Research of image detection and matching algorithms

Yufei Bai

Dublin International College, Beijing University of Industry, Beijing, China

yufei.bai@ucdconnect.ie

Abstract. Image matching, a fundamental computer vision method, serves as a crucial pillar for more complex vision applications. The general adoption of feature-based image registration technologies has been accelerated by advances in computing hardware and vision theory. As the current research in this field is not very sufficient, this paper gives an overview of the relevant aspects. At the beginning, this article first introduces the research background, the research achievements and the application in different fields of image feature detection and matching. The main body discusses the most current advancements in this subject, including feature points, local features, global features, matching, and optimization, after examining the classical detection algorithms from recent decades and referencing the most recent machine learning algorithm headed by depth learning, and shows the advantages and disadvantages of the algorithms. Finally, the paper summarizes and prospects the full text.

Keywords: feature detection algorithms, local descriptor, deep learning, machine learning

1. Introduction

Image feature detection and matching has been a key topic of image processing and the foundation of computer vision since the 1970s. A two-dimensional picture of a plane is received by the human eye or camera. The unwavering aim in this discipline has been three-dimensional reconstruction, comprehension, and mastery of the world. Researchers have made significant progress using mathematical techniques to analyze this process for decades, from the comprehension of texture and color in the twentieth century to the feature extraction of lines, points, and faces in the twenty-first century. Global automation has steadily developed as a result of the world's continually rising scientific and technical level in recent years. Artificial intelligence technology has emerged as a popular study topic in many nations because it has the potential to make a machine-federated computer see, comprehend, and act like a human person. One of the most crucial perception technologies is visual perception. In the field of artificial intelligence research, it is essential. As visual perception technology is promoted, computer vision technology is developing quickly. At the same time, one of the hottest study areas in the entire subject of computer vision is how to identify distinctions and connections between various visual targets and process the observed information in accordance with certain requirements. Feature matching joins two picture targets with the same or comparable properties as a fundamental and crucial technique. It serves as the connection between low-level and high-level vision. It is a successful method for recognizing, incorporating, and recovering high-dimensional structures from low-dimensional images [1]. Deep learning-driven artificial intelligence algorithms have also achieved significant advancements in the field of computer images and had a significant impact on the fields of image feature matching and detection. According to recent research, deep learning is becoming more prevalent in all aspects of picture feature matching and detection, from manual feature selection to data-driven learning. Additionally, due to the powerful ability of deep learning technology to learn and communicate in-depth features, deep convolution network-based feature matching technology has also drawn significant attention and offers a fresh approach to the problem of solving picture matching issues.

In many domains, feature matching is a fundamental and crucial technology. Feature matching issues can be classified into various times, different viewpoints, and different sensors, or template image matching, depending on the variations of data collecting or imaging settings. Each type of image acquisition has corresponding application purposes: 1) Feature matching based on different imaging times. It is generally used for scene change detection, security monitoring and target tracking, and disease tracking in medical diagnosis and treatment. 2) Feature matching based on different perspectives. Its main purpose is to match sequence images of the same target or scene taken from different perspectives, such as restoring camera posture and establishing camera trajectory, three-dimensional reconstruction of target or scene, and stitching of remote sensing panoramic images. 3) Feature matching based on different sensors. This type of matching is often used for multimode image matching in medical image analysis, security, and military fields such as infrared visible registration, image registration, and fusion with different resolutions and spectral information in remote sensing image processing. 4) Template-based feature matching. This type is often used for template recognition, difference detection, or content retrievals, such as visual-based pattern recognition (character recognition, license plate recognition), image retrieval, disease diagnosis in medical image analysis, sample classification, and matching and positioning of aerial or satellite images in other known geographic information maps from remote sensing images. Therefore, in-depth study on it has important practical application value.

This paper first introduces the basic process of image feature detection and matching, then respectively introduces the content of traditional and learning-based image feature detection and matching, and their differences. Finally, a summary will be given to illustrate the achievements and shortcomings of image feature detection and matching.

2. The Main Steps of the Feature Detection and Matching of Images.

The feature detection and matching of images are always divided into five main steps: Image Preprocessing, Feature Point Detection, Local descriptor extraction, Global descriptor extraction and Feature Matching.

●The first step is to have the pretreatment of the images. According to different algorithms used in this progress, the images need different ways to be handled, such as graying, noise elimination.

●After the images processed, the feature detection algorithms begin to work. The algorithms could choose some feature points to represent the whole big image. The selections of the feature points follow different standards and could make the image sparse.

●When the feature points are chosen, local descriptor could be extracted. Some small geometric domains could be found and selected around the feature points. To express the features of these domains, each domains has a unique vector which is called local descriptor.

●These vectors are used to identify different domains and become the basic information of the following steps. Of course, when the image is processed, it’s possible to find some features. These features represent the high-level characteristics of the images and could be used to search the images. Also, these descriptors could be found in the local descriptors.

●After the descriptors are found, the features of the images could be used to match the other image and find the match points. The feature detection and matching of images could be done.

3. Traditional Image Feature Detection and Matching

In early years, instead of machine learning and high technology methods, feature detection and matching algorithms are mainly based on some fundamental theories of mathematics. These algorithms combine linear algebra and mathematical vectors. However, most of them have some shortages and can only fit some of the positions, which asks users to choose the best one for their program.

3.1. Image Feature Detection Methods

The most famous way of feature detection is Corner Detection. It's mainly known as the intersection of two lines. In a narrow sense, the local neighborhood of the corner should include two boundaries of two different areas in two different directions. Well, the corner in the real world could be shown as crossings, junctions and other similar things. As the corner is a stable corner which could be maintained as before after changing the visual angle and could always exist, it is a very excellent feature point to identify the images. However, most of the corner detection algorithms are based on some unique points of the images instead of corner points. In most of the algorithms, the special feature points have coordinates and some other mathematical characteristics, such as gradients, greyscale.

The other useful way is blob detection. A blob is an area, which varies a lot with the surrounding pixels in gray scales or colors. It's a more universal meaning feature point. Compared to the corner points, blobs are more stable and could avoid the influence of noise better.

A fundamental way to find the blobs is to use Laplacian, which is a simple isotropic differential operator. It has rotation invariance, but it is quite sensitive to the noise. In that case, in 1980, Marr and Hildreth combine the Laplacian and Gaussian low-pass filtering and put forward a new way to do that job -- Laplace and Guassian [2]. Laplace and Guassian gives researchers a quite useful fundamental algorithm to do further job, but it still has some disadvantages such as the huge calculation amount and the low processing speed.

A famous algorithm is called SIFT (Scale-invariant feature transform) algorithm (Figure 1). It uses Gauss Pyramid and DOG (Difference of Gaussian) Pyramid to have the processed images and then compared the points in different levels and find the extremum points [3]. Then give the directions and positions. After that, there could be 128 values to build a vector. It has rotational invariance but has less feature points.

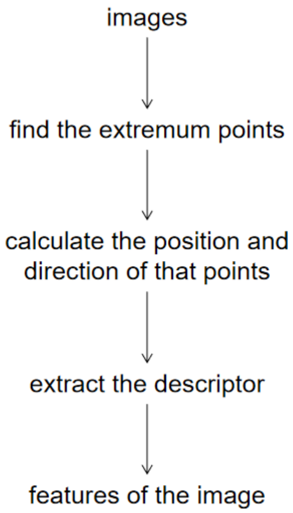

Fig.1. The process of SIFT.

Based on SIFT algorithm, SURF (Speeded-Up Robust Features) algorithm is similar as SIFT algorithm and in a faster speed. The SURF algorithm builds pyramid from determinant image of Hessian matrix, which could reduce the calculating time. The other algorithm based on SIFT is ORB (Oriented Fast and Rotated Brie) algorithm. It produced only one picture on every level of the pyramid. It also uses FAST (Features fromaccelerated segment test) corner algorithm to detect the original point. At last, there could be 256 values to describe the vector. There is also a good algorithm called Harris corner algorithm. It uses matrix models to find out the result and calculate the two feature vectors. The Harris algorithm don't have the scale invariance, while it could have a reply to the changes of light and contrast ratio.

3.2. Algorithms Using Local Descriptor

The progress of matching is to compare the pieces of images instead of just points. Though the feature points are found, the comparation is still about the areas around these points. So, the problem is changed into how to represent these feature areas. An obvious solution is to use feature vectors to represent the areas. These feature vectors are called local descriptors.

The local descriptors need invariance and distinguishability. Because the local descriptors need to stay stable when dealing with problems such as image transform, the first thing when building a local descriptor is to make sure the distinguishability of it. In wide baseline matching, it's important to consider the invariance of feature descriptors to view angle changes, scale changes, rotation changes [4]. In shape recognition and object retrieval, considering the invariance of feature descriptors to shapes is the most significant.

Many kinds of local descriptors are designed and used now. Among the traditional descriptors, SIFT descriptor can't be ignored (figure 2). It uses distribution characteristics of gradient direction of the pixels besides feature point. After that, every feature points could have direction parameters which ensure the rotation invariance and scale invariance. With calculating the surrounding areas, a 128 values vector could build and normalize the features to decrease the influence of the light.

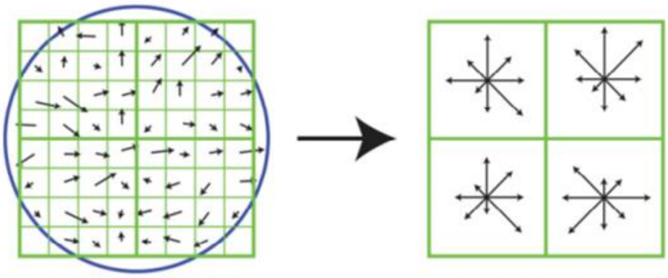

Fig.2. SIFT descriptor.

With all these important functions, SIFT descriptor has an obvious shortage: huge amount of calculation and low efficiency. Though it has 128 values of the vector, many of the values are useless. In that case, many functions are focused on how to reduce the dimensions and improve the efficiency. BRIEF (Binary Robust Independent Elementary Features) algorithm simplifies the whole process of building the local descriptors [5]. It doesn't need to calculate the complex feature descriptors like SIFT, while it just produces a two-value string to represent the features. It chooses several point pairs in the surrounding area of the feature point and then compared the brightness value of every point pair. The bigger one gets the value 1 and the lower one gets the value 0. After the comparation, a long binary string has been produced. This string could usually be in length of 128, 256 or 512. BRIEF algorithm could work in a high speed and is very simple to use. However, it can't be used in large scale rotation, which may need to work together with other algorithms.

3.3. Feature Points Matching Process

After the feature descriptors are produced, the feature points of two images could begin matching. The fundamental way is to use Brute-force matcher, which calculate the distance of every line of the feature descriptors and compare the distance to get the results. In different situations, the distances are calculated in different ways, such as Hanming distance in ORB algorithm. This function is quite simple to realize but the time and space complexity is so high. In that case, FLAAN matching (Flann-based matcher) is used to simplify the work. It uses fast approximate nearest neighbor search algorithm to match. This algorithm may not exactly find the best result, so it always needs indexes.

During the Brute-force matcher, there are always some wrong results, which could mainly be classified in two kinds: False-positive matches and False negative matches. False-positive matches means that two wrong feature points are matched, while False-negative matches means that two matching points can't be discovered by the algorithms. To solve these problems, some optimization algorithms are used. RANSAC (Random Sample Consensus) is one of the most widely used uniformity optimization algorithm [6]. It uses hypotheticality and randomness to solve the problem. When handling a data set, it uses iteration to build mathematical models and estimate the parameters. Because it's quite uncertain, it could reduce the amount of calculation but may not get the right result. In that case, it needs more times of iteration. RANSAC algorithm could be widely used in many kinds of uniformity optimization.

4. Learning-based Image Feature Detection and Matching

The selection of empirical parameters, lack of access to context information, and manual design characteristics all have a significant impact on the matching effect and are not appropriate for use in complex situations. Machine learning algorithms driven by in-depth learning have become the foundation of research and application over the past ten years as a result of improving computing performance and the popularity of large-scale data labeling datasets. Traditional hand labeling descriptors have gradually given way to data-driven learning algorithms.

4.1. Learning-based Key Point Detection

A time-invariant learning detector was proposed by Verdie et al. It employs a set of training images taken from the same scene from the same angle during various seasons and times, creates training datasets through DoG (Difference of Gaussian), trains using a unique piecewise linear regression function, and optimizes using PCA (Principal Component Analysis) to achieve better repeatable feature point detection than SIFT under varying lighting conditions.

The Quad-Networks technique trains the neural network to rank points in an invariant transformation using unsupervised data representation. It takes the points of interest from the top/bottom of the ranking and changes the problem of the learning interest point detector into the problem of the learning ranking point.

Key.Net is a technique that compounds manually added feature points onto an image of the original size after filtering them through a CNN (Convolutional Neural Network) network [7]. Figure 3 depicts the network structure.

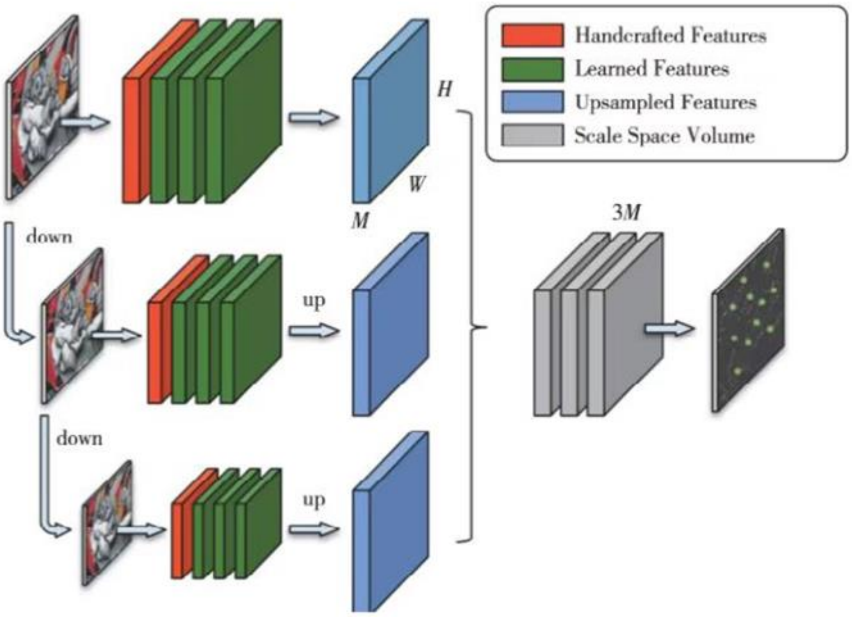

Figure 3. Network structure diagram for Key Net [7].

4.2. Learning-based Local Descriptor Detection

In order to extract the features of local images, local descriptors play this role. A whole image can typically be divided into equal blocks, and each block is referred to as a patch. All manual feature selection algorithms, in the eyes of deep learning researchers, fall far short of feature extraction networks like CNN. Consequently, it makes sense to use CNN in place of manual feature descriptors.

A CNN network with an easy structure and effective feature extraction was proposed by L2-NET [8]. To make sure that the network can access billions of training samples in only a few epochs, it suggested a progressive sampling technique. It also eliminated the distance threshold setting by concentrating on the relative distance between patch characteristics. The network structure of L2-NET is shown in Figure 4.

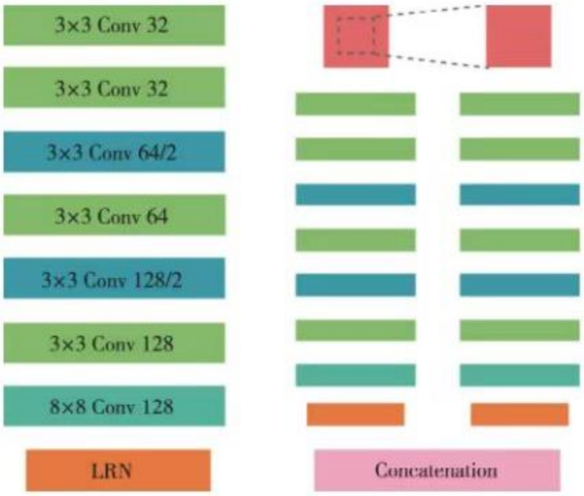

Fig. 4. CNN layers for L2-Net.

Using only one loss function, learning more potent descriptors, conducting numerous in-depth tests on picture matching, retrieval, wide baseline, etc., and attaining the most cutting-edge outcomes in actual tasks, HardNet has further enhanced the foundation of L2-NET. There is an algorithm for image feature extraction that can go into the practical realm.

There is a significant issue with the selection of patch blocks regardless of the dataset used to train the network. There are various alternatives for descriptors. The key point detector typically needs to accurately predict its size, shape, and orientation in advance. To extract the "support region," the reference advises using a log-polar sampling approach. This model can use a bigger support region without being hidden and can match descriptors at a larger scale than before.

4.3. Learning-based Global Descriptor Detection

In a way, target detection and image categorization are also global descriptors. Choosing to extract image-based global descriptors becomes crucial for general image retrieval, particularly for large-scale image retrieval.

A straightforward and understandable global description algorithm is the BoW (Bag-of-Words) algorithm. Assuming that the word order, grammar, syntax, and other components of a document are ignored and that it is simply taken into account as a collection of words, the BoVW (Bag of Visual Words) algorithm is brought into the image field [9].

The N D-dimensional features in an image are first extracted by the VLAD (Vector of Local Aggregated Descriptors) technique, and then all N D-dimensional feature images are extracted [10]. K global features are produced after the D-feature map is clustered using K-means to produce K cluster centers and after collecting all of the feature residuals from each cluster. By erasing the differences in the distribution of the features of the image itself and only keeping the differences between the local features and the distribution of the cluster centers, these K global features represent a distribution of local features within the cluster, producing a global descriptor of a specific size. VLAD encoding is another name for the final encoding. In order to extract global descriptors based on VLAD, NetVLAD uses CNN. The main issue with NetVLAD is that the output feature is too huge in size, making it challenging to process or fit. NeXtVLAD builds upon NetVLAD by going further. It incorporates the notion of ResNet's metamorphosis by ResNeXt. In order to achieve greater fitting and dimension reduction, it divides the FC network into three sections before using NetVLAD aggregation and breaks down high-dimensional features into a collection of relatively low-dimensional vectors.

Additionally, the notion of directly extracting global features from images using CNNs is more widely used. The classification convolution neural network, first proposed by the Neural Code descriptor [11], is trained using a sizable categorized dataset, such as Image-Net. A high-level descriptor of the visual content and semantic level of a picture can be used directly from the output value of the full-connected layer near the top.

5. Conclusion

In recent years, the detection and matching algorithms of images are using in most of industries and are still developing in a rapid speed. With the growing needs of image matching, artificial matching is disappeared and people are depending on machines now. Traditional algorithms are mainly based on mathematical calculations and models, which are simplification and abstraction of the real-world images. In that case, traditional algorithms are lack of generalization and robustness. To improve the algorithms, people focus more on improving the calculation ability of the machine and big data for the deep learning. Machine learning could mainly use an intact end-to-end internet to get the results. With a higher efficiency and faster speed, algorithms based on machine learning are becoming the main steam.

However, because machine learning is based on the collected data, how to collect enough data is becoming the biggest problem. It’s possible to get a universally data set for the whole world, so in some situation the results may not be so reliable. Also, deep learning needs higher calculation ability than traditional ones, which may need much more GPU (Graphics Processing Unit) or CPU (Central Processing Unit / Processor) to work for the process, having a higher power waste and needing better equipment.

Till now, though there are some shortages of the algorithms, people are still working on them and trying to find a better universality method to work on it. Combining traditional algorithms and machine learning algorithms, a new framework of image detection and matching system is building and could be the basic part for the future image processing revolution.

References

[1] Zitova, B., Flusser, J. 2003. Image registration methods: a survey. *Imag. Vis. Comput.*, **21(11)**, 977–1000.

[2] Marr, D., & Hildreth, E. 1980. Theory of edge detection. *Royal Soc. London Biol. Sci.*, **207(1167)**, 187-217.

[3] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Inter. J. Com. Vis.*, **60(2)**, 91-110.

[4] Tola, E., Lepetit, V., & Fua, P. 2008. A fast local descriptor for dense matching. *Conf. Com. Vis. Pat. Rec.* 1-8.

[5] Calonder, M., Lepetit, V., Strecha, C., & Brief, F. P. 2019 Binary robust independent elementary features. *Euro. Conf. Comput. Vis.* 778-792.

[6] Ke, Y., & Sukthankar, R. 2004. PCA-SIFT: A more distinctive representation for local image descriptors. *Conf. Com. Vis. Pat. Rec.* **2**, 506-513.

[7] Barroso-Laguna, A., Riba, E.，Ponsa, D.，et al. 2019 Key．Net: keypoint detection by handcrafted and learned CNN filters． *arXiv preprint，arXiv: 1904. 00889*

[8] Tian, Y R., Fan, B., Wu, F C. 2017 L2-Net: deep learning of discriminative patch descriptor in euclidean space. *Conf. Com. Vis. Pat. Rec.* 661-669.

[9] Sivic, J., Zisserman, A. 2003 Video Google: a text retrieval approach to object matching in video.

Inter. Conf. Comput. Vis. 1470.

[10] Jégou, H., Douze, M., Schmid, C., et al．2010 Aggregating local descriptors into a compact image representation. *Inter. Conf. Comput. Vis.* 3304-3311.

[11] Babenko, A., Slesarev, A., Chigorin, A., et al．2014. Neural codes for image retrieval. *Euro. Conf. Comput. Vis.* 584-599.