |  |  |  |  |  |  |  |  |  |  | IPFP |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | 101 | RRWM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | TM |  |  |  |  |  |  |  |  |  |  |
|  | 0.8 |  |  |  |  |  |  |  |  |  |  |  | GNCCP | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | FGM |  |  |  |  |  |  |  |  |  |  |
| Accuracy |  |  |  |  |  |  |  |  |  |  |  | 100 |  |  |  |  |  |  |  |  |  |  |  |
| 0.7 |  |  |  |  |  |  |  |  |  |  | Time/s |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | SM |  |  |  |  |  |  |  |  |  | 10-1 |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.6 | SMAC |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | IPFP |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.5 | RRWM | |  |  |  |  |  |  |  |  | 10-2 |  |  |  |  |  |  |  |  |  |  |  |
|  | TM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | GNCCP | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | FGM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.4 |  |  |  |  |  |  |  |  |  |  | 10-3 |  |  |  |  |  |  |  |  |  |  |  |
|  | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |  |
|  |  |  |  |  |  | #Outlier |  |  |  |  |  |  |  |  |  |  | #Outlier |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  |  |  |  |  | 102 |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | SM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | SMAC |  |  |  |  |  |  |  |  |  |  |
|  | 0.9 |  |  |  |  |  |  |  |  |  |  | 101 | IPFP |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | RRWM | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | TM |  |  |  |  |  |  |  |  |  |  |
|  | 0.8 |  |  |  |  |  |  |  |  |  |  |  | GNCCP | |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | FGM |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Accuracy |  |  |  |  |  |  |  |  |  |  |  | 100 |  |  |  |  |  |  |  |  |  |  |  |
| 0.7 |  |  |  |  |  |  |  |  |  |  | Time/s |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | 10-1 |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.6 | SM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | SMAC |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | IPFP |  |  |  |  |  |  |  |  |  | 10-2 |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.5 | RRWM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | TM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | GNCCP | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | FGM |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.4 |  |  |  |  |  |  |  |  |  |  | 10-3 |  |  |  |  |  |  |  |  |  |  |  |
|  | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 |  |
|  |  |  |  |  |  | #Outlier |  |  |  |  |  |  |  |  |  |  | #Outlier |  |  |  |  |  |  |

**Fig. 9** Quantitative evaluation on*car*and*motorbike*dataset. Top row(from left to right): example of equal-size matching with ground-truth correspondence, car image matching results and the runtime statistics.

Bottom row (from left two right): example of unequal-size matching with ground-truth correspondence, motorbike matching results and the runtime statistics

123

International Journal of Computer Vision



**Fig. 10** Examples of the Chinese character dataset

the results still exhibit the efficacy of utilizing higher-order information in GM.

The Chinese character dataset has four hand-written Chi-nese characters with marked features wherein each character has 10 samples. We create matching instances between all pairs of samples for each character, i.e. 45 instances each. The average performance is summarized in Table [2](#page39) and the example is shown in Fig. [10](#page38). The scene is relatively challeng-ing, and we use simple edge distances to construct affinity matrix, resulting in the relatively low accuracy for all meth-ods. However, the superior performers are still evident. FGM and TM perform similarly, but TM is more efficient.

**6.5 Results on Pose Estimation**

The camera pose estimation aims to determine the posi-tion and orientation of the camera with respect to the object or scene, which is a significant step in 3-D computer vision tasks, such as SfM, SLAM, and visual localization for self-driving cars and augmented reality. Here, the camera pose estimation of traditional approaches estimates the pose from a set of 2-D versus 3-D matches between pixels in a query image and 3-D points in a scene model. However, the 3-D model is typically obtained via SfM, thus leading to poten-tially inaccurate pose estimates. To address this problem, one alternative is to perform a set of 2-D versus 2-D correspon-dences between two or more images of the same scene.

To estimate the camera pose, the putative sparse feature correspondences must also be constructed with off-the-shelf feature matcher, such as SIFT. Moreover, the most classi-cal pipeline is the combination of SIFT and RANSAC. The geometric model can be estimated and converted into the rel-ative camera pose, i.e., rotation matrix and translation matrix. Many advanced handcrafted methods and trainable ones are considered as good options for their superior performance. Here, we integrate some typical mismatch removal meth-ods between SIFT and RANSAC, while some learning-based methods can intrinsically output the transform matrix from their networks, which can be directly used for this task. In addition, two different datasets, including indoor and outdoor scenes, are used in this experiment. The performance is char-acterized by the mean average precision (mAP), as depicted in Table [3](#page40). The experiments of this part are performed on a server with 2.00 GHz Intel Xeon CPU, 128 GB memory.

In the following, we briefly introduce the datasets and evaluation metrics to be used and provide quantitative com-parisons and analyses.

*Outdoor scenes.* We adopt the Yahoo’s YFCC100Mdataset (Thomee et al. [2016](#page54)), with 100 million publicly accessible tourist photos from the Internet and subsequently curate into 72 image sequences for SfM. From this dataset, 68 sequences are selected as valid raw data. Next, we use the Visual SfM (Heinly et al. [2015](#page48)) to recover the camera poses and generate the ground-truth. This dataset is divided into disjoint subsets for training (60%), validation (20%), and test (20%). For fairness, all learning-based methods are re-trained on the same training set.

*Indoor scenes.* We adopt the SUN3D dataset (Xiao et al.[2013](#page55)), which is an RGBD video dataset with camera poses computed by generalized bundle adjustment. Specifically, all samples in this dataset are subsampled from videos of every 10 frames of feature office-like scenes. This dataset is extremely challenging for sparse correspondence methods due to the few distinctive features, heavy repetitive elements, and substantial self-occlusions. Zhang et al. ([2019b](#page56)) reported that some sequences in this dataset do not provide camera poses. Thus, these sequences are dropped and 239 sequences are finally obtained as valid data. Similar to the data of out-door scenes, the SUN3D dataset is split into disjoint subsets for training (60%), validation (20%), and testing (20%).

*Evaluation Metrics.* Once potential inliers are obtained,it is possible to efficiently estimate the rotation and transla-tion vectors by RANSAC. The performance can be evaluated using the angular difference between the estimated and ground-truth vectors; i.e., the closest arc distance in degrees as the error metric. First, a curve should be generated by clas-sifying whether each pose as accurate or not. The precision should be computed with respect to the given angle threshold from 0◦ to 180◦, and a normalized cumulative curve should be built. Second, the area under curve (AUC) is computed up to a maximum threshold of 5◦, 10◦, or 20◦. Since the curve itself can measure precision, its AUC can be regarded as the metric of mAP.

Several traditional mismatch removal methods, i.e., GMS (Bian et al. [2017](#page44)), ICF (Li and Hu [2010](#page49)), LPM (Ma et al. [2019d](#page51)), SM (Leordeanu and Hebert [2005](#page49)) and VFC (Ma et al. [2014](#page51)) are used for evaluation of the pose estimation task, in addition to two deep-learning-based methods, i.e., LFGC (Moo Yi et al. [2018](#page51)) and OAN (Zhang et al. [2019b](#page56)). For these methods, pose estimation results are obtained by a sub-sequent RANSAC procedure. In addition, plain RANSAC (Fischler and Bolles [1981](#page47)) is also included for comparison. As shown in Table [3](#page40), on the adopted dataset, the perfor-mances of traditional methods are very limited due to the dominant outliers. In contrast, the deep-learning-based meth-ods seem to significantly outperform the traditional methods, resilient to the high outlier ratio.

123

|  |
| --- |
| International Journal of Computer Vision |

**Table 2** Evaluation by average accuracy and runtime on Chinese character dataset (best in bold)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | SM (Leordeanu | SMAC (Cour | IPFP (Leordeanu | RRWM (Cho | TM (Duchenne | GNCCP (Liu and | FGM (Zhou and |
|  | and Hebert | et al. 2007) | et al. 2009) | et al. 2010) | et al. 2011) | Qiao 2014) | De la Torre |
|  | 2005) |  |  |  |  |  | 2015) |
|  |  |  |  |  |  |  |  |
| Character1 (Acc) | 0.2151 | 0.2690 | 0.3325 | 0.5548 | **0**.**7611** | 0.5508 | 0.6048 |
| Character1 (Time) | 0.0045 | 0.0293 | 0.0089 | 0.0408 | 0.2834 | 0.3138 | 2.1719 |
| Character2 (Acc) | 0.3449 | 0.4464 | 0.6580 | 0.8097 | 0.7729 | 0.8879 | **0**.**8986** |
| Character2 (Time) | 0.0033 | 0.0121 | 0.0064 | 0.0129 | 0.2300 | 0.1351 | 1.1638 |
| Character3 (Acc) | 0.2413 | 0.2595 | 0.3889 | 0.5151 | **0**.**9000** | 0.5040 | 0.6500 |
| Character3 (Time) | 0.0039 | 0.0284 | 0.0081 | 0.0343 | 0.2835 | 0.2932 | 1.9731 |
| Character4 (Acc) | 0.2077 | 0.2338 | 0.2879 | 0.5082 | 0.5787 | 0.4242 | **0**.**6116** |
| Character4 (Time) | 0.0033 | 0.0113 | 0.0062 | 0.0218 | 0.2290 | 0.2189 | 1.3926 |
|  |  |  |  |  |  |  |  |

|  |
| --- |
| 123 |

|  |
| --- |
| 123 |

**Table 3** mAP performance of representative methods for pose estimation on*YFCC100M*and*SUN3D*datasets

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | degree | GMS (Bian et al. | ICF (Li and Hu | LPM (Ma et al. | SM (Leordeanu | VFC (Ma et al. | RANSAC | (Fis- | LFGC (Moo Yi | OAN | (Zhang |
|  |  | 2017) | 2010) | 2019d) | and Hebert 2005) | 2014) | chler and | Bolles | et al. 2018) | et al. 2019b) | |
|  |  |  |  |  |  |  | 1981) |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| YFCC100M | 5 | 1.61 | 3*.*71 | 6*.*26 | 3*.*77 | 7*.*79 | 9*.*08 |  | 47.98 | 52.18 |  |
|  | 10 | 3.30 | 7*.*76 | 11*.*65 | 7*.*79 | 13*.*87 | 14*.*28 |  | – | – |  |
|  | 20 | 7.04 | 16*.*13 | 21*.*79 | 16*.*48 | 24*.*13 | 22*.*80 |  | – | – |  |
| SUN3D | 5 | 0.39 | 3*.*48 | 6*.*36 | 5*.*08 | 7*.*65 | 2*.*85 |  | 15.98 | 17.50 |  |
|  | 10 | 1.22 | 6*.*65 | 11*.*14 | 9*.*13 | 12*.*54 | 5*.*61 |  | – | – |  |
|  | 20 | 3.94 | 13*.*52 | 19*.*52 | 16*.*79 | 20*.*78 | 11*.*22 |  | — | – |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

|  |
| --- |
| International Journal of Computer Vision |

|  |
| --- |
| International Journal of Computer Vision |

**Table 4** Performance by maximum recall rate (%) at precision equals to 100% for loop closure detection (best in bold)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| dataset | LPM (Ma et al. | GMS (Bian et al. | GS (Liu and Yan | SM (Leordeanu and Hebert 2005) | ICF (Li and Hu | RANSAC | LORANSAC | VFC (Ma et al. |
|  | 2019d) | 2017) | 2010) |  | 2010) | (Fischler and | (Lebeda et al. | 2014) |
|  |  |  |  |  |  | Bolles 1981) | 2012) |  |
|  |  |  |  |  |  |  |  |  |
| Lip6Indoor | 91.82 | 88.64 | 87.73 | 91*.*36 | 88.18 | **93**.**18** | **93**.**18** | 90.45 |
| Lip6Outdoor | 54.89 | 54.23 | 54.06 | 55*.*22 | 51.41 | 56.22 | **56**.**55** | 54.06 |
| NewCollege | 84.99 | 84.26 | 84.75 | 85*.*96 | 64.89 | 85.47 | 84.99 | **86**.**44** |
| CityCentre | 73.08 | 70.05 | 71.66 | 71*.*3 | 45.28 | 74.33 | **75**.**04** | 71.12 |
|  |  |  |  |  |  |  |  |  |

|  |
| --- |
| 123 |

International Journal of Computer Vision

**6.6 Results on Loop-Closure Detection**

Appearance-based loop-closure detection is a fundamental component in visual SLAM. The essence involves recogniz-ing previously visited areas of the environment. This task is crucial in reducing the drift of the estimated trajectory caused by the accumulative error and contributes to global consistent mapping.

Appearance-based loop-closure detection only uses image similarity to identify previously visited places. This category commonly starts with the construction of a set of putative correspondences by a feature operator, such as SIFT, between the current image and each previously visited image. Then, the closed loop is determined on the basis of the number of accurate matches using mismatch removal methods. This solution is simple but relatively effective.

Moreover, the computational requirement in directly real-izing feature matching between the current image and each previously visited image would be largely increased. To ensure the real-time performance of loop-closure detection, we use a two-step approach. In the first step, loop-closure candidates are selected by the BoW method with presup-posed score threshold, which is fast and easy to implement. However, the BoW method only considers whether or not a feature exists and neglects the spatial arrangement of the fea-tures, thereby leading to perceptual aliasing problem. Thus, in the second step, a robust feature matching algorithm is required to determine whether a loop-closure candidate is a true loop-closure event.

To evaluate the effectiveness and compare the perfor-mance of the loop-closure detection methods based on feature matching, we conduct extensive experiments on four differ-ent datasets, including *NewCollege*, *CityCentre*, *Lip6Indoor*, and *Lip6Outdoor*. The performance is characterized by the maximum recall that can be achieved at 100% precision, as shown in Table [4](#page41). The experiments are performed on a desk-top with 2*.*6 GHz Intel Core CPU, 16*G B* memory.

The *NewCollege* and *CityCentre* datasets are obtained from the work of Cummins and Newman Cummins and Newman ([2008](#page46)). The *NewCollege* dataset contains 1*,* 073 images with size of 640 × 480, and the *CityCentre* dataset contains 1*,* 237 images with size of 640 × 480. The images were recorded by means of the vision system of a wheeled robotic platform while traversing 2*.*2*km* through a college’s campus grounds and adjoining parks with buildings, roads, gardens, cars, and people. The environment is outdoor and dynamic.

The *Lip6Indoor* and *Lip6Outdoor* datasets are obtained from Angeli et al. ([2008](#page44)). The *Lip6Indoor* dataset has 388 images with size of 240 × 192; it is an indoor image sequence with strong perceptual aliasing problem. While the *Lip6Outdoor* dataset has 1*,* 063 images with size of 240×192; it is a long outdoor image sequence of a street with

many buildings, cars, and people. Both image sequences are grabbed with a single-monocular handheld camera. In addi-tion, a binary matrix is defined as the ground truth for each dataset, whose rows and columns correspond to images at different time indices. Each element in this binary matrix denotes the presence (set to 1) or absence (set to 0) of a loop-closure event between the corresponding frame pair.

To generate consistent maps, the loop-closure detection module should obtain true positive detections to provide information for the back-end optimization, thereby reducing the drift of the estimated trajectory caused by accumulative error. However, the loop-closure detection result must also include no false positive detections as this can affect the per-formance of a full SLAM system and result in a completely inaccurate map result. In summary, the loop closure mech-anisms should work at 100% precision while maintaining high recall rate. In such cases, the evaluation of loop-closure detection algorithm is performed in terms of precision-recall metrics. Here, precision is the ratio of the number of true pos-itive loop-closure detections to the number of total positive loop-closure detections identified by the system, and recall is the ratio between the true positive loop closure detections and the total actual loop-closure events defined by the ground truth of dataset. Combining the analysis and the curve, we focus on the maximum recall that can be achieved at 100% precision, indicating that the loop-closure detection result includes no false positive detection and avoids the influence in a full SLAM system.

Some of the representative mismatch removal methods are adopted for comparison in our experiment. The quan-titative comparisons, with respect to maximum recall rate at precision of 100% on different datasets, are presented in Table [4](#page41). From the results, we can see that the methods that pursue relaxed geometric constraints, i.e., LPM (Ma et al. [2019d](#page51)), GMS (Bian et al. [2017](#page44)), GS (Liu and Yan [2010](#page50)), SM (Leordeanu and Hebert [2005](#page49)), ICF (Li and Hu [2010](#page49)) and VFC (Ma et al. [2014](#page51)), are less favored in this task. In comparison, the resampling methods that exploit parametric models of the correspondences, i.e., RANSAC (Fischler and Bolles [1981](#page47)) and LORANSAC (Lebeda et al. [2012](#page49)), can give better results for loop-closure detection.

**7 Conclusions and Future Trends**

Image matching has played a significant role in various visual applications and has attracted considerable attention. Researchers have also achieved significant progress in this field in the past few decades. Therefore, we provide a compre-hensive review of the existing image matching methods–from handcrafted to trainable ones–in order to provide better reference and understanding for the researchers in this com-munity.

123

International Journal of Computer Vision

Image matching can be briefly classified into area- and feature-based matching. Area-based methods are used to achieve dense matching without detecting any salient fea-ture points from the images. They are more welcomed in high overlapping image matching (such as medical image registration) and narrow-baseline stereo (such as binocular stereo matching). The deep learning-based techniques have drawn increasing attention for such a pipeline. Therefore, we provide a brief review of these types of methods in Sect. [4](#page16) and focus more on the learning-based methods.

The feature-based image matching can effectively address the limitations in large viewpoint, wide baseline, and seri-ous non-rigid image matching problems. It can be used in a pipeline of salient feature detection, discriminative descrip-tion, and reliable matching, often including transformation model estimation. Following this procedure, feature detec-tion can extract the distinctive structure from the image. Meanwhile, feature description may be regarded as an image representation method, which is widely used for image cod-ing and similarity measurement. The matching step can be extended into different types of matching forms, such as graph matching, point set registration, descriptor match-ing and mismatch removal, as well as the matching task in 3-D cases. These are more flexible and applicable than area-based methods, thereby receiving considerable atten-tion in image matching area. Therefore, we review them with the core idea that they are used from traditional techniques to classical learning and deep learning. Moreover, to pro-vide a comprehensive understanding of the significance in image matching, we introduce several applications related to image matching. We also provide comprehensive and objec-tive comparisons and analyses of these classical and deep learning-based techniques through extensive experiments on representative datasets.

Despite the considerable development in both theory and performance, image matching remains an open problem with challenges for further efforts.

– The two-stage strategy for feature matching, which has been widely adopted in the literature, performs mismatch removal on only a small set of potential correspondences with sufficiently similar descriptors. However, this may lead to restricted performance in recall, which can be problematic for some scenarios.

– In a different scenario, correspondences are sought not between projections of physically the same points in different images, but between semantic analogs across different instances within a category. This requires new paradigms for feature matching in feature description and mismatch removal.

– Joint matching of multiple images has been proven to drastically boost the matching performance of pairwise matching and has attracted considerable attention in

recent years. However, the complexity is still the main concern of the problem. Thus, practical and efficient algo-rithms are required.

– In recent years, deep learning schemes have rapidly evolved and shown tremendous improvements in many research fields related to computer vision. However, in the literature of feature matching, most works have applied deep learning techniques to feature detection and descrip-tion. Thus, the potential capacity for accurate feature matching can be further explored in the future.

– Image matching among multi-modal images is still an unsolved problem. In the future, deep learning techniques can be used for better feature detection and description performance.

– Feature matching is a fundamental task in computer vision. However, its application has not been sufficiently explored. Thus, one promising research direction is to customize modern feature matching techniques to sat-isfy different requirements of practical vision tasks, e.g., SfM and SLAM.

**Acknowledgements** This work was partly supported by the NationalNatural Science Foundation of China under Grant Nos. 61773295 and 61972250, Natural Science Foundation of Hubei Province under Grant No. 2019CFA037, and National Key Research and Development Pro-gram of China under Grant No. 2018AAA0100704.

**Compliance with ethical standards**

**Conflict of Interest** The authors declare no conflict of interest.

**Open Access** This article is licensed under a Creative CommonsAttribution 4.0 International License, which permits use, sharing, adap-tation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indi-cate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copy-right holder. To view a copy of this licence, visit [http://creativecomm](http://creativecommons.org/licenses/by/4.0/) [ons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

**References**

Aanæs, H., Dahl, A. L., & Pedersen, K. S. (2012). Interesting interest points. *International Journal of Computer Vision*, *97*(1), 18–35.

Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., & Dahl, A. B. (2016). Large-scale data for multiple-view stereopsis. *International Jour-nal of Computer Vision*, *120*(2), 153–168.

Abdel-Hakim, A. E., & Farag, A. A. (2006). Csift: A sift descriptor with color invariant characteristics. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1978–1983.

123

International Journal of Computer Vision

Adamczewski, K., Suh, Y., & Mu Lee, K. (2015). Discrete tabu search for graph matching. In *Proceedings of the IEEE international con-ference on computer vision*, pp. 109–117.

Adams, W. P., & Johnson, T. A. (1994). Improved linear programming-based lower bounds for the quadratic assignment problem. *DIMACS Series in Discrete Mathematics and Theoretical Com-puter Science*, *16*, 43–77.

Aflalo, Y., Dubrovina, A., & Kimmel, R. (2016). Spectral general-ized multi-dimensional scaling. *International Journal of Computer* *Vision*, *118*(3), 380–392.

Agrawal, M., Konolige, K., & Blas, M. R. (2008). Censure: Center surround extremas for realtime feature detection and matching. In *Proceedings of the European conference on computer vision*, pp.102–115.

Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). Freak: Fast retina key-point. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 510–517.

Alcantarilla, P. F., Bartoli, A., & Davison, A. J. (2012) Kaze features. In *Proceedings of the European conference on computer vision*, pp. 214–227.

Alcantarilla, P. F., & Solutions, T. (2011). Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *34*(7), 1281–1298.

Aldana-Iuit, J., Mishkin, D., Chum, O., & Matas, J. (2016). In the sad-dle: Chasing fast and repeatable features. In *Proceedings of the* *international conference on pattern recognition*, pp. 675–680.

Almohamad, H., & Duffuaa, S. O. (1993). A linear programming approach for the weighted graph matching problem. *IEEE Trans-actions on Pattern Analysis and Machine Intelligence*, *15*(5),522–525.

Amberg, B., Romdhani, S., & Vetter, T. (2007). Optimal step nonrigid ICP algorithms for surface registration. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1–8.

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J. A. (2008). A fast and incremental method for loop-closure detection using bags of visual words. In: *IEEE transactions on robotics*, pp. 1027–1037.

Arandjelovi´c, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 2911–2918.

Arar, M., Ginger, Y., Danon, D., Bermano, A. H., & Cohen-Or, D. (2020). Unsupervised multi-modal image registration via geome-try preserving image-to-image translation. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recognition*,pp. 13,410–13,419.

Aubry, M., Schlickewei, U., & Cremers, D. (2011). The wave kernel signature: A quantum mechanical approach to shape analysis. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 1626–1633.

Avrithis, Y., & Tolias, G. (2014). Hough pyramid matching: Speeded-up geometry re-ranking for large scale image retrieval. *International* *Journal of Computer Vision*, *107*(1), 1–19.

Awrangjeb, M., & Lu, G. (2008). Robust image corner detection based on the chord-to-point distance accumulation technique. *IEEE* *Transactions on Multimedia*, *10*(6), 1059–1072.

Awrangjeb, M., Lu, G., & Fraser, C. S. (2012). Performance compar-isons of contour-based corner detectors. *IEEE Transactions on* *Image Processing*, *21*(9), 4167–4179.

Babai, L. (2018). Groups, graphs, algorithms: The graph isomorphism problem. In *Proceedings of the international congress of mathe-maticians*, pp. 3319–3336.

Balntas, V., Johns, E., Tang, L., & Mikolajczyk, K. (2016a). Pn-net: Conjoined triple deep network for learning local image descriptors. arXiv preprint [arXiv:1601.05030](http://arxiv.org/abs/1601.05030).

Balntas, V., Lenc, K., Vedaldi, A., & Mikolajczyk, K. (2017). Hpatches:

A benchmark and evaluation of handcrafted and learned local

descriptors. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 5173–5182.

Balntas, V., Riba, E., Ponsa, D., & Mikolajczyk, K. (2016b). Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British machine vision con-ference*, pp. 1–11.

Barath, D., & Matas, J. (2018). Graph-cut ransac. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.6733–6741.

Barath, D., Ivashechkin, M., & Matas, J. (2019a). Progressive napsac: Sampling from gradually growing neighborhoods. arXiv preprint [arXiv:1906.02295](http://arxiv.org/abs/1906.02295).

Barath, D., Matas, J., & Noskova, J. (2019b). Magsac: Marginalizing sample consensus. In *Proceedings of the IEEE conference on com-puter vision and pattern recognition*, pp. 10,197–10,205.

Barath, D., Noskova, J., Ivashechkin, M., & Matas, J. (2020). Magsac++, a fast, reliable and accurate robust estimator. In *Pro-ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1304–1312.

Barroso-Laguna, A., Riba, E., Ponsa, D., & Mikolajczyk, K. (2019). Key.net: Keypoint detection by handcrafted and learned CNN filters. In *Proceedings of the IEEE international conference on* *computer vision*, pp. 5836–5844.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Proceedings of the European conference on computer* *vision*, pp. 404–417.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Bazin, J.C., Seo, Y., & Pollefeys, M. (2012). Globally optimal consensus set maximization through rotation search. In *Proceedings of the* *Asian conference on computer vision*, pp. 539–551.

Bellavia, F., & Colombo, C. (2020). Is there anything new to say about sift matching? *International Journal of Computer Vision*, *128*(3), 1847–1866.

Belongie, S., Malik, J., & Puzicha, J. (2001). Shape context: A new descriptor for shape matching and object recognition. In *Advances* *in neural information processing systems*, pp. 831–837.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *4*, 509–522.

Bernard, F., Theobalt, C., & Moeller, M. (2018). Ds\*: Tighter lifting-free convex relaxations for quadratic matching problems. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 4310–4319.

Bernard, F., Thunberg, J., Swoboda, P., & Theobalt, C. (2019). Hippi: Higher-order projected power iterations for scalable multi-matching. In *Proceedings of the IEEE international conference on* *computer vision*, pp. 10,284–10,293.

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-d shapes. In *Sensor fusion IV: Control paradigms and data struc-tures*, Vol. 1611, pp. 586–607.

Bhattacharjee, D., & Roy, H. (2019). Pattern of local gravitational force (plgf): A novel local image descriptor. In *IEEE transactions on* *pattern analysis and machine intelligence*.

Bhowmik, A., Gumhold, S., Rother, C., & Brachmann, E. (2020). Rein-forced feature points: Optimizing feature detection and description for a high-level task. In *Proceedings of the IEEE/CVF conference* *on computer vision and pattern recognition*, pp. 4948–4957.

Bian, J., Lin, W. Y., Matsushita, Y., Yeung, S. K., Nguyen, T. D., & Cheng, M. M. (2017). Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 4181–4190.

Biasotti, S., Cerri, A., Bronstein, A., & Bronstein, M. (2016). Recent trends, applications, and perspectives in 3d shape similarity assess-

123

International Journal of Computer Vision

ment. In *Computer graphics forum*, Vol. 35, Wiley Online Library, pp. 87–119.

Blais, G., & Levine, M. D. (1995). Registering multiview range data to create 3d computer objects. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *17*(8), 820–824.

Bonny, M. Z., & Uddin, M. S. (2016). Feature-based image stitching algorithms. In *Proceedings of the international workshop on com-putational intelligence*, pp. 198–203.

Boscaini, D., Masci, J., Melzi, S., Bronstein, M. M., Castellani, U.,

* Vandergheynst, P. (2015). Learning class-specific descriptors for deformable shapes using localized spectral convolutional net-works. In *Computer graphics forum*, Vol. 34, Wiley Online Library, pp. 13–23.

Boscaini, D., Masci, J., Rodolà, E., & Bronstein, M. (2016). Learning shape correspondence with anisotropic convolutional neural net-works. In *Advances in neural information processing systems*, pp. 3189–3197.

Brachmann, E., & Rother, C. (2019). Neural-guided RANSAC: Learn-ing where to sample model hypotheses. In *Proceedings of the IEEE* *international conference on computer vision*, pp. 4322–4331.

Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., & Rother, C. (2017). Dsac-differentiable RANSAC for camera localization. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 6684–6692.

Bronstein, M. M., & Kokkinos, I. (2010). Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.1704–1711.

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2006). General-ized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of* *Sciences*, *103*(5), 1168–1172.

Brown, M., Hua, G., & Winder, S. (2010). Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *33*(1), 43–57.

Brown, M., & Lowe, D. G. (2007). Automatic panoramic image stitch-ing using invariant features. *International Journal of Computer* *Vision*, *74*(1), 59–73.

Caelli, T., & Kosinov, S. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *26*(4), 515–519.

Caetano, T. S., McAuley, J. J., Cheng, L., Le, Q. V., & Smola, A. J. (2009). Learning graph matching. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *31*(6), 1048–1058.

Cai, H., Mikolajczyk, K., & Matas, J. (2010). Learning linear discrimi-nant projections for dimensionality reduction of image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 338–352.

Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features. In *Proceedings of the* *European conference on computer vision*, pp. 778–792.

Campbell, D., & Petersson, L. (2015). An adaptive data representation for robust point-set registration and merging. In *Proceedings of* *the IEEE international conference on computer vision*, pp. 4292–4300.

Campbell, D., & Petersson, L. (2016). Gogma: Globally-optimal gaus-sian mixture alignment. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 5685–5694.

Canny, J. (1987). A computational approach to edge detection. In *Read-ings in computer vision*, Elsevier, pp. 184–203.

Cao, S. Y., Shen, H. L., Chen, S. J., & Li, C. (2020). Boosting structure consistency for multispectral and multimodal image registration. *IEEE Transactions on Image Processing*, *29*, 5147–5162.

Castellani, U., Cristani, M., Fantoni, S., & Murino, V. (2008). Sparse points matching by combining 3d mesh saliency with statistical

descriptors. In *Computer graphics forum*, Vol. 27, Wiley Online Library, pp. 643–652.

Chang, J. R., & Chen, Y. S. (2018). Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 5410–5418.

Chang, W., & Zwicker, M. (2009). Range scan registration using reduced deformable models. In *Computer graphics forum*, Vol. 28, Wiley Online Library, pp. 447–456.

Chang, M. C., & Kimia, B. B. (2011). Measuring 3d shape similarity by graph-based matching of the medial scaffolds. *Computer Vision* *and Image Understanding*, *115*(5), 707–720.

Chen, Q., & Koltun, V. (2015). Robust nonrigid registration by convex optimization. In *Proceedings of the IEEE international conference* *on computer vision*, pp. 2039–2047.

Chen, Y., Guibas, L., & Huang, Q. (2014). Near-optimal joint object matching via convex relaxation. In *Proceedings of the interna-tional conference on machine learning*, pp. 100–108.

Chen, Y. C., Huang, P. H., Yu, L. Y., Huang, J. B., Yang, M. H., & Lin, Y. Y. (2018). Deep semantic matching with foreground detection and cycle-consistency. In *Proceedings of the Asian conference on* *computer vision*, pp. 347–362.

Chen, J., Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2013). Rlbp: Robust local binary pattern. In *Proceedings of the British machine* *vision conference*.

Chen, J., Wang, L., Li, X., & Fang, Y. (2019). Arbicon-net: Arbitrary continuous geometric transformation networks for image registra-tion. In *Advances in neural information processing systems*, pp. 3410–3420.

Chen, H. M., Arora, M. K., & Varshney, P. K. (2003a). Mutual information-based image registration for remote sensing data. *International Journal of Remote Sensing*, *24*(18), 3701–3706.

Chen, H., & Bhanu, B. (2007). 3d free-form object recognition in range images using local surface patches. *Pattern Recognition Letters*, *28*(10), 1252–1262.

Chen, Q. S., Defrise, M., & Deconinck, F. (1994). Symmetric phase-only matched filtering of Fourier-Mellin transforms for image registration and recognition. *IEEE Transactions on Pattern Anal-ysis and Machine Intelligence*, *16*(12), 1156–1168.

Chen, C., Li, Y., Liu, W., & Huang, J. (2015). Sirf: Simultaneous satel-lite image registration and fusion in a unified framework. *IEEE* *Transactions on Image Processing*, *24*(11), 4213–4224.

Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., et al. (2009). Wld: A robust local image descriptor. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *32*(9), 1705–1720.

Chen, J., Tian, J., Lee, N., Zheng, J., Smith, R. T., & Laine, A. F. (2010). A partial intensity invariant feature descriptor for multimodal retinal image registration. *IEEE Transactions on Biomedical Engi-neering*, *57*(7), 1707–1718.

Chen, H. M., Varshney, P. K., & Arora, M. K. (2003b). Perfor-mance of mutual information similarity measure for registration of multitemporal remote sensing images. *IEEE Transactions on* *Geoscience and Remote Sensing*, *41*(11), 2445–2454.

Chertok, M., & Keller, Y. (2010). Efficient high order matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(12), 2205–2215.

Chetverikov, D., Stepanov, D., & Krsek, P. (2005). Robust Euclidean alignment of 3d point sets: The trimmed iterative closest point algorithm. *Image and Vision Computing*, *23*(3), 299–309.

Cho, M., & Lee, K. M. (2012). Progressive graph matching: Making a move of graphs via probabilistic voting. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.398–405.

Cho, M., Lee, J., & Lee, K. M. (2010). Reweighted random walks for graph matching. In *Proceedings of the European conference on* *computer vision*, pp. 492–505.

123

International Journal of Computer Vision

Chopra, S., Hadsell, R., LeCun, Y., et al. (2005). Learning a similar-ity metric discriminatively, with application to face verification. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 539–546.

Choy, C. B., Gwak, J., Savarese, S., & Chandraker, M. (2016). Univer-sal correspondence network. In *Advances in neural information* *processing systems*, pp. 2414–2422.

Choy, C., Lee, J., Ranftl, R., Park, J., & Koltun, V. (2020). High-dimensional convolutional networks for geometric pattern recog-nition. In *Proceedings of the IEEE/CVF conference on computer* *vision and pattern recognition*, pp. 11,227–11,236.

Chui, H., & Rangarajan, A. (2003). A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, *89*(2–3), 114–141.

Chum, O., & Matas, J. (2005). Matching with prosac-progressive sam-ple consensus. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 220–226.

Chum, O., Matas, J., & Kittler, J. (2003). Locally optimized ransac. In *Proceedings of the joint pattern recognition symposium*, Springer,pp. 236–243.

Churchill, D., & Vardy, A. (2013). An orientation invariant visual hom-ing algorithm. *Journal of Intelligent & Robotic Systems*, *71*(1), 3–29.

Cook, D. J., & Holder, L. B. (2006). *Mining graph data*. New York:

Wiley.

Cour, T., Srinivasan, P., & Shi, J. (2007). Balanced graph matching. In *Advances in neural information processing systems*, pp. 313–320.

Cummins, M., & Newman, P. (2008). Fab-map: Probabilistic localiza-tion and mapping in the space of appearance. *The International* *Journal of Robotics Research*, *27*(6), 647–665.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE conference on com-puter vision and pattern recognition*, pp. 886–893.

Danelljan, M., Meneghetti, G., Shahbaz Khan, F., & Felsberg, M. (2016). A probabilistic framework for color-based point set regis-tration. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 1818–1826.

Datar, M., Immorlica, N., Indyk, P., & Mirrokni, V. S. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on computational geometry*, pp. 253–262.

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O. (2007). Monoslam: Real-time single camera slam. *IEEE Transactions on* *Pattern Analysis and Machine Intelligence*, *6*, 1052–1067.

Dawn, S., Saxena, V., & Sharma, B. (2010). Remote sensing image reg-istration techniques: A survey. In *Proceedings of the international* *conference on image and signal processing*, pp. 103–112.

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., & Isgum, I. (2019). A deep learning framework for unsuper-vised affine and deformable image registration. *Medical Image* *Analysis*, *52*, 128–143.

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., & Isgum, I. (2017). End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep learning in medi-cal image analysis and multimodal learning for clinical decision support*, Springer, pp. 204–212.

Deng, H., Birdal, T., & Ilic, S. (2018). Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.195–205.

Deng, H., Zhang, W., Mortensen, E., Dietterich, T., & Shapiro, L. (2007). Principal curvature-based region detector for object recog-nition. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 1–8.

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2016). Deep image homography estimation. arXiv preprint [arXiv:1606.03798](http://arxiv.org/abs/1606.03798).

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *Pro-ceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236.

Dong, J., & Soatto, S. (2015). Domain-size pooling in local descrip-tors: Dsp-sift. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 5097–5106.

Dorai, C., & Jain, A. K. (1997). Cosmos-a representation scheme for 3d free-form objects. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *19*(10), 1115–1130.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceed-ings of the IEEE international conference on computer vision*, pp.2758–2766.

Duan, Y., Lu, J., Wang, Z., Feng, J., & Zhou, J. (2017). Learning deep binary descriptor with multi-quantization. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.1183–1192.

Duchenne, O., Bach, F., Kweon, I. S., & Ponce, J. (2011). A tensor-based algorithm for high-order graph matching. *IEEE Transactions on* *Pattern Analysis and Machine Intelligence*, *33*(12), 2383–2395.

Du, Q., Fan, A., Ma, Y., Fan, F., Huang, J., & Mei, X. (2018). Infrared and visible image registration based on scale-invariant piifd feature and locality preserving matching. *IEEE Access*, *6*, 64107–64121.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., & Sattler, T. (2019). D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 8092–8101.

Dym, N., Maron, H., & Lipman, Y. (2017). Ds++: A flexible, scal-able and provably tight relaxation for matching problems. arXiv preprint [arXiv:1705.06148](http://arxiv.org/abs/1705.06148).

Egozi, A., Keller, Y., & Guterman, H. (2012). A probabilistic approach to spectral graph matching. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *35*(1), 18–27.

Elad, A., & Kimmel, R. (2003). On bending invariant signatures for surfaces. *IEEE Transactions on Pattern Analysis and Machine* *Intelligence*, *25*(10), 1285–1295.

Elbaz, G., Avraham, T., & Fischer, A. (2017). 3d point cloud registration for localization using a deep neural network auto-encoder. In *Pro-ceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4631–4640.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., & Burgard, W. (2012). An evaluation of the rgb-d slam system. In *Proceedings* *of the IEEE international conference on robotics and automation*,pp. 1691–1696.

Erin Liong, V., Lu, J., Wang, G., Moulin, P., & Zhou, J. (2015). Deep hashing for compact binary codes learning. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.2475–2483.

Erlik Nowruzi, F., Laganiere, R., & Japkowicz, N. (2017). Homogra-phy estimation from image pairs with hierarchical convolutional networks. In *Proceedings of the IEEE international conference on* *computer vision*, pp. 913–920.

Evangelidis, G. D., & Horaud, R. (2018). Joint alignment of multiple point sets with batch and incremental expectation-maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1397–1410.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisser-man, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fan, B., Kong, Q., Wang, X., Wanga, Z., Xiang, S., Pan, C., et al. (2019). A performance evaluation of local features for image-based 3d reconstruction. *IEEE Transactions on Image Processing*, *28*(10), 4774–4789.

123

International Journal of Computer Vision

Fan, B., Wu, F., & Hu, Z. (2011). Rotationally invariant descriptors using intensity order pooling. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *34*(10), 2031–2045.

Ferrante, E., & Paragios, N. (2017). Slice-to-volume medical image registration: A survey. *Medical Image Analysis*, *39*, 101–123.

Ferraz, L., & Binefa, X. (2012). A sparse curvature-based detector of affine invariant blobs. *Computer Vision and Image Understanding*, *116*(4), 524–537.

Fey, M., Lenssen, J. E., Morris, C., Masci, J., & Kriege, N. M. (2020). Deep graph matching consensus. In *International conference on* *learning representations*.

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

Fitzgibbon, A. W. (2003). Robust registration of 2d and 3d point sets.

*Image and Vision Computing*, *21*(13–14), 1145–1153.

Flint, A., Dick, A., & Van Den Hengel, A. (2007). Thrift: Local 3d structure recognition. In *Proceedings of the biennial conference* *on digital image computing techniques and applications*, pp. 182–188.

Fogel, F., Jenatton, R., Bach, F., & d’Aspremont, A. (2013). Convex relaxations for permutation problems. In *Advances in neural infor-mation processing systems*, pp. 1016–1024.

Foroosh, H., Zerubia, J. B., & Berthod, M. (2002). Extension of phase correlation to subpixel registration. *IEEE Transactions on Image* *Processing*, *11*(3), 188–200.

Forssén, P. E. (2007). Maximally stable colour regions for recognition and matching. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 1–8.

Fragoso, V., Sen, P., Rodriguez, S., & Turk, M. (2013). Evsac: accel-erating hypotheses generation by modeling matching scores with extreme value theory. In *Proceedings of the IEEE international* *conference on computer vision*, pp. 2472–2479.

Frome, A., Huber, D., Kolluri, R., Bülow, T., & Malik, J. (2004). Rec-ognizing objects in range data using regional point descriptors. In *Proceedings of the European conference on computer vision*, pp.224–237.

Gao, W., & Tedrake, R. (2019). Filterreg: Robust and efficient prob-abilistic point-set registration using Gaussian filter and twist parameterization. In *Proceedings of the IEEE conference on com-puter vision and pattern recognition*, pp. 11,095–11,104.

Gauglitz, S., Höllerer, T., & Turk, M. (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *International* *Journal of Computer Vision*, *94*(3), 335–360.

Gay-Bellile, V., Bartoli, A., & Sayd, P. (2008). Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 87–104.

Gelfand, N., Mitra, N. J., Guibas, L. J., & Pottmann, H. (2005). Robust global registration. In *Symposium on geometry processing*, Vol. 2, Vienna, Austria, p. 5.

Georgakis, G., Karanam, S., Wu, Z., Ernst, J., & Kosecká, J. (2018). End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 1965–1973.

Ghosh, D., & Kaabouch, N. (2016). A survey on image mosaicing techniques. *Journal of Visual Communication and Image Repre-sentation*, *34*, 1–11.

Gil, A., Mozos, O. M., Ballesta, M., & Reinoso, O. (2010). A compara-tive evaluation of interest point detectors and local descriptors for visual slam. *Machine Vision and Applications*, *21*(6), 905–920.

Gionis, A., Indyk, P., Motwani, R., et al. (1999). Similarity search in high dimensions via hashing. In *Proceedings of the international* *conference on very large databases*, pp. 518–529.

Giraldo, L. G. S., Hasanbelliu, E., Rao, M., & Principe, J. C. (2017). Group-wise point-set registration based on rényi’s second order entropy. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 2454–2462.

Glaunes, J., Trouvé, A., & Younes, L. (2004). Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 712–718.

Gojcic, Z., Zhou, C., Wegner, J. D., Guibas, L. J., & Birdal, T. (2020). Learning multiview 3d point cloud registration. In *Proceedings of* *the IEEE/CVF conference on computer vision and pattern recog-nition*, pp. 1759–1769.

Gold, S., & Rangarajan, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *18*(4), 377–388.

Gold, S., Rangarajan, A., Lu, C. P., Pappu, S., & Mjolsness, E. (1998). New algorithms for 2d and 3d point matching: Pose estimation and correspondence. *Pattern Recognition*, *31*(8), 1019–1031.

Golyanik, V., Aziz Ali, S., & Stricker, D. (2016). Gravitational approach for point set registration. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 5802–5810.

Gong, Y., Kumar, S., Rowley, H. A., & Lazebnik, S. (2013). Learning binary codes for high-dimensional data using bilinear projections. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 484–491.

Gong, Y., Lazebnik, S., Gordo, A., & Perronnin, F. (2012). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *35*(12), 2916–2929.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adver-sarial nets. In *Advances in neural information processing systems*, pp. 2672–2680.

Granger, S., & Pennec, X. (2002). Multi-scale em-icp: A fast and robust approach for surface registration. In *Proceedings of the European* *conference on computer vision*, pp. 418–432.

Guo, Y., Bennamoun, M., Sohel, F., Lu, M., Wan, J., & Kwok, N. M. (2016). A comprehensive performance evaluation of 3d local fea-ture descriptors. *International Journal of Computer Vision*, *116*(1), 66–89.

Guo, Y., Sohel, F., Bennamoun, M., Lu, M., & Wan, J. (2013). Rotational projection statistics for 3d local surface description and object recognition. *International Journal of Computer Vision*, *105*(1), 63–86.

Guo, Y., Sohel, F., Bennamoun, M., Wan, J., & Lu, M. (2015). A novel local surface feature for 3d object recognition under clutter and occlusion. *Information Sciences*, *293*, 196–213.

Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE Transac-tions on Image Processing*, *19*(6), 1657–1663.

Gupta, R., Patil, H., & Mittal, A. (2010). Robust order-based methods for feature description. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 334–341.

Han, X., Leung, T., Jia, Y., Sukthankar, R., & Berg, A. C. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 3279–3286.

Han, K., Rezende, R. S., Ham, B., Wong, K. Y. K., Cho, M., Schmid, C., & Ponce, J. (2017). Scnet: Learning semantic correspondence. In *Proceedings of the IEEE international conference on computer* *vision*, pp. 1831–1840.

Harris, C. G., Stephens, M., et al. (1988). A combined corner and edge detector. In *Proceedings of the Alvey vision conference*, pp. 147– 151.

123

International Journal of Computer Vision

Hartmann, W., Havlena, M., & Schindler, K. (2014). Predicting match-ability. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 9–16.

Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: A survey. *Machine Vision and Applications*, *31*(1), 8.

Hayat, N., & Imran, M. (2019). Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter. *Journal of* *Visual Communication and Image Representation*, *62*, 295–308.

He, K., Lu, Y., & Sclaroff, S. (2018). Local descriptors optimized for average precision. In *Proceedings of the IEEE conference on com-puter vision and pattern recognition*, pp. 596–605.

Heikkilä, M., Pietikäinen, M., & Schmid, C. (2009). Description of interest regions with local binary patterns. *Pattern Recognition*, *42*(3), 425–436.

Heinly, J., Dunn, E., & Frahm, J. M. (2012). Comparative evaluation of binary features. In *Proceedings of the European conference on* *computer vision*, pp. 759–773.

Heinly, J., Schonberger, J. L., Dunn, E., & Frahm, J. M. (2015). Recon-structing the world\* in six days\*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 3287–3295.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., & Navab, N. (2012). Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In*Proceedings of the Asian conference on computer vision*, pp. 548–562.

Horaud, R., Forbes, F., Yguel, M., Dewaele, G., & Zhang, J. (2011). Rigid and articulated point registration with expectation condi-tional maximization. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *33*(3), 587–602.

Hu, N., Huang, Q., Thibert, B., & Guibas, L. J. (2018). Distributable consistent multi-object matching. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 2463–2471.

Huang, Q. X., & Guibas, L. (2013). Consistent shape maps via semidef-inite programming. In *Computer graphics forum*, Vol. 32, Wiley Online Library, pp. 177–186.

Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., & Yang, R. (2018). The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition workshops*, pp. 954–960.

Huang, D., Shan, C., Ardabilian, M., Wang, Y., & Chen, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics,* *Part C (Applications and Reviews)*, *41*(6), 765–781.

Iglesias, J. P., Olsson, C., & Kahl, F. (2020). Global optimality for point set registration using semidefinite programming. In *Proceedings of* *the IEEE/CVF conference on computer vision and pattern recog-nition*, pp. 8287–8295.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial trans-former networks. In *Advances in neural information processing* *systems*, pp. 2017–2025.

Jégou, H., Douze, M., & Schmid, C. (2010). Improving bag-of-features for large scale image search. *International Journal of Computer* *Vision*, *87*(3), 316–336.

Jiang, B., Tang, J., Ding, C., & Luo, B. (2017b). Binary constraint preserving graph matching. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 4402–4409.

Jiang, B., Tang, J., Ding, C., Gong, Y., & Luo, B. (2017a). Graph match-ing via multiplicative update algorithm. In *Advances in neural* *information processing systems*, pp. 3187–3195.

Jiang, Z., Wang, T., & Yan, J. (2020b). Unifying offline and online multi-graph matching via finding shortest paths on supergraph. In *IEEE transactions on pattern analysis and machine intelligence*.

Jiang, X., Ma, J., Jiang, J., & Guo, X. (2020a). Robust feature matching using spatial clustering with heavy outliers. *IEEE Transactions on* *Image Processing*, *29*, 736–746.

Jiang, B., Zhao, H., Tang, J., & Luo, B. (2014). A sparse nonnega-tive matrix factorization technique for graph matching problems. *Pattern Recognition*, *47*(2), 736–747.

Jian, B., & Vemuri, B. C. (2011). Robust point set registration using Gaussian mixture models. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *33*(8), 1633–1645.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., & Trulls, E. (2020). Image matching across wide baselines: From paper to practice. arXiv preprint [arXiv:2003.01587](http://arxiv.org/abs/2003.01587).

Johnson, K., Cole-Rhodes, A., Zavorin, I., & Le Moigne, J. (2001). Mutual information as a similarity measure for remote sensing image registration. In *Geo-spatial image and data exploitation II*, pp. 51–61.

Johnson, A. E., & Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on* *Pattern Analysis and Machine Intelligence*, *21*(5), 433–449.

Ke, Y., Sukthankar, R., et al. (2004). Pca-sift: A more distinctive repre-sentation for local image descriptors. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 506–513.

Kedem, D., Tyree, S., Sha, F., Lanckriet, G. R., & Weinberger, K. Q. (2012). Non-linear metric learning. In *Advances in neural infor-mation processing systems*, pp. 2573–2581.

Kezurer, I., Kovalsky, S. Z., Basri, R., & Lipman, Y. (2015). Tight relaxation of quadratic matching. In *Computer graphics forum*, Vol. 34, Wiley Online Library, pp. 115–128.

Khoury, M., Zhou, Q. Y., & Koltun, V. (2017). Learning compact geometric features. In *Proceedings of the IEEE international con-ference on computer vision*, pp. 153–161.

Kim, S., Lin, S., JEON, S. R., Min, D., & Sohn, K. (2018). Recurrent transformer networks for semantic correspondence. In *Advances* *in neural information processing systems*, pp. 6126–6136.

Kim, V.G., Lipman, Y., & Funkhouser, T. (2011). Blended intrinsic maps. In *ACM transactions on graphics*, Vol. 30, ACM, p. 79.

Kim, V. G., Li, W., Mitra, N. J., DiVerdi, S., & Funkhouser, T. A. (2012).

Exploring collections of 3d models using fuzzy correspondences.

*ACM Transactions on Graphics*, *31*(4), 54–1.

Kimmel, R., Zhang, C., Bronstein, A., & Bronstein, M. (2011). Are mser features really interesting? *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *33*(11), 2316–2320.

Kim, S., Min, D., Lin, S., & Sohn, K. (2020). Discrete-continuous trans-formation matching for dense semantic correspondence. *IEEE* *Transactions on Pattern Analysis and Machine Intelligence*, *42*(1),59–73.

Klein, S., Staring, M., & Pluim, J. P. (2007). Evaluation of optimiza-tion methods for nonrigid medical image registration using mutual information and b-splines. *IEEE Transactions on Image Process-ing*, *16*(12), 2879–2890.

Kluger, F., Brachmann, E., Ackermann, H., Rother, C., Yang, M. Y.,

* Rosenhahn, B. (2020). Consac: Robust multi-model fitting by conditional sample consensus. In *Proceedings of the IEEE/CVF* *conference on computer vision and pattern recognition*, pp. 4634–4643.

Komorowski, J., Czarnota, K., Trzcinski, T., Dabala, L., & Lynen, S. (2018). Interest point detectors stability evaluation on apolloscape dataset. In *Proceedings of the European conference on computer* *vision*, pp. 727–739.

Kovnatsky, A., Bronstein, M. M., Bresson, X., & Vandergheynst, P. (2015). Functional correspondence by matrix completion. In *Pro-ceedings of the IEEE conference on computer vision and pattern recognition*, pp. 905–914.

Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F. C., Miao, S., Maier, A. K., Ayache, N., Liao, R., & Kamen, A. (2017).

123

International Journal of Computer Vision

Robust non-rigid registration through agent-based action learning. In *Proceedings of the international conference on medical image* *computing and computer-assisted intervention*, pp. 344–352.

Kulis, B., & Darrell, T. (2009). Learning to hash with binary reconstruc-tive embeddings. In *Advances in neural information processing* *systems*, pp. 1042–1050.

Kulis, B., & Grauman, K. (2009). Kernelized locality-sensitive hashing for scalable image search. In *Proceedings of the IEEE international* *conference on computer vision*, pp. 2130–2137.

Kumar, B., Carneiro, G., Reid, I., et al. (2016). Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 5385–5394.

Laskar, Z., & Kannala, J. (2018). Semi-supervised semantic matching. In *Proceedings of the European conference on computer vision* *workshop*, pp. 1–11.

Lawin, F. J., Danelljan, M., Khan, F., Forssén, P. E., & Felsberg, M. (2018). Density adaptive point set registration. In *Proceedings of* *the IEEE international conference on computer vision*, pp. 3829–3837.

Lawler, E. L. (1963). The quadratic assignment problem. *Management* *Science*, *9*(4), 586–599.

Lazaridis, G., & Petrou, M. (2006). Image registration using the Walsh transform. *IEEE Transactions on Image Processing*, *15*(8), 2343– 2357.

Le Moigne, J., Campbell, W. J., & Cromp, R. F. (2002). An automated parallel image registration technique based on the correlation of wavelet features. *IEEE Transactions on Geoscience and Remote* *Sensing*, *40*(8), 1849–1864.

Lebeda, K., Matas, J., & Chum, O. (2012). Fixing the locally optimized ransac–full experimental evaluation. In *Proceedings of the British* *machine vision conference*, pp. 1–11.

Lee, J., Cho, M., & Lee, K. M. (2010). A graph matching algorithm using data-driven markov chain monte carlo sampling. In *Proceedings* *of the international conference on pattern recognition*, pp. 2816–2819.

Lee, J., Cho, M., & Lee, K. M. (2011). Hyper-graph matching via reweighted random walks. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 1633–1640.

Lee, S., Lim, J., & Suh, I. H. (2020). Progressive feature matching: Incremental graph construction and optimization. In *IEEE trans-actions on image processing*.

Lê-Huu, D. K., & Paragios, N. (2017). Alternating direction graph matching. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 4914–4922.

Lenc, K., & Vedaldi, A. (2014). Large scale evaluation of local image feature detectors on homography datasets. In *Proceedings of the* *British machine vision conference*.

Lenc, K., & Vedaldi, A. (2016). Learning covariant feature detectors. In *Proceedings of the European conference on computer vision*, pp. 100–117.

Leordeanu, M., & Hebert, M. (2005). A spectral technique for corre-spondence problems using pairwise constraints. In *Proceedings of* *the IEEE international conference on computer vision*, pp. 1482–1489.

Leordeanu, M., Hebert, M., & Sukthankar, R. (2009). An integer pro-jected fixed point method for graph matching and map inference. In *Advances in neural information processing systems*, pp. 1114– 1122.

Leordeanu, M., Sukthankar, R., & Hebert, M. (2012). Unsupervised learning for graph matching. *International Journal of Computer* *Vision*, *96*(1), 28–45.

Leutenegger, S., Chli, M., & Siegwart, R. (2011). Brisk: Binary robust invariant scalable keypoints. In *Proceedings of the IEEE interna-tional conference on computer vision*, pp. 2548–2555.

Levi, G. (1973). A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo*, *9*(4), 341.

Li, H., & Hartley, R. (2007). The 3d–3d registration problem revisited. In *Proceedings of the IEEE international conference on computer* *vision*, pp. 1–8.

Li, X., Han, K., Li, S., & Prisacariu, V. A. (2020). Dual-resolution correspondence networks. arXiv preprint [arXiv:2006.08844](http://arxiv.org/abs/2006.08844).

Li, H., Shen, T., & Huang, X. (2009). Global optimization for alignment of generalized shapes. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 856–863.

Li, H., Sumner, R. W., & Pauly, M. (2008). Global correspondence optimization for non-rigid registration of depth scans. In *Computer* *graphics forum*, Vol. 27, Wiley Online Library, pp. 1421–1430.

Lian, W., Zhang, L., & Yang, M. H. (2017). An efficient globally optimal algorithm for asymmetric point matching. *IEEE Transactions on* *Pattern Analysis and Machine Intelligence*, *39*(7), 1281–1293.

Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., & Comaniciu, D. (2017). An artificial agent for robust image registration. In *Proceedings of the thirty-first AAAI conference on* *artificial intelligence*, pp. 4168–4175.

Liao, Q., Sun, D., & Andreasson, H. (2020). Point set registration for 3d range scans using fuzzy cluster-based metric and efficient global optimization. In *IEEE transactions on pattern analysis and* *machine intelligence*.

Li, X., & Hu, Z. (2010). Rejecting mismatches by correspondence func-tion. *International Journal of Computer Vision*, *89*(1), 1–17.

Li, Z., Mahapatra, D., Tielbeek, J. A., Stoker, J., van Vliet, L. J., & Vos, F. M. (2015). Image registration based on autocorrelation of local structure. *IEEE Transactions on Image Processing*, *35*(1), 63–75.

Lin, W. Y. D., Cheng, M. M., Lu, J., Yang, H., Do, M. N., & Torr, P. (2014). Bilateral functions for global motion modeling. In *Pro-ceedings of the European conference on computer vision*, pp.341–356.

Lin, W. Y., Liu, S., Jiang, N., Do, M. N., Tan, P., & Lu, J. (2016b). Repmatch: Robust feature matching and pose for reconstructing modern cities. In *Proceedings of the European conference on com-puter vision*, pp. 562–579.

Lin, W. Y., Liu, S., Matsushita, Y., Ng, T. T., & Cheong, L. F. (2011). Smoothly varying affine stitching. In *Proceedings of the IEEE con-ference on computer vision and pattern recognition*, pp. 345–352.

Lin, K., Lu, J., Chen, C. S., & Zhou, J. (2016a). Learning compact binary descriptors with unsupervised deep neural networks. In *Pro-ceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1183–1192.

Lindeberg, T. (1998). Feature detection with automatic scale selection.

*International Journal of Computer Vision*, *30*(2), 79–116.

Lin, W. Y., Wang, F., Cheng, M. M., Yeung, S. K., Torr, P. H., Do, M. N., et al. (2017). Code: Coherence based decision boundaries for feature correspondence. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *40*(1), 34–47.

Lipman, Y., & Funkhouser, T. (2009). Möbius voting for surface corre-spondence. *ACM Transactions on Graphics*, *28*(3), 72.

Lipman, Y., Yagev, S., Poranne, R., Jacobs, D. W., & Basri, R. (2014). Feature matching with bounded distortion. *ACM Transactions on* *Graphics*, *33*(3), 26.

Litany, O., Remez, T., Rodolà, E., Bronstein, A., & Bronstein, M. (2017). Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE international* *conference on computer vision*, pp. 5659–5667.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Litman, R., & Bronstein, A. M. (2014). Learning spectral descriptors for deformable shape correspondence. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *36*(1), 171–180.

123

International Journal of Computer Vision

Liu, H., & Yan, S. (2010). Common visual pattern discovery via spatially coherent correspondences. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 1609–1616.

Liu, Y., & Zhang, H. (2012). Indexing visual features: Real-time loop closure detection using a tree structure. In *Proceedings of the IEEE* *international conference on robotics and automation*, pp. 3613–3618.

Liu, Y., Feng, R., & Zhang, H. (2015a). Keypoint matching by outlier pruning with consensus constraint. In *Proceedings of the IEEE* *international conference on robotics and automation*, pp. 5481–5486.

Liu, W., Wang, J., Ji, R., Jiang, Y. G., & Chang, S. F. (2012a). Supervised hashing with kernels. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 2074–2081.

Liu, Y., Wang, C., Song, Z., & Wang, M. (2018b). Efficient global point cloud registration by matching rotation invariant features through translation search. In *Proceedings of the European conference on* *computer vision*, pp. 448–463.

Liu, R., Yang, C., Sun, W., Wang, X., & Li, H. (2020). Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recogni-tion*, pp. 12,757–12,766.

Liu, X., Ai, Y., Zhang, J., & Wang, Z. (2018a). A novel affine and con-trast invariant descriptor for infrared and visible image registration. *Remote Sensing*, *10*(4), 658.

Liu, Y., Chen, X., Peng, H., & Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. *Information* *Fusion*, *36*, 191–207.

Liu, H., Guo, B., & Feng, Z. (2005). Pseudo-log-polar Fourier transform for image registration. *IEEE Signal Processing Letters*, *13*(1), 17– 20.

Liu, Y., Liu, S., & Wang, Z. (2015b). Multi-focus image fusion with dense sift. *Information Fusion*, *23*, 139–155.

Liu, M., Pradalier, C., & Siegwart, R. (2013). Visual homing from scale with an uncalibrated omnidirectional camera. *IEEE Transactions* *on Robotics*, *29*(6), 1353–1365.

Liu, Z. Y., & Qiao, H. (2014). GNCCP–graduated nonconvexityand concavity procedure. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *36*(6), 1258–1267.

Liu, Z. Y., Qiao, H., & Xu, L. (2012b). An extended path following algo-rithm for graph-matching problem. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *34*(7), 1451–1456.

Li, Y., Wang, S., Tian, Q., & Ding, X. (2015). A survey of recent advances in visual feature detection. *Neurocomputing*, *149*, 736– 751.

Loeckx, D., Slagmolen, P., Maes, F., Vandermeulen, D., & Suetens, P. (2009). Nonrigid image registration using conditional mutual information. *IEEE Transactions on Image*, *29*(1), 19–29.

Loiola, E. M., de Abreu, N. M. M., Boaventura-Netto, P. O., Hahn, P., & Querido, T. (2007). A survey for the quadratic assignment problem. *European Journal of Operational Research*, *176*(2), 657–690.

Lowe, D.G., et al. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE international conference on* *computer vision*, pp. 1150–1157.

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

Lowry, S., & Andreasson, H. (2018). Logos: Local geometric support for high-outlier spatial verification. In *Proceedings of the IEEE* *international conference on robotics and automation*, pp. 7262–7269.

Luo, W., Schwing, A. G., & Urtasun, R. (2016). Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 5695–5703.

Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., & Quan, L. (2019). Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 2527–2536.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., & Quan, L. (2020). Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on* *computer vision and pattern recognition*, pp. 6589–6598.

Ma, J., Zhao, J., Jiang, J., Zhou, H., Zhou, Y., Wang, Z., & Guo, X. (2018b). Visual homing via guided locality preserving matching. In *Proceedings of the IEEE international conference on robotics* *and automation*, pp. 7254–7261.

Ma, J., Zhao, J., Tian, J., Tu, Z., & Yuille, A. L. (2013b). Robust esti-mation of nonrigid transformation for point set registration. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 2147–2154.

Ma, J., Chen, C., Li, C., & Huang, J. (2016a). Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, *31*, 100–109.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Image*, *16*(2), 187–198.

Mainali, P., Lafruit, G., Yang, Q., Geelen, B., Van Gool, L., & Lauw-ereins, R. (2013). Sifer: Scale-invariant feature detector with error resilience. *International Journal of Computer Vision*, *104*(2), 172– 197.

Mair, E., Hager, G. D., Burschka, D., Suppa, M., & Hirzinger, G. (2010). Adaptive and generic corner detection based on the accelerated segment test. In *Proceedings of the European conference on com-puter vision*, pp. 183–196.

Maiseli, B., Gu, Y., & Gao, H. (2017). Recent developments and trends in point set registration methods. *Journal of Visual Communication* *and Image Representation*, *46*, 95–106.

Ma, J., Jiang, X., Jiang, J., Zhao, J., & Guo, X. (2019a). LMR: Learning a two-class classifier for mismatch removal. *IEEE Transactions on* *Image Processing*, *28*(8), 4045–4059.

Ma, J., Jiang, J., Liu, C., & Li, Y. (2017a). Feature guided Gaussian mixture model with semi-supervised em and local geometric con-straint for retinal image registration. *Information Sciences*, *417*, 128–142.

Ma, J., Jiang, J., Zhou, H., Zhao, J., & Guo, X. (2018a). Guided locality preserving feature matching for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(8),4435–4447.

Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., et al. (2020). Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, *54*, 85–98.

Ma, J., Qiu, W., Zhao, J., Ma, Y., Yuille, A. L., & Tu, Z. (2015). Robust *l*2*e* estimation of transformation for non-rigid registration. *IEEE Transactions on Signal Processing*, *63*(5), 1115–1129.

Marimon, D., Bonnin, A., Adamek, T., & Gimeno, R. (2010). Darts: Efficient scale-space extraction of daisy keypoints. In *Proceedings* *of the IEEE conference on computer vision and pattern recogni-tion*, pp. 2416–2423.

Maron, H., & Lipman, Y. (2018). (probably) concave graph matching. In *Advances in Neural information processing systems*, pp. 406–416.

Maron, H., Dym, N., Kezurer, I., Kovalsky, S., & Lipman, Y. (2016). Point registration via efficient convex relaxation. *ACM Transac-tions on Graphics*, *35*(4), 73.

Masci, J., Boscaini, D., Bronstein, M., & Vandergheynst, P. (2015). Geodesic convolutional neural networks on Riemannian mani-folds. In *Proceedings of the IEEE international conference on* *computer vision workshops*, pp. 37–45.

Masood, A., & Sarfraz, M. (2007). Corner detection by sliding rectan-gles along planar curves. *Computers & Graphics*, *31*(3), 440–448.

123

International Journal of Computer Vision

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image* *and Vision Computing*, *22*(10), 761–767.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., et al. (2017b). Remote sensing image registration with modified sift and enhanced feature matching. *IEEE Geoscience and Remote Sensing Letters*, *14*(1), 3–7.

Ma, J., Wu, J., Zhao, J., Jiang, J., Zhou, H., & Sheng, Q. Z. (2019b). Nonrigid point set registration with robust transformation learn-ing under manifold regularization. *IEEE Transactions on Neural* *Networks and Learning Systems*, *30*(12), 3584–3597.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional net-works for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 4040–4048.

Ma, J., Yu, W., Liang, P., Li, C., & Jiang, J. (2019c). Fusiongan: A gen-erative adversarial network for infrared and visible image fusion. *Information Fusion*, *48*, 11–26.

Ma, J., Zhao, J., Jiang, J., Zhou, H., & Guo, X. (2019d). Locality preserv-ing matching. *International Journal of Computer Vision*, *127*(5), 512–531.

Ma, J., Zhao, J., Ma, Y., & Tian, J. (2015a). Non-rigid visible and infrared face registration via regularized gaussian fields criterion. *Pattern Recognition*, *48*(3), 772–784.

Ma, J., Zhao, J., Tian, J., Bai, X., & Tu, Z. (2013a). Regularized vector field learning with sparse approximation for mismatch removal. *Pattern Recognition*, *46*(12), 3519–3532.

Ma, J., Zhao, J., Tian, J., Yuille, A. L., & Tu, Z. (2014). Robust point matching via vector field consensus. *IEEE Transactions on Image* *Processing*, *23*(4), 1706–1721.

Ma, J., Zhao, J., & Yuille, A. L. (2016b). Non-rigid point set registration by preserving global and local structures. *IEEE Transactions on* *Image Processing*, *25*(1), 53–64.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., & Tian, J. (2015b). Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Transactions on Geoscience and Remote* *Sensing*, *53*(12), 6469–6481.

Mian, A., Bennamoun, M., & Owens, R. (2010). On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, *89*(2–3), 348–361.

Miao, S., Piat, S., Fischer, P., Tuysuzoglu, A., Mewes, P., Mansi, T., & Liao, R. (2018). Dilated fcn for multi-agent 2d/3d medical image registration. In *Proceedings of the thirty-second AAAI conference* *on artificial intelligence*, pp. 4694–4701.

Mikolajczyk, K., & Schmid, C. (2001). Indexing based on scale invari-ant interest points. In *Proceedings of the IEEE international* *conference on computer vision*, pp. 525–531.

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

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *27*(10), 1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., et al. (2005). A comparison of affine region detec-tors. *International Journal of Computer Vision*, *65*(1–2), 43–72.

Mishchuk, A., Mishkin, D., Radenovic, F., & Matas, J. (2017). Working hard to know your neighbor’s margins: Local descriptor learning loss. In *Advances in neural information processing systems*, pp. 4826–4837.

Mishkin, D., Radenovic, F., & Matas, J. (2017). Learning dis-criminative affine regions via discriminability. arXiv preprint [arXiv:1711.06704](http://arxiv.org/abs/1711.06704).

Mishkin, D., Radenovic, F., & Matas, J. (2018). Repeatability is not enough: Learning affine regions via discriminability. In *Proceed-ings of the European conference on computer vision*, pp. 284–300.

Mitra, R., Doiphode, N., Gautam, U., Narayan, S., Ahmed, S., Chan-dran, S., & Jain, A. (2018). A large dataset for improving patch matching. arXiv preprint [arXiv:1801.01466](http://arxiv.org/abs/1801.01466).

Mok, T. C., & Chung, A. (2020). Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of* *the IEEE/CVF conference on computer vision and pattern recog-nition*, pp. 4644–4653.

Mokhtarian, F., & Suomela, R. (1998). Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Anal-ysis and Machine Intelligence*, *20*(12), 1376–1381.

Möller, R., Krzykawski, M., & Gerstmayr, L. (2010). Three 2d-warping schemes for visual robot navigation. *Autonomous Robots*, *29*(3–4), 253–291.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 5115–5124.

Moo Yi, K., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., & Fua, P. (2018). Learning to find good correspondences. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition*,pp. 2666–2674.

Moo Yi, K., Verdie, Y., Fua, P., & Lepetit, V. (2016). Learning to assign orientations to feature points. In *Proceedings of the IEEE confer-ence on computer vision and pattern recognition*, pp. 107–116.

Moravec, H. P. (1977). *Techniques towards automatic visual obstacle* *avoidance*.

Morel, J. M., & Yu, G. (2009). Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, *2*(2), 438–469.

Mukherjee, D., Wu, Q. J., & Wang, G. (2015). A comparative experi-mental study of image feature detectors and descriptors. *Machine* *Vision and Applications*, *26*(4), 443–466.

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. (2015). ORB-SLAM: A versatile and accurate monocular slam system. *IEEE Transac-tions on Robotics*, *31*(5), 1147–1163.

Mustafa, A., Kim, H., & Hilton, A. (2018). Msfd: Multi-scale segmentation-based feature detection for wide-baseline scene reconstruction. *IEEE Transactions on Image Processing*, *28*(3), 1118–1132.

Myronenko, A., & Song, X. (2010). Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine* *Intelligence*, *32*(12), 2262–2275.

Nasuto, D., & Craddock, J. B. R. (2002). Napsac: High noise, high dimensional robust estimation-it’s in the bag. In *Proceedings of* *the British machine vision conference*, pp. 458–467.

Ni, K., Jin, H., & Dellaert, F. (2009). Groupsac: Efficient consensus in the presence of groupings. In *Proceedings of the IEEE interna-tional conference on computer vision*, pp. 2193–2200.

Norouzi, M., & Blei, D. M. (2011). Minimal loss hashing for compact binary codes. In *Proceedings of the international conference on* *machine learning*, pp. 353–360.

Nüchter, A., Lingemann, K., Hertzberg, J., & Surmann, H. (2007). 6d SLAM–3d mapping outdoor environments. *Journal of Field* *Robotics*, *24*(8–9), 699–722.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine* *Intelligence*, *24*(7), 971–987.

Ono, Y., Trulls, E., Fua, P., & Yi, K. M. (2018). LF-NET: Learning local features from images. In *Advances in neural information* *processing systems*, pp. 6234–6244.

123

International Journal of Computer Vision

Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., & Guibas, L. (2012). Functional maps: A flexible representation of maps between shapes. *ACM Transactions on Graphics*, *31*(4), 30.

Pachauri, D., Kondor, R., & Singh, V. (2013). Solving the multi-way matching problem by permutation synchronization. In *Advances* *in neural information processing systems*, pp. 1860–1868.

Pais, G. D., Ramalingam, S., Govindu, V. M., Nascimento, J. C., Chel-lappa, R., & Miraldo, P. (2020). 3dregnet: A deep neural network for 3d point registration. In *Proceedings of the IEEE/CVF confer-ence on computer vision and pattern recognition*, pp. 7193–7203.

Pan, W. H., Wei, S. D., & Lai, S. H. (2008). Efficient NCC-based image matching in Walsh-Hadamard domain. In *Proceedings of the Euro-pean conference on computer vision*, pp. 468–480.

Pang, J., Sun, W., Ren, J. S., Yang, C., & Yan, Q. (2017). Cascade resid-ual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE international conference on* *computer vision*, pp. 887–895.

Papazov, C., & Burschka, D. (2011). Stochastic global optimization for robust point set registration. *Computer Vision and Image Under-standing*, *115*(12), 1598–1609.

Park, J., Zhou, Q. Y., & Koltun, V. (2017). Colored point cloud registra-tion revisited. In *Proceedings of the IEEE international conference* *on computer vision*, pp. 143–152.

Parra Bustos, A., Chin, T. J., & Suter, D. (2014). Fast rotation search with stereographic projections for 3d registration. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.3930–3937.

Paul, S., & Pati, U. C. (2016). Remote sensing optical image registration using modified uniform robust sift. *IEEE Geoscience and Remote* *Sensing Letters*, *13*(9), 1300–1304.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *12*(7), 629–639.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 1–8.

Piasco, N., Sidibé, D., Demonceaux, C., & Gouet-Brunet, V. (2018). A survey on visual-based localization: On the benefit of heteroge-neous data. *Pattern Recognition*, *74*, 90–109.

Pilet, J., Lepetit, V., & Fua, P. (2008). Fast non-rigid surface detection, registration and realistic augmentation. *International Journal of* *Computer Vision*, *76*(2), 109–122.

Pinheiro, A. M., & Ghanbari, M. (2010). Piecewise approximation of contours through scale-space selection of dominant points. *IEEE* *Transactions on Image Processing*, *19*(6), 1442–1450.

Plötz, T., & Roth, S. (2018). Neural nearest neighbors networks. In *Advances in Neural information processing systems*, pp. 1087–1098.

Poggi, M., Pallotti, D., Tosi, F., & Mattoccia, S. (2019). Guided stereo matching. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 979–988.

Pohl, C., & Van Genderen, J. L. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applica-tions. *International Journal of Remote Sensing*, *19*(5), 823–854.

Pokrass, J., Bronstein, A. M., Bronstein, M. M., Sprechmann, P., & Sapiro, G. (2013). Sparse modeling of intrinsic correspondences. In *Computer graphics forum*, Vol. 32, Wiley Online Library, pp. 459–468.

Pomerleau, F., Colas, F., Siegwart, R., & Magnenat, S. (2013). Com-paring ICP variants on real-world data sets. *Autonomous Robots*, *34*(3), 133–148.

Poursaeed, O., Yang, G., Prakash, A., Fang, Q., Jiang, H., Hariharan, B.,

* Belongie, S. (2018). Deep fundamental matrix estimation with-out correspondences. In *Proceedings of the European conference* *on computer vision workshop*, pp. 1–13.

Qi, C.R., Su, H., Mo, K., & Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 652–660.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pp. 5099–5108.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Raguram, R., Chum, O., Pollefeys, M., Matas, J., & Frahm, J. M. (2012). USAC: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 2022–2038.

Ramer, U. (1972). An iterative procedure for the polygonal approxima-tion of plane curves. *Computer Graphics and Image Processing*, *1*(3), 244–256.

Ramisa, A., Goldhoorn, A., Aldavert, D., Toledo, R., & de Mantaras, R. L. (2011). Combining invariant features and the ALV hom-ing method for autonomous robot navigation based on panoramas. *Journal of Intelligent & Robotic Systems*, *64*(3–4), 625–649.

Ranftl, R., & Koltun, V. (2018). Deep fundamental matrix estimation. In *Proceedings of the European conference on computer vision*, pp. 284–299.

Reddy, B. S., & Chatterji, B. N. (1996). An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE* *Transactions on Image Processing*, *5*(8), 1266–1271.

Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., & Humenberger, M. (2019). R2d2: Repeatable and reliable detector and descriptor. arXiv preprint [arXiv:1906.06195](http://arxiv.org/abs/1906.06195).

Revaud, J., Weinzaepfel, P., Harchaoui, Z., & Schmid, C. (2016). Deep-matching: Hierarchical deformable dense matching. *International* *Journal of Computer Vision*, *120*(3), 300–323.

Richardson, A., & Olson, E. (2013). Learning convolutional filters for interest point detection. In *Proceedings of the IEEE international* *conference on robotics and automation*, pp. 631–637.

Robertson, C., & Fisher, R. B. (2002). Parallel evolutionary registration of range data. *Computer Vision and Image Understanding*, *87*(1– 3), 39–50.

Rocco, I., Arandjelovic, R., & Sivic, J. (2017). Convolutional neural network architecture for geometric matching. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition*,pp. 6148–6157.

Rocco, I., Cimpoi, M., Arandjelovi´c, R., Torii, A., Pajdla, T., & Sivic, J. (2018). Neighbourhood consensus networks. In *Advances in* *neural information processing systems*, pp. 1651–1662.

Rodola, E., Bronstein, A. M., Albarelli, A., Bergamasco, F., & Torsello, A. (2012). A game-theoretic approach to deformable shape match-ing. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 182–189.

Rodolà, E., Cosmo, L., Bronstein, M. M., Torsello, A., & Cremers, D. (2017). Partial functional correspondence. In *Computer graphics* *forum*, Vol. 36, Wiley Online Library, pp. 222–236.

Rodolà, E., Rota Bulo, S., Windheuser, T., Vestner, M., & Cremers, D. (2014). Dense non-rigid shape correspondence using random forests. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 4177–4184.

Rodola, E., Torsello, A., Harada, T., Kuniyoshi, Y., & Cremers, D. (2013). Elastic net constraints for shape matching. In *Proceed-ings of the IEEE international conference on computer vision*, pp.1169–1176.

Rosenfeld, A., & Weszka, J. S. (1975). An improved method of angle detection on digital curves. *IEEE Transactions on Computers*, *100*(9), 940–941.

123

International Journal of Computer Vision

Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *Proceedings of the European conference on* *computer vision*, pp. 430–443.

Rosten, E., Porter, R., & Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *32*(1), 105–119.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. R. (2011). Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE* *international conference on computer vision*, pp. 2564–2571.

Rustamov, R. M. (2007). Laplace-Beltrami eigenfunctions for defor-mation invariant shape representation. In *Proceedings of the* *Eurographics symposium on geometry processing*, pp. 225–233.

Rusu, R. B., Blodow, N., & Beetz, M. (2009). Fast point feature his-tograms (fpfh) for 3d registration. In *Proceedings of the IEEE* *international conference on robotics and automation*, pp. 3212–3217.

Rusu, R. B., Blodow, N., Marton, Z. C., & Beetz, M. (2008). Aligning point cloud views using persistent feature histograms. In *Proceed-ings of the IEEE/RSJ international conference on intelligent robots and systems*, pp. 3384–3391.

Sahillioglu, Y., & Yemez, Y. (2011). Coarse-to-fine combinatorial matching for dense isometric shape correspondence. In *Computer* *graphics forum*, Vol. 30, Wiley Online Library, pp. 1461–1470.

Salakhutdinov, R., & Hinton, G. (2009). Semantic hashing. *Interna-tional Journal of Approximate Reasoning*, *50*(7), 969–978.

Salti, S., Lanza, A., & Di Stefano, L. (2013). Keypoints from symmetries by wave propagation. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 2898–2905.

Salti, S., Tombari, F., Spezialetti, R., & Di Stefano, L. (2015). Learn-ing a descriptor-specific 3d keypoint detector. In *Proceedings of* *the IEEE international conference on computer vision*, pp. 2318–2326.

Sandhu, R., Dambreville, S., & Tannenbaum, A. (2010). Point set registration via particle filtering and stochastic dynamics. *IEEE* *Transactions on Pattern Analysis and Machine Intelligence*, *32*(8),1459–1473.

Sarlin, P.E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision* *and pattern recognition*, pp. 4938–4947.

Savinov, N., Seki, A., Ladicky, L., Sattler, T., & Pollefeys, M. (2017). Quad-networks: Unsupervised learning to rank for interest point detection. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 1822–1830.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. In *TNN*.

Schellewald, C., & Schnörr, C. (2005). Probabilistic subgraph matching based on convex relaxation. In *Proceedings of the international* *workshop on energy minimization methods in computer vision and pattern recognition*, pp. 171–186.

Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revis-ited. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 4104–4113.

Schonberger, J. L., Hardmeier, H., Sattler, T., & Pollefeys, M. (2017). Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 1482–1491.

Schroeter, D., & Newman, P. (2008). On the robustness of visual hom-ing under landmark uncertainty. In *Proceedings of the intelligent* *autonomous systems*, pp. 278–287.

Scott, G. L., & Longuet-Higgins, H. C. (1991). An algorithm for associ-ating the features of two images. *Proceedings of the Royal Society* *of London. Series B: Biological Sciences*, *244*(1309), 21–26.

Shah, R., Srivastava, V., & Narayanan, P. (2015). Geometry-aware fea-ture matching for structure from motion applications. In *Proceed-*

*ings of the IEEE winter conference on applications of computer vision*, pp. 278–285.

Shaked, A., & Wolf, L. (2017). Improved stereo matching with con-stant highway networks and reflective confidence learning. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 4641–4650.

Shakhnarovich, G. (2005). Learning task-specific similarity. Ph.D. the-sis, Massachusetts Institute of Technology.

Shapiro, L. S., & Brady, J. M. (1992). Feature-based correspondence: An eigenvector approach. *Image and Vision Computing*, *10*(5), 283–288.

Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., Cheng, M., & He, Z. (2019). RF-NET: An end-to-end image matching network based on receptive field. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 8132–8140.

Shi, J., & Tomasi, C. (1993). Good features to track. Technical report, Cornell University.

Silva, L., Bellon, O. R. P., & Boyer, K. L. (2005). Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *27*(5), 762–776.

Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., & Komodakis, N. (2016). A deep metric for multimodal registra-tion. In *Proceedings of the international conference on medical* *image computing and computer-assisted intervention*, pp. 10–18.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Learning local feature descriptors using convex optimization. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *36*(8), 1573–1585.

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., & Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international* *conference on computer vision*, pp. 118–126.

Sipiran, I., & Bustos, B. (2011). Harris 3d: A robust extension of the Har-ris operator for interest point detection on 3d meshes. *The Visual* *Computer*, *27*(11), 963.

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE* *international conference on computer vision*, pp. 1–8.

Smith, S. M., & Brady, J. M. (1997). Susan: A new approach to low level image processing. *International Journal of Computer Vision*, *23*(1), 45–78.

Sofka, M., Yang, G., & Stewart, C. V. (2007). Simultaneous covariance driven correspondence (CDC) and transformation estimation in the expectation maximization framework. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1–8.

Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B. P., Isgum, I., & Star-ing, M. (2017). Nonrigid image registration using multi-scale 3d convolutional neural networks. In *Proceedings of the international* *conference on medical image computing and computer-assisted intervention*, pp. 232–239.

Sotiras, A., Davatzikos, C., & Paragios, N. (2013). Deformable med-ical image registration: A survey. *IEEE Transactions on Medical* *Imaging*, *32*(7), 1153.

Strecha, C., Lindner, A., Ali, K., & Fua, P. (2009). Training for task spe-cific keypoint detection. In *Joint pattern recognition symposium*, Springer, pp. 151–160.

Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., & Thoennessen, U. (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings of the IEEE* *Conference on computer vision and pattern recognition*, pp. 1–8.

Strecha, C., Bronstein, A., Bronstein, M., & Fua, P. (2012). Ldahash: Improved matching with smaller descriptors. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *34*(1), 66–78.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of RGB-D slam systems. In *Pro-*

123

International Journal of Computer Vision

*ceedings of the IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580.

Suh, Y., Cho, M., & Lee, K. M. (2012). Graph matching via sequen-tial Monte Carlo. In *Proceedings of the European conference on* *computer vision*, pp. 624–637.

Sun, J., Ovsjanikov, M., & Guibas, L. (2009). A concise and prov-ably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, Vol. 28, Wiley Online Library, pp. 1383–1392.

Sweeney, C., Hollerer, T., & Turk, M. (2015). Theia: A fast and scalable structure-from-motion library. In *Proceedings of the ACM inter-national conference on multimedia*, pp. 693–696.

Swoboda, P., Kuske, J., & Savchynskyy, B. (2017). A dual ascent frame-work for Lagrangean decomposition of combinatorial problems. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 1596–1606.

Swoboda, P., Mokarian, A., Theobalt, C., Bernard, F., et al. (2019). A convex relaxation for multi-graph matching. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.11,156–11,165.

Swoboda, P., Rother, C., Abu Alhaija, H., Kainmuller, D., & Savchyn-skyy, B. (2017). A study of lagrangean decompositions and dual ascent solvers for graph matching. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1607–1616.

Takita, K., Aoki, T., Sasaki, Y., Higuchi, T., & Kobayashi, K. (2003). High-accuracy subpixel image registration based on phase-only correlation. *IEICE Transactions on Fundamentals of Electronics,* *Communications and Computer Sciences*, *86*(8), 1925–1934.

Tang, F., Lim, S. H., Chang, N. L., & Tao, H. (2009). A novel feature descriptor invariant to complex brightness changes. In *Proceedings* *of the IEEE conference on computer vision and pattern recogni-tion*, pp. 2631–2638.

Tevs, A., Berner, A., Wand, M., Ihrke, I., & Seidel, H. P. (2011). Intrin-sic shape matching by planned landmark sampling. In *Computer* *graphics forum*, Vol. 30, Wiley Online Library, pp. 543–552.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73.

Tian, Y., Fan, B., & Wu, F. (2017). L2-net: Deep learning of discrimi-native patch descriptor in Euclidean space. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.661–669.

Tian, Y., Yan, J., Zhang, H., Zhang, Y., Yang, X., & Zha, H. (2012). On the convergence of graph matching: Graduated assignment revisited. In *Proceedings of the European conference on computer* *vision*, pp. 821–835.

Tian, Y., Yu, X., Fan, B., Wu, F., Heijnen, H., & Balntas, V. (2019). Sosnet: Second order similarity regularization for local descrip-tor learning. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 11,016–11,025.

Toews, M., & Wells, W. (2009). Sift-rank: Ordinal description for invari-ant feature correspondence. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 172–177.

Tola, E., Lepetit, V., & Fua, P. (2010). Daisy: An efficient dense descrip-tor applied to wide-baseline stereo. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *32*(5), 815–830.

Tombari, F., Salti, S., & Di Stefano, L. (2010a). Unique shape context for 3d data description. In *Proceedings of the ACM workshop on* *3D object retrieval*, pp. 57–62.

Tombari, F., Salti, S., & Di Stefano, L. (2010b). Unique signatures of histograms for local surface description. In *Proceedings of the* *European conference on computer vision*, pp. 356–369.

Tombari, F., Salti, S., & Di Stefano, L. (2013). Performance evalua-tion of 3d keypoint detectors. *International Journal of Computer* *Vision*, *102*(1–3), 198–220.

Torr, P. H. (2003). Solving Markov random fields using semi definite programming. In *Proceeding of AISTATS*, pp. 1–8.

Torr, P., & Zisserman, A. (1998). Robust computation and parametriza-tion of multiple view relations. In *Proceedings of the international* *conference on computer vision*, pp. 727–732.

Torresani, L., Kolmogorov, V., & Rother, C. (2012). A dual decompo-sition approach to feature correspondence. *IEEE Transactions on* *Pattern Analysis and Machine Intelligence*, *35*(2), 259–271.

Torr, P. H., & Zisserman, A. (2000). Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision* *and Image Understanding*, *78*(1), 138–156.

Trajkovi´c, M., & Hedley, M. (1998). Fast corner detection. *Image and* *Vision Computing*, *16*(2), 75–87.

Tron, R., Zhou, X., Esteves, C., & Daniilidis, K. (2017). Fast multi-image matching via density-based clustering. In *Proceedings of* *the IEEE international conference on computer vision*, pp. 4057–4066.

Truong, P., Danelljan, M., & Timofte, R. (2020). Glu-net: Global-local universal network for dense flow and correspondences. In *Proceed-ings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6258–6268.

Trzcinski, T., & Lepetit, V. (2012). Efficient discriminative projections for compact binary descriptors. In *Proceedings of the European* *conference on computer vision*, pp. 228–242.

Trzcinski, T., Christoudias, M., Fua, P., & Lepetit, V. (2013). Boosting binary keypoint descriptors. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 2874–2881.

Trzcinski, T., Christoudias, M., Lepetit, V., & Fua, P. (2012). Learning image descriptors with the boosting-trick. In *Advances in neural* *information processing systems*, pp. 269–277.

Trzcinski, T., Christoudias, M., & Lepetit, V. (2014). Learning image descriptors with boosting. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *37*(3), 597–610.

Tsin, Y., & Kanade, T. (2004). A correlation-based approach to robust point set registration. In *Proceedings of the European conference* *on computer vision*, pp. 558–569.

Tuytelaars, T., & Van Gool, L. (2004). Matching widely separated views based on affine invariant regions. *International Journal of Com-puter Vision*, *59*(1), 61–85.

Tuytelaars, T., Mikolajczyk, K., et al. (2008). Local invariant fea-ture detectors: A survey. *Foundations and Trends® in Computer* *Graphics and Vision*, *3*(3), 177–280.

Ufer, N., & Ommer, B. (2017). Deep semantic feature matching. In *Pro-ceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6914–6923.

Umeyama, S. (1988). An eigen decomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *10*(5), 695–703.

Unnikrishnan, R., & Hebert, M. (2008). Multi-scale interest regions from unorganized point clouds. In *Proceedings of the IEEE con-ference on computer vision and pattern recognition workshops*,pp. 1–8.

van Wyk, B. J., & van Wyk, M. A. (2004). A POCS-based graph matching algorithm. *IEEE Transactions on Pattern Analysis and* *Machine Intelligence*, *26*(11), 1526–1530.

Van Kaick, O., Zhang, H., Hamarneh, G., & Cohen-Or, D. (2011). A survey on shape correspondence. In *Computer graphics forum*, Vol. 30, Wiley Online Library, pp. 1681–1707.

Verdie, Y., Yi, K., Fua, P., & Lepetit, V. (2015). Tilde: A temporally invariant learned detector. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 5279–5288.

Vongkulbhisal, J., De la Torre, F., & Costeira, J. P. (2017). Dis-criminative optimization: Theory and applications to point cloud registration. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 4104–4112.

123

International Journal of Computer Vision

Vongkulbhisal, J., Irastorza Ugalde, B., De la Torre, F., & Costeira, J. P. (2018). Inverse composition discriminative optimization for point cloud registration. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*, pp. 2993–3001.

Wang, J., & Zhang, M. (2020). Deepflash: An efficient network for learning-based medical image registration. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recognition*,pp. 4444–4452.

Wang, C., Bronstein, M. M., Bronstein, A. M., & Paragios, N. (2011). Discrete minimum distortion correspondence problems for non-rigid shape matching. In *Proceedings of the international con-ference on scale space and variational methods in computer vision*,pp. 580–591.

Wang, Z., Fan, B., & Wu, F. (2011). Local intensity order pattern for feature description. In *Proceedings of the international conference* *on computer vision*, pp. 603–610.

Wang, H., Guo, J., Yan, D. M., Quan, W., & Zhang, X. (2018b). Learning 3d keypoint descriptors for non-rigid shape matching. In *Proceed-ings of the European conference on computer vision*, pp. 3–19.

Wang, J., Kumar, S., & Chang, S. F. Semi-supervised hashing for scal-able image retrieval. In *Proceedings of the IEEE conference on* *computer vision and pattern recognition*.

Wang, T., Liu, H., Li, Y., Jin, Y., Hou, X., & Ling, H. (2020). Learning combinatorial solver for graph matching. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recognition*,pp. 7568–7577.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., & Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 1386–1393.

Wang, G., Wang, Z., Chen, Y., Zhou, Q., & Zhao, W. (2016). Context-aware Gaussian fields for non-rigid point set registration. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 5811–5819.

Wang, F., Xue, N., Yu, J. G., & Xia, G. S. (2020). Zero-assignment constraint for graph matching with outliers. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recognition*,pp. 3033–3042.

Wang, F., Xue, N., Zhang, Y., Bai, X., & Xia, G. S. (2018a). Adaptively transforming graph matching. In *Proceedings of the European con-ference on computer vision*, pp. 625–640.

Wang, R., Yan, J., & Yang, X. (2019). Learning combinatorial embed-ding networks for deep graph matching. In *ICCV*.

Wang, Q., Zhou, X., & Daniilidis, K. (2018). Multi-image semantic matching by mining consistent features. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.685–694.

Wang, J., Zhou, F., Wen, S., Liu, X., & Lin, Y. (2017). Deep metric learning with angular loss. In *Proceedings of the IEEE interna-tional conference on computer vision*, pp. 2593–2601.

Wang, Z., Fan, B., Wang, G., & Wu, F. (2015). Exploring local and overall ordinal information for robust feature description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(11), 2198–2211.

Wang, G., Wang, Z., Chen, Y., & Zhao, W. (2015). Robust point match-ing method for multimodal retinal image registration. *Biomedical* *Signal Processing and Control*, *19*, 68–76.

Wei, L., Huang, Q., Ceylan, D., Vouga, E., & Li, H. (2016). Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pat-tern recognition*, pp. 1544–1553.

Wei, X., Zhang, Y., Gong, Y., & Zheng, N. (2018). Kernelized subspace pooling for deep local descriptors. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1867–1875.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine* *Learning Research*, *10*(Feb), 207–244.

Weiss, Y., Torralba, A., & Fergus, R. (2009) Spectral hashing. In *Advances in neural information processing systems*, pp. 1753–1760.

Windheuser, T., Vestner, M., Rodolà, E., Triebel, R., & Cremers, D.

(2014). Optimal intrinsic descriptors for non-rigid shape analysis.

In *Proceedings of the British machine vision conference*.

Wohlhart, P., & Lepetit, V. (2015). Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 3109–3118.

Wu, Y., Lim, J., & Yang, M. H. (2015b). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1834–1848.

Wu, J., Zhang, H., & Guan, Y. (2014). Visual loop closure detection by matching binary visual features using locality sensitive hashing. In *Proceeding of the world congress on intelligent control and* *automation*, pp. 940–945.

Wu, C. Visualsfm: A visual structure from motion system. Retrieved November 16, 2018 from <http://ccwu.me/vsfm/doc.html>.

Wu, G., Kim, M., Wang, Q., Munsell, B. C., & Shen, D. (2015a). Scalable high-performance image registration framework by unsu-pervised deep feature representations learning. *IEEE Transactions* *on Biomedical Engineering*, *63*(7), 1505–1516.

Xiao, J., Owens, A., & Torralba, A. (2013). Sun3d: A database of big spaces reconstructed using SFM and object labels. In *Proceed-ings of the IEEE international conference on computer vision*, pp.1625–1632.

Xie, J., Wang, M., & Fang, Y. (2016). Learned binary spectral shape descriptor for 3d shape correspondence. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.3309–3317.

Yan, J., Li, Y., Liu, W., Zha, H., Yang, X., & Chu, S. M. (2014). Graduated consistency-regularized optimization for multi-graph matching. In *Proceedings of the European conference on com-puter vision*, pp. 407–422.

Yan, J., Ren, Z., Zha, H., & Chu, S. (2016a). A constrained clustering based approach for matching a collection of feature sets. In *Pro-ceedings of the international conference on pattern recognition*,pp. 3832–3837.

Yan, J., Tian, Y., Zha, H., Yang, X., Zhang, Y., & Chu, S. M. (2013). Joint optimization for consistent multiple graph matching. In *Pro-ceedings of the IEEE international conference on computer vision*,pp. 1649–1656.

Yan, J., Xu, H., Zha, H., Yang, X., Liu, H., & Chu, S. (2015c). A matrix decomposition perspective to multiple graph matching. In *Proceedings of the IEEE international conference on computer vision*, pp. 199–207.

Yan, J., Yin, X. C., Lin, W., Deng, C., Zha, H., & Yang, X. (2016b). A short survey of recent advances in graph matching. In *Proceedings* *of the ACM on international conference on multimedia retrieval*,pp. 167–174.

Yan, J., Zhang, C., Zha, H., Liu, W., Yang, X., & Chu, S. M. (2015d). Discrete hyper-graph matching. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.1520–1528.

Yan, J., Cho, M., Zha, H., Yang, X., & Chu, S. M. (2015a). Multi-graph matching via affinity optimization with graduated consistency reg-ularization. *IEEE Transactions on Pattern Analysis and Machine* *Intelligence*, *38*(6), 1228–1242.

Yang, M., Wu, F., & Li, W. (2020). Waveletstereo: Learning wavelet coefficients of disparity map in stereo matching. In *Proceedings of* *the IEEE/CVF conference on computer vision and pattern recog-nition*, pp. 12,885–12,894.

123

International Journal of Computer Vision

Yang, X., Kwitt, R., Styner, M., & Niethammer, M. (2017b). Quicksil-ver: Fast predictive image registration-a deep learning approach. *NeuroImage*, *158*, 378–396.

Yang, J., Li, H., Campbell, D., & Jia, Y. (2016). Go-ICP: A globally opti-mal solution to 3d ICP point-set registration. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *38*(11), 2241–2254.Yang, K., Pan, A., Yang, Y., Zhang, S., Ong, S., & Tang, H. (2017a).

Remote sensing image registration using multiple image features.

*Remote Sensing*, *9*(6), 581.

Yan, J., Wang, J., Zha, H., Yang, X., & Chu, S. (2015b). Consistency-driven alternating optimization for multigraph matching: A unified approach. *IEEE Transactions on Image Processing*, *24*(3), 994– 1009.

Yao, Y., Deng, B., Xu, W., & Zhang, J. (2020). Quasi-Newton solver for robust non-rigid registration. In *Proceedings of the IEEE/CVF* *conference on computer vision and pattern recognition*, pp. 7600–7609.

Ye, Y., Shan, J., Bruzzone, L., & Shen, L. (2017). Robust registration of multimodal remote sensing images based on structural similar-ity. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(5), 2941–2958.

Yew, Z. J., & Lee, G. H. (2020). RPM-NET: Robust point matching using learned features. In *Proceedings of the IEEE/CVF confer-ence on computer vision and pattern recognition*, pp. 11,824–11,833.

Yi, K. M., Trulls, E., Lepetit, V., & Fua, P. (2016). Lift: Learned invariant feature transform. In *Proceedings of the European conference on* *computer vision*, pp. 467–483.

Yin, Z., & Shi, J. (2018). Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE confer-ence on computer vision and pattern recognition*, pp. 1983–1992.

Yu, T., Wang, R., Yan, J., & Li, B. (2020a). Learning deep graph match-ing with channel-independent embedding and Hungarian attention. In *International conference on learning representations*.

Yu, T., Yan, J., & Li, B. (2020b). Determinant regularization for gradient-efficient graph matching. In *Proceedings of the* *IEEE/CVF conference on computer vision and pattern recogni-tion*, pp. 7123–7132.

Yu, T., Yan, J., Wang, Y., Liu, W., et al. (2018). Generalizing graph matching beyond quadratic assignment model. In *Advances in neu-ral information processing systems*, pp. 861–871.

Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Proceedings of the* *IEEE conference on computer vision and pattern recognition*, pp.4353–4361.

Zaharescu, A., Boyer, E., Varanasi, K., & Horaud, R. (2009). Surface feature detection and description with applications to mesh match-ing. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 373–380.

Zanfir, A., & Sminchisescu, C. (2018). Deep learning of graph matching. In *Proceedings of the IEEE conference on computer vision and* *pattern recognition*, pp. 2684–2693.

Zaslavskiy, M., Bach, F., & Vert, J. P. (2009). A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern* *Analysis and Machine Intelligence*, *31*(12), 2227–2242.

Zass, R., & Shashua, A. (2008). Probabilistic graph and hypergraph matching. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, IEEE, pp. 1–8.

Zbontar, J., & LeCun, Y. (2015). Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pp. 1592–1599.

Zbontar, J., & LeCun, Y. (2016). Stereo matching by training a convo-lution neural network to compare image patches. *The Journal of* *Machine Learning Research*, *17*(1), 2287–2318.

Zeng, Z., Chan, T. H., Jia, K., & Xu, D. (2012). Finding correspondence from multiple images via sparse and low-rank decomposition. In *Proceedings of the European conference on computer vision*, pp.325–339.

Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., & Funkhouser, T. (2017). 3dmatch: Learning local geometric descriptors from RGB-D reconstructions. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*, pp. 1802–1811.

Zeng, Y., Wang, C., Wang, Y., Gu, X., Samaras, D., & Paragios, N. (2010). Dense non-rigid surface registration using high-order graph matching. In *Proceedings of the IEEE conference on com-puter vision and pattern recognition*, pp. 382–389.

Zhang, H. (2011). Borf: Loop-closure detection with scale invariant visual features. In *Proceedings of the IEEE international confer-ence on robotics and automation*, pp. 3125–3130.

Zhang, L., & Rusinkiewicz, S. (2018). Learning to detect features in tex-ture images. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 6325–6333.

Zhang, F., Prisacariu, V., Yang, R., & Torr, P. H. (2019a). Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition*,pp. 185–194.

Zhang, Z., Shi, Q., McAuley, J., Wei, W., Zhang, Y., & Van Den Hengel, A. (2016). Pairwise matching through max-weight bipartite belief propagation. In *Proceedings of the IEEE conference on computer* *vision and pattern recognition*, pp. 1202–1210.

Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., & Liao, H. (2019b). Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE* *international conference on computer vision*, pp. 5845–5854.

Zhang, S., Yang, Y., Yang, K., Luo, Y., & Ong, S. H. (2017a). Point set registration with global-local correspondence and transformation estimation. In *Proceedings of the IEEE international conference* *on computer vision*, pp. 2669–2677.

Zhang, X., Yu, F. X., Karaman, S., & Chang, S. F. (2017b). Learning discriminative and transformation covariant local feature detec-tors. In *Proceedings of the IEEE conference on computer vision* *and pattern recognition*, pp. 6818–6826.

Zhang, X., Yu, F. X., Kumar, S., & Chang, S. F. (2017c). Learning spread-out local feature descriptors. In *Proceedings of the IEEE* *international conference on computer vision*, pp. 4595–4603.

Zhang, X., Qu, Y., Yang, D., Wang, H., & Kymer, J. (2015). Laplacian scale-space behavior of planar curve corners. *IEEE Transactions* *on Pattern Analysis and Machine Intelligence*, *37*(11), 2207–2217.

Zhang, X., Wang, H., Smith, A. W., Ling, X., Lovell, B. C., & Yang, D. (2010). Corner detection based on gradient correlation matrices of planar curves. *Pattern Recognition*, *43*(4), 1207–1223.

Zhao, J., & Ma, J. (2017). Visual homing by robust interpolation for sparse motion flow. In *Proceedings of the IEEE/RSJ international* *conference on intelligent robots and systems*, pp. 1282–1288.

Zhao, C., Cao, Z., Li, C., Li, X., & Yang, J. (2019). Nm-net: Mining reli-able neighbors for robust feature correspondences. In *Proceedings* *of the IEEE conference on computer vision and pattern recogni-tion*, pp. 215–224.

Zhao, Q., Karisch, S. E., Rendl, F., & Wolkowicz, H. (1998). Semidef-inite programming relaxations for the quadratic assignment prob-lem. *Journal of Combinatorial Optimization*, *2*(1), 71–109.

Zheng, L., Yang, Y., & Tian, Q. (2018). Sift meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis* *and Machine Intelligence*, *40*(5), 1224–1244.

Zhong, Y. (2009). Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Proceedings of the IEEE international* *conference on computer vision workshops*, pp. 689–696.

Zhou, W., Li, H., & Tian, Q. (2017). Recent advance in content-based image retrieval: A literature survey. arXiv preprint [arXiv:1706.06064](http://arxiv.org/abs/1706.06064).

123

International Journal of Computer Vision

Zhou, W., Li, H., Lu, Y., & Tian, Q. (2011). Large scale image search with geometric coding. In *Proceedings of the ACM international* *conference on multimedia*, pp. 1349–1352.

Zhou, W., Lu, Y., Li, H., Song, Y., & Tian, Q. (2010). Spatial coding for large scale partial-duplicate web image search. In *Proceedings of* *the ACM international conference on multimedia*, pp. 511–520.

Zhou, X., Zhu, M., & Daniilidis, K. (2015). Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE interna-tional conference on computer vision*, pp. 4032–4040.

Zhou, L., Zhu, S., Luo, Z., Shen, T., Zhang, R., Zhen, M., Fang, T., & Quan, L. (2018). Learning and matching multi-view descriptors for registration of point clouds. In *Proceedings of the European* *conference on computer vision*, pp. 505–522.

Zhou, F., & De la Torre, F. (2015). Factorized graph matching. *IEEE* *Transactions on Pattern Analysis and Machine Intelligence*, *38*(9),1774–1789.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer* *vision*, pp. 2223–2232.

Zieba, M., Semberecki, P., El-Gaaly, T., & Trzcinski, T. (2018). Bingan: Learning compact binary descriptors with a regularized GAN. In *Advances in neural information processing systems*, pp. 3608–3618.

Zitnick, C. L., & Ramnath, K. (2011). Edge foci interest points. In *Proceedings of the IEEE international conference on computer vision*, pp. 359–366.

Zitova, B., & Flusser, J. (2003). Image registration methods: A survey.

*Image and Vision Computing*, *21*(11), 977–1000.

**Publisher’s Note** Springer Nature remains neutral with regard to juris-dictional claims in published maps and institutional affiliations.

123