Research on feature point matching algorithm improvement using depth prediction

Yongbin Chen1, Guitang Wang1, Liming Wu1
1Guangdong University of Technology, Guangzhou, 510006, People’s Republic of China
E-mail: wanggt@gdut.edu.cn

Abstract: Feature point matching plays an important role in feature-based image registration such as the scale-invariant feature transform algorithm. Feature-based image registration is widely used in visual simultaneous localisation and mapping, augmented reality, self-driving etc. The most meaningful study on feature matching is to improve the accuracy and efficiency and this study pays attention to improving the accuracy by removing the mismatching feature points. Since most of the existed feature-based image registration algorithms are not so strong and efficient enough in mismatch removing, in this study, the authors propose a novel mismatch removal algorithm by incorporating depth prediction into feature matching to improve the performance. In this approach, the depth maps are predicted in pixel-wise through the given red–green–blue images using a deep learning algorithm. Experimental results show that their method outperforms conventional ones in mismatch removing.

1 Introduction

Image registration is mainly to align the adjacent frames which are taken in different times or different views, then determine the relationship between the matching images in the three-dimensional (3D) environment. Feature-based image matching is the foundation of many applications related to computer vision, such as augmented reality, visual odometry, visual simultaneous localisation and mapping (SLAM), object tracking, self-driving etc. The effect of feature matching directly impacts the performance of these applications. However, due to the diversity of the surrounding environment, feature matching is still the hot spots and difficulties of computer vision and pattern recognition.

Currently, image matching methods mainly have two types, which are feature-based and grey-based approaches. Feature-based approaches have some advantages, including low redundancy, singularity, and invariance to image transformation. There are some feature-based methods, such as scale-invariant feature transform (SIFT), oriented features from accelerated segment test (FAST), binary robust independent elementary features (ORB), speeded up robust features (SURF) etc. Due to the different demands, all of them have strengths and weaknesses.

Rublee et al. [1] presented ORB in 2011. ORB feature matching is a combination of amended features from accelerated segment test (FAST) [2] and binary robust independent elementary features (BRIFE) [3] description algorithm. The corners are detected in each layer of the pyramid of FAST and used Harris Corner to screen the quality feature points. ORB features are limited affine changes and invariant to rotation and scale.

Bay et al. [4] proposed the SURF algorithm in 2008, which depends on image Gauss scale space analysis. The SURF algorithm is based on the determinant of the Hessian matrix, and the integral image is used to improve the effectiveness of the feature detection. The 64 bins descriptor of SURF describes each detected feature and has the distribution of Haar wavelet responses in a certain neighbourhood. SURF is characterised by scale and rotation invariant rather than affine transformation. However, the SURF descriptor can be extended to 128 bin values with larger view changes. The main advantage of SURF on SIFT is its low computation cost.

SIFT algorithm is one of the most famous feature-based image registrations, which was first proposed by Lowe [5] in the International Conference on Computer Vision in 1999. It aims at detecting the invariant feature points, which also provides some message on the object recognition. In 2004, the SIFT feature detector, which is the most complete was granted by Lowe [6]. Due to the most attractive advantage of invariance to image transformation, SIFT has been widely researched in the recent few years. The SIFT descriptor is one of the best feature matching method. However, it does not have enough robustness and distinctiveness in feature match especially the real-time implementation. Some authors presented some methods to improve the performance of SIFT, such as an enhanced SIFT, edge-SIFT, layer parallel SIFT and so on. Both of them pay attention to improving the accuracy and efficiency of the SIFT algorithm.

Owing to the complicated environment, it is difficult to achieve the goal just using the pixel-wise red–green–green (RGB) information of adjacent images.

Single view depth prediction is an ancient subject in computer vision. In recent years, the related technologies have been well developed. One of the most attractive technologies is Visual SLAM, which uses camera motion to predict camera pose. Based on different time intervals and image frames, triangulation is used to estimate depth information. Without the ideal surrounding assumption, depth prediction from a single view of a general scene is not robust.

In recent years, convolutional neural networks have been used for learning the complicated relationship between depth pixels and colour pixels [7–11]. Due to a large number of parameters involved in deep networks, these methods encompass a higher complexity, these methods comprise higher intricacy. However, the deep learning algorithm improves the accuracy of the standard benchmark datasets.

In this study, we present a novel method to improve the accuracy of feature matching by combining depth estimation with SIFT; the feature matching is limited to the pixel region of a similar value of depth in the adjacent images, which can reduce the mismatching points. Experimental results show that our novel method increases the accuracy of feature matching.

This paper is organised as follows: Section 2 gives an overview of the algorithm. Section 3 introduces the architecture including the depth estimation algorithm and traditional SIFT algorithm. Section 4 shows the experiments and analysis and the final is the conclusions of this paper.

2 System design

In this study, we propose an algorithm, which combines SIFT feature matching and depth prediction to boost the performance of
image matching. With the help of depth prediction, we can segment the image by the pixel-wise depth information so that it is able to narrow the matching region between the adjacent frames and reduce the mismatching feature points. Fig. 1 is a brief presentation of this system.

As is shown in Fig. 1, the improved SIFT algorithm extracts the feature points with the traditional feature detector and removes the mismatching points by narrowing the matching pixel region which is segmented by the depth prediction algorithm MonoDepth. In other words, the matched points are bound to have a similar depth value between the adjacent images.

3 Architecture

Our method has two major parts, the first one is the feature matching and the other is the depth prediction. In this study, the SIFT algorithm is used for feature extracting and matching while the MonoDepth algorithm is used for depth prediction. The introduction of SIFT and MonoDepth is as follow.

3.1 Scale-invariant feature transform (SIFT)

SIFT is an algorithm for detecting and describing the features in the image, which searches extreme points in the spatial scale and extracts their position. This algorithm was published by David Lowe in 1999 and summarised in 2004, and some improved-SIFT [12, 13] are presented recently.

SIFT features are based on locating the feature points on the object and are independent of image size and rotation. It has a good tolerance for light changes, noise, and small viewing angles. It is easy to identify objects with little misunderstanding though in a large number feature database. The use of SIFT feature descriptions is also quite often for the detection of partial object occlusion, and even more than three SIFT object features are sufficient to calculate the direction and position. With the computer hardware conditions present, the recognition speed can be approximated to real-time operations.

The steps of SIFT are shown as follows.

3.1.1 Extremum detection of scale space: The difference of gauss pyramid between two adjacent scales, which generated the DoG space pyramid and established the DoG space pyramid

\[ L(x, y, \sigma) = G(x, y, \sigma) \otimes I(x, y), \]  

(1)

where \( I(x, y) \) indicates the original image. \( \sigma \) indicates the scale factor, the smaller the value of \( \sigma \) is, the less smooth the image is and the corresponding proportion is smaller. \( \otimes \) is the convolution operation and \( (x, y) \) represents the row and column in the image, meanwhile, \( G(x, y, \sigma) \) is a typical Gaussian function.

A DoG scale-space is presented for detecting stable key points effectively, as is shown in (2) that different scales of Gaussian differential kernels are used

\[ \text{DoG}(x, y, k\sigma) = L(x, y, k\sigma) - L(x, y, \sigma). \]  

(2)

3.1.2 Key points localisation: Each sampling point needs to be compared with all its adjacent points to find the extreme points of the scale space to determine whether it is more or less than the adjacent points in its image and scale domains.

The point is considered as a feature point of the image if it is the minimum or maximum in the 26 domains of DoG.

3.1.3 Low contrast points: The position and scale of the key points, which can reach sub-pixel accuracy, are accurately determined by simulating the 3D function, and the low contrast and the unstable edge response points are removed. To enhance the matching stability and the anti-noise ability, it uses an approximate Harris Corner Detector.

Each candidate point applies the scale space Taylor expansion formula; the low contrast points will be removed from the candidate points

\[ D(X) = D(X_0) + \frac{\partial D}{\partial X}(X - X_0) + \frac{1}{2} (X - X_0)^T \frac{\partial^2 D}{\partial X^2}(X - X_0), \]  

(3)

where \((X - X_0)^T\) is the offset of this point, by calculating the derivative of the function \(D\) concerning point \(X\), we find the exact location of the extreme critical point \(X\). Also, the derivative of function \(D\) is set to zero. It is as shown in the following equation:

\[ \hat{X} = -\left(\frac{\partial D}{\partial X}\right)^{-1} \frac{\partial D}{\partial X}. \]  

(4)

Substituting (4) into (3), we have

\[ D(X) = D(X_0) + \frac{1}{2} \frac{\partial^2 D}{\partial X^2} \hat{X}. \]  

(5)

The candidate points with (5) will be abandoned in which absolute value is beyond the threshold.

3.1.4 Unstable points with strong responses along edges: The extremes of a poorly defined Gaussian difference operator have larger principal curvatures across the edge and less primary curvature in the vertical direction. It obtained the main curvature by a \(2 \times 2\) Hessian matrix \(H\)

\[ H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \]  

(6)

It will be considered as a candidate point once its value \((\text{Trace}(H^2))/(\text{Det}(H))\) is less than a certain threshold.

3.1.5 Principal orientation generation: Now, the key point of each image is decided, it will calculate the direction of each feature point, and further calculations are performed according to this direction

\[ m(x, y) = \sqrt{(L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2}, \]  

(7)

\[ \theta(x, y) = \arctan\left(\frac{L(x, y + 1) - L(x, y - 1))}{L(x + 1, y) - L(x - 1, y))}\right), \]  

(8)

where \(L(x, y)\) is a point of the image, and as is shown in (7) and (8), it is the gradient and direction.
In the feature-based image registration, the process of matching is according to the similarity or difference between the location of the feature pixels and the brightness. However, the method based on the pixel RGB value will produce a large number of mismatching points, which cannot achieve better results. The SIFT algorithm is used to extract and match the feature points, and then a part of the mismatching points can be removed by random sample consensus (RANSAC), which can improve the matching accuracy. However, it is difficult to determine the cost function and threshold in RANSAC. If the threshold $T$ is too small, the estimated parameter model will be unstable, and if the threshold $T$ is set too large, more external points will be judged to be internal points, resulting in the error of the model estimation. Combining the depth value of a 2D image to assist image registration is one of the solutions.

In this section, we are going to introduce a state-of-the-art unsupervised depth prediction method MonoDepth, which is proposed by Godard et al. [14] in the Conference on Computer Vision and Pattern Recognition (CVPR) 2017. One of the most important highlights is that MonoDepth uses binocular stereo data, which is easy to capture, rather than using hard to obtain labelled depth data. The approach of depth estimation exploits a novel loss function to improve prediction, which achieves the consistency between the predicted depth maps and camera views during training. Results show that the MonoDepth is superior to the fully supervised depth prediction which means that it is not necessary to require the expensive ground truth data for training in the future. Fig. 2 shows the system framework of MonoDepth and Fig. 3 shows the performance of depth estimation

$$C_x = \alpha_p (C_{sp} + C_{dp}) + \alpha_d (C_{sp} + C_{dp}) + \alpha_h (C_{sp} + C_{dp})$$

where $C_{sp}$ makes the reconstructed image similar to the training input image, $C_{dp}$ smoothed the disparities, and $C_h$ let the prediction of the difference between left and right to be consistent. Only the left image can be fed through the convolutional layers though each major term containing both left and right image variants. Then, it proposed that each component of the loss is in the left image (e.g. $C_{sp}$). The right side e.g. $C_{sp}$ needs to be swaged to the left and sampled in the opposite direction.

To improve the SIFT algorithm, we incorporate MonoDepth prediction into feature matching to narrow the feature matching pixel region. The core of MonoDepth is open access and we modify the source code to adapt to our improved SIFT algorithm. It mainly follows (10)

$$\begin{align*}
D(P_1, P_2) < \text{Threshold}_{\text{spatial}} \\
D(P_1, P_2) < \text{Threshold}_{\text{depth}}(P_1 \in I_1, P_2 \in I_2).
\end{align*}$$

where vector $P_1$ represents the matched points in IMAGE $I_1$, vector $P_2$ represents the matched points in image $I_2$. When the spatial distance between matched points $P_1$ and $P_2$ is less than Threshold$_{\text{spatial}}$ and the depth value between $P_1$ and $P_2$ is less than Threshold$_{\text{depth}}$, the matched points are defined in the same pixel region, which means achieving the good matched result based on MonoDepth algorithm.

## 4 Experimental evaluation

In this portion, the performance of depth predicted-based SIFT is given, and the experimental results are presented under some pairs of images. Images for the experiment are picked from KITTI Dataset [15] which captures outdoor scenes with a high resolution of 376 × 1241. Both datasets provide ground truth camera poses for evaluation. It is widely used in Visual SLAM and semantic segmentation. The performance results are implemented on Intel(R) Core(TM) i7CPU and NVIDIA GeForce-1070 with Ubuntu16.04-LTE. The MonoDepth is based on graphics processing unit using CUDA-9.1 SDK.

There are ten couple images for testing our algorithm. Firstly, the depth value of images is predicted using MonoDepth algorithm, and the depth value and the pixel region are also obtained. Secondly, the SIFT algorithm is applied to extract the images features, and compute the descriptors. Finally, the feature points are matched under the related pixel region in the images of a similar depth value. Examination results show that the mismatching ration can be reduced by narrowing the matching region. The performance of depth estimation is shown in Fig. 4.

The first column and the third column are the adjacent origin images, and the second and fourth columns are the results of depth estimation with MonoDepth. Different colour represents the different distance from the camera; it shows that the pixel-wise estimation is excellent. Besides, there is some depth value which is not by reality, so the performance of depth predicted-based feature matching will be optimised with the progress of the depth estimation algorithm.

We utilise OpenCV-3.20 SDK for SIFT feature extraction and descriptors. When matching without MonoDepth, the matching points, whose matching distance is greater than three times minimal distance, are removed to reduce mismatching preliminarily. When matching with MonoDepth, SIFT algorithm of OpenCV is used to extract feature points and compute descriptors which is similar to matching without MonoDepth. However, when
it matches with adjacent images, it should compare the depth distance between the matched feature points, and we set the Threshold value as 35, which means that the distance between the good matched points to the camera must be <35.

Table 1 illustrates their quantitative evaluation. Under the same conditions, the RANSAC algorithm is as a criterion for judgment which computes the inliers points and counts each pair images’ inliers ratio.

Table 1: Quantitative comparison of feature matching with/without MonoDepth

| Image pairs | All detected points | Matched Points | Points | Inliers Points | Points | Inliers Points | Ratios, % |
|-------------|---------------------|----------------|--------|----------------|--------|----------------|----------|
|             |                     | Without Mono   | Mono   | Without Mono   | Mono   | Without mono   | Mono     |
| 1           | 8056                | 3454           | 3330   | 2022           | 2122   | 58             | 64       |
| 2           | 7894                | 3664           | 2861   | 1621           | 1432   | 44             | 50       |
| 3           | 8310                | 1604           | 1099   | 1194           | 1005   | 74.4           | 91.4     |
| 4           | 8301                | 5098           | 4718   | 3643           | 3560   | 71.5           | 75.5     |
| 5           | 8083                | 1297           | 1058   | 776            | 647    | 59.8           | 61.2     |
| 6           | 8225                | 4529           | 3056   | 2704           | 1877   | 59.7           | 61.4     |
| 7           | 8567                | 3143           | 2650   | 1128           | 1020   | 35.9           | 38.5     |
| 8           | 5370                | 4021           | 3143   | 1628           | 1738   | 41             | 55.3     |
| 9           | 8189                | 4480           | 3165   | 2202           | 1885   | 49.2           | 59.6     |
| 10          | 7580                | 1578           | 908    | 1148           | 801    | 72.8           | 88.2     |

it matches with adjacent images, it should compare the depth distance between the matched feature points, and we set the Threshold value as 35, which means that the distance between the good matched points to the camera must be <35.

Table 1 illustrates their quantitative evaluation. Under the same conditions, the RANSAC algorithm is as a criterion for judgment which computes the inliers points and counts each pair images’ inliers ratio.

Since we apply the same SIFT algorithm in OpenCV-3.2 to extract feature points in both algorithms with and without MonoDepth, the numbers of ‘AllDetectPoints’ are the same. The ‘FeatureMatchedPoints’ of the method with MonoDepth is a bit less than the method without MonoDepth while the Inliers Ratio is in contrast. This means that the proposed method with depth predicted-based SIFT has better results. Experimental results show that the feature matching with depth estimation has better accuracy than the traditional one. Regrettably, it costs a little more time. However, in some case, applications pay more attention to accuracy rather than efficiency such as the 3D reconstruction of augmented reality. Our method provides more accurate location information.

5 Conclusion

In this study, we present a novel method to improve the accuracy of feature matching by combining depth estimation with SIFT, the feature matching is limited to the pixel region of the similar value of depth in the adjacent images which can reduce the mismatching points. Experimental results show that the novel method increases the accuracy of feature matching. There is no doubt that the ability of depth prediction still has room for improvement. With the rapid development of deep learning and the hardware computing power, depth estimation and feature matching may have a breakthrough, and the image registration will achieve better results. We wish to point out that, maybe its accuracy cannot be compared with the well-optimised methods exploiting additional optimisation algorithm for depth estimation such as Visual SLAM. However, our method does not rely heavily on surface features to predict depth, so it has the potential to be applied on low-texture surface scenes.
6 Acknowledgments
This study was funded by the Guangzhou Science and Technology Plan Project (nos. 201802010021 and 201806010128) and the Guangdong Science and Technology Plan Project (no. 2017A090905047).

7 References
[1] Rublee, E. R.: ‘ORB: an efficient alternative to SIFT or SURF’. Int. Conf. on Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 2564–2571
[2] Rosten, E.T.D.: ‘Fusing points and lines for high performance tracking’. IEEE Int. Conf. on Computer Vision (ICCV), Beijing, China, 2005, pp. 1508–1515
[3] Fua, M.C.L.S.: ‘BRIEF: binary robust independent elementary features’. European Conf. on Computer Vision, Crete, Greece, 2010, pp. 778–792
[4] Herbert, B., Andreas, E., Gool, L.V.: ‘Speeded-up robust features (SURF)’. Comput. Vis. Image Underst., 2008, 110, pp. 346–359
[5] Lowe, D.G.: ‘Object recognition from local scale-invariant features’. Int. J. Comput. Vis., 1999, 2, pp. 1150–1157
[6] Lowe, D.G.: ‘Distinctive image features from scale-invariant keypoints’, Int. J. Comput. Vis., 2004, 60, pp. 91–110
[7] Eigen, D., Fergus, R.: ‘Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture’. 2015 IEEE Int. Conf. on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2650–2658
[8] Eigen, D., Puhrsch, C., Fergus, R.: ‘Depth map prediction from a single image using a multi-scale deep network’. Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014
[9] Liu, F., Shen, C., Lin, G.: ‘Deep convolutional neural fields for depth estimation from a single image’. 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014
[10] Li, B., Shen, Ch.H., Dai, Y.Ch.: ‘Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs’. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1119–1127
[11] Wang, P., Shen, X., Lin, Z.: ‘Joint semantic segmentation and depth estimation with deep convolutional networks’. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015
[12] Yang, P., Ji, F.B.: ‘A fast and high accuracy registration method for multi-source images’, Optim., 2015, 126, pp. 3061–3065
[13] Deng, H. Y.: ‘Optimization of SIFT algorithm for fast-image feature extraction in line-scanning ophthalmoscope’. Optim., 2018, 152, pp. 21–28
[14] Godard, C., Mac Aodha, O., Brostow, G.J.: ‘Unsupervised monocular depth estimation with left-right consistency’. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, AC, 2016
[15] Geiger, A., Lenz, P., Urtasun, R.: ‘Are we ready for autonomous driving? The KITTI vision benchmark suite’. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2012