Proceedings of the 2009 IEEE
International Conference on Robotics and Biomimetics
December 19 -23, 2009, Guilin, China

3D Reconstruction Based on SIFT and Harris Feature Points

Keju Peng, Xin Chen, Dongxiang Zhou, and Yunhui Liu

Abstract—**This paper presents a new 3D reconstruction method using feature points extracted by the SIFT and Harris corner detector. Since the SIFT feature points can be detected stably and relatively accurately, the proposed algorithm first uses the SIFT matching points to calculate the fundamental matrix. On the other hand many of the feature points detected by the SIFT are not what we need for reconstruction, so by combining the SIFT feature points with the Harris corners it is possible to obtain more vivid and detailed 3D information. Experiments have been conducted to validate the proposed method.**

I. Introduction

3D reconstruction is a classical problem in the field of computer vision. With the rapid development of information technology, people have more and more needs for 3D model. In the initial stage of this study, people who studied the 3D model based on image sequences all first tried to calibrate the camera, get the intrinsic and extrinsic parameters of the camera, then through counter-projecting the matching points from two images to the 3D space a 3D reconstruction can be achieved; But it was a cumbersome process of the traditional camera calibration, each capture needs a new calibration. So, in some circumstances that the camera parameters need real time changes or could not use calibration object, this method was less practical and very difficult to achieve. Reconstruction based on uncalibrated image sequences do not need to determine the camera parameters in advance, it just captures a group of image sequences through a freely moving ordinary camera then the 3D coordinates of the object can be obtained after the intrinsic and extrinsic parameters have been calculated. This method has been a hot issue for years because it has the advantages of convenient, real time and low cost.

Reconstruction based on two images is the foundation of reconstruction based on image sequences, this paper carried out the reconstruction based on two freely captured images with invariable intrinsic parameters of the camera. How to extract and match the feature points accurately and stably is a key step for 3D reconstruction. On the current terms, the feature point detection algorithms which are effective and widely used should be the Harris [1] corner extraction algorithm and the SIFT (Scale Invariant Feature Transform) [2] feature extraction algorithm, but both of them have their own shortcomings. There is not a perfect solution to meet the needs of the researchers so far. Most of the previous literatures just use one of the two algorithms in their study; this paper makes an attempt at a combination of both of them to achieve a 3D reconstruction. Since the SIFT feature points can be detected stably and relatively accurately, the proposed algorithm first uses the SIFT matching points to calculate the fundamental matrix and then enhance the robustness of the fundamental matrix by RANSAC(Random Sample Consensus) [3] iterative algorithm; We use the optimized matrix to remove the false matches of the initial SIFT matching points and the Harris corners, finally, we reconstruct the optimized matching points and perform triangulation and texture mapping of the 3D spatial points. In this paper, the basic idea of the proposed algorithm is: Feature points detected by the SIFT are very accurate but many of them are not what we need for reconstruction, corners can better express the basic shape of objects, so by combining the SIFT feature points with the Harris corners it is possible to obtain more vivid and detailed 3D information. Compared with the method without corners, reconstruction of the proposed method is more closer to the real object.

Keju Peng is with College of Electronic Science and engineering, National University of Defense Technology, Changsha, 410073, China. Phone: +86-731-84573497; fax: +86-731-84514427; e-mail: keju009@nudt.edu.cn.

Xin Chen is with College of Electronic Science and engineering, National University of Defense Technology, Changsha, 410073, China.

Dongxiang Zhou is with College of Electronic Science and engineering, National University of Defense Technology, Changsha, 410073, China.

Yunhui Liu is with Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong，Hong Kong，China.

II. Extraction and Matching of SIFT Feature Points

SIFT is presented by David G Lowe; after summing up the existing technologies based on invariant, he proposed a feature matching algorithm in 2004 based on scale space, the algorithm can maintain invariance to image zooming, rotation or even affine transformation; SIFT is a successful algorithm in the field of feature matching, it can extract stable feature points. SIFT has been proven to be the most robust local invariant detector and descriptor among the others with respect to geometrical changes [4], Therefore, we utilize SIFT feature points to calculate the fundamental matrix and reconstruct the object.

The extraction of the SIFT feature points include the following steps: the detection of the extreme point in scale space, accurate localization of key point, assigning the main orientation of key point and creating the key point descriptor.

1) Scale space extreme detection: first build the DOG(Difference of Gaussian) pyramid of the image, compare a pixel with its 26 neighbors in 3×3 regions at the current and adjacent scales, $D(x, y, \sigma)$ represents the difference of the two adjacent scale images：

$$D(x,y,\sigma)=(G(x,y,k\sigma)-G(x,y,\sigma))*I(x,y)=L(x,y,k\sigma)-L(x,y,\sigma) \quad (1)$$

978-1-4244-4775-6/09/$25.00 © 2009 IEEE.

If a point is the maximum or minimum in the 26 neighbors among the current scale and the upper and lower scales of the DOG scale space, then we consider the point as a feature point in the scale.

2) Assigning the orientation for each key point, so that the detector has a character of rotation invariance.

$$m(x,y)=\sqrt{\left(L(x+1,y)-L(x-1,y)\right)^2+\left(L(x,y+1)-L(x,y-1)\right)^2} \quad (2)$$

$$\theta(x,y)=\tan^{-1}\left(\frac{\left(L(x,y+1)-L(x,y-1)\right)}{\left(L(x+1,y)-L(x-1,y)\right)}\right) \quad (3)$$

Formula (2) and (3) are the gradient magnitude and the orientation of pixel *(x,y)* at its scale *L(x,y)*. In actual calculation, a gradient histogram is formed from the gradient orientations of sample points within a region around the key point. The scope of the gradient histogram is $0\sim360^0$, each 10^0 represents a direction, so there are 36 directions in all. The highest peak in the histogram is detected as the dominant direction of the key point.

3） The creation of the key point descriptor. First the coordinates of the descriptor and the gradient orientations are rotated to the key point's orientation to ensure rotation invariance. The next step is to sample points within a 16×16 region around the key point, and then in each of the 4×4 plot calculate the histograms with 8 orientation bins. After accumulating the gradient magnitudes of the 4×4 region to the orientation histograms, we can create a seed point; each seed point is a 8-demensional vector, so each key point can create a 128 element feature vector.

SIFT uses the Euclidean distance between two feature vectors as the similarity criteria of the two key points and uses the nearest neighbor algorithm to match each other. Suppose a feature point is selected in image 1, the nearest neighbor feature point is defined as the key point with minimum metric distance in image 2. By comparing the distance of the closest neighbor to that of the second-closest neighbor we can obtain a more effective method to achieve more correct matches. Given a threshold, if the ratio of the distance between the closest neighbor and the second-closest neighbor is less than the threshold, then we think we have obtained a correct match.

III. EPIPOLAR GEOMETRY AND FUNDAMENTAL MATRIX

In two views of the same object taken from different view points, the relative positions of the matching points are restricted by epipolar geometry [5]. Epipolar geometry can be accurately expressed by fundamental matrix.

The epipolar geometry is shown in figure 1, I_1 and I_2 are two image planes taken from the same camera, P is a spatial point, C and C' are the optical center of the camera in different viewpoints, the connection line of C and C' intersects each image plane at the epipoles e and e' respectively, m and m' are the points projected from point P on two image planes. For two uncalibrated images captured from the same pinhole camera, epipolar geometry restriction is the basic relationship between them. As is shown in the figure, we can see that the corresponding point of m in I_2 is bound to lay on its epipolar line l_m'. According to epipolar geometry, the search area on two-dimensional plane is down to one-dimensional beeline, we can also optimize the initial matching points hereby.

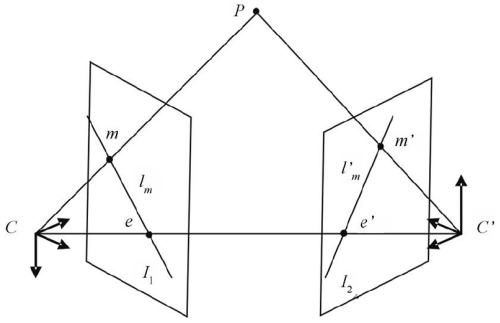

Fig. 1. Illustration of epipolar geometry.

Mathematically, the epipolar line restriction relationship can be described by fundamental matrix(3×3), suppose $m(x,y,1)^T$ is a point on I_1, its epipolar line equation on I_2 is $l_m'=Fm$, according to the epipolar geometry, $m'(x',y',1)^T$ must lie on the epipolar line lm' [8], the algebraic representation is:

$$m'^{T}Fm=0 \quad (4)$$

From (4)， we can see that fundamental matrix F can be solved if sufficient corresponding points are found in two images, F has nine parameters in total. After normalizing one of the non-zero parameters, there are eight unknown parameters left. So F can be solved as long as at least eight pairs of matching points are known. This is the 8-point algorithm to solve fundamental matrix. This paper first uses the 8-point algorithm to calculate the fundamental matrix and then enhances the robustness of the fundamental matrix by RANSAC iterative algorithm, once the fundamental matrix is obtained, we can use epipolar geometry relationship to refine the initial matching points. Points that have a long distance from the epipolar line are removed from the match list, thus we can further ensure the accuracy of the matching points.

IV. DETECTION AND MATCHING OF CORNERS

Corner is the point where the gray of the image dramatically changes or the junction of the contour boundary. It reflects the important information of the image. To extract corner can give prominence to the important information. This paper uses Harris corner detection algorithm to extract corners, Harris algorithm uses the rate of change of gray scale of the image to determine the corner. The proposition of this method is inspired by self-correlation function in signal processing. By carrying on the first order difference to the image, a matrix M which is linked to self-correlation function is given. Eigenvalues of matrix M are the first order curvatures of self-correlation function, if both of the curvature values are large then the point can be regarded as a feature point.

The formula of Harris operator only refers to the first derivative of the image:

$$M=G(\tilde{\sigma})\otimes\begin{bmatrix} G_x^2 & G_xG_y \\ G_xG_y & G_y^2 \end{bmatrix} \quad (5)$$

In (5), $G(\tilde{\sigma}) = \exp\left[-(x^2 + y^2 / 2\tilde{\sigma}^2)\right]$, it is a Gaussian smoothing filter, the purpose is to eliminate the unexpected image points to avoid these points being selected as the feature points. G_x is the gradient in the direction of gray scale x, G_y is the gradient in the direction of gray scale y. Harris corner can be defined as the maximum in local area by the following formula:

$$I = Det(M) - kTrace^2(M), k = 0.04 \quad (6)$$

In (6), *Det* and *Trace(M)* are the determinant and the trace of matrix M respectively, K is a default constant.

According to (5), calculate the first derivative in horizontal and vertical direction, as well as the product of them for each image point, after which three new images can be obtained. The attribute values of each pixel in three images can be represented by g_x, g_y, and $g_x g_y$ respectively. Then we carry out Gaussian filtering of the three images and calculate the interesting values of each corresponding point in the original image.

Harris algorithm thinks that, feature points are the corresponding pixel points with maximal interesting values in local area. Therefore, after calculating the interesting value of each point, we should extract all of the points with maximal local interesting values in the original image. In actual operation, we can extract maximum value within a 3×3 window around each pixel orderly, if the interesting value of the central pixel is the maximum, then the point is regarded as a feature point.

This paper performed the initial matching of corners by means of Zero-mean Normalized Cross Correlation(ZNCC) [6], It is a stereo matching method based on local area, however, ZNCC has a relatively large error rate in the deformative area because it just uses the gray information of the image. We also optimized the initial matching corners by the fundamental matrix refined after RANSAC, this paper performed the 3D reconstruction by means of the optimized SIFT matching points together with the corners.

V. CAMERA CALIBRATION AND 3D RECONSTRUCTION

Pinhole camera model is the commonly used model. The perspective projection transform from the point $\tilde{X} = [X, Y, Z, 1]^T$ in Euclidean 3D space to the 2D image point $\tilde{x} = [u, v, 1]^T$ can be described with a projective matrix P (3×4), here X and x are their homogeneous coordinates.

$$\tilde{x} = P\tilde{X} = K[R \mid t]\tilde{X} \quad (7)$$

Where K is the intrinsic parameter matrix of the camera; R and t, respectively, are the rotation matrix and translation vector of the camera relative to the reference coordinate system.

This article captured the same scene with the same digital camera in different views. So we can keep the intrinsic parameters of the camera unchanged, K can be obtained by zhang's method [7] in advance, let the camera coordinate system of the first image as the world coordinate system, then its projective matrix can be expressed as [4]:

$$P_1 = K[I \mid 0] = [K \mid 0] \quad (8)$$

Where I is the identity matrix (3×3), here we suppose there is no rotation and translation of the first image. Relative to the world coordinate system formed by the first image, the second image's rotation matrix and translation vector can be written as R and t, so the second image's projective matrix is:

$$P_2 = K[R \mid t] \quad (9)$$

As the camera has been calibrated above, so K and P_1 are known; only after the second image's R and t are calculated that we can obtain P_2. The 3D coordinates of the corresponding feature points can be calculated from the matching points of the two images by triangulation.

As the fundamental matrix contains the information of intrinsic and extrinsic parameters of the two images, therefore, we can obtain the essential matrix from the fundamental matrix, then the extrinsic parameters of the second image can be obtained after Singular Value Decomposition (SVD) [10] of the essential matrix. Essential matrix is defined as $E=K^TFK$. Theoretically [9], one of the eigenvalues of the essential matrix should be zero during the SVD, and the other two eigenvalues equal to each other. But in practice, the eigenvalue may not be zero due to the impact of the noise, so we can set the smallest eigenvalue to be zero, and set the average value of the other two eigenvalues as the two equal eigenvalues. A new essential matrix E' can be obtained from this diagonal matrix, decompose E' by SVD again, we can obtain two 3-order matrices (U and V) and a 3-order diagonal matrix S:

$$E'=USV^T, \quad \text{suppose } W = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (10)$$

The possible values for rotation matrix R are $R=UWV^T$ or UW^TV^T and the possible values for translation vector t may be u_3 or $-u_3$, in which u_3 is the final column of the matrix U. Therefore, there are four situations for the projective matrix P_2, they are:

$$\begin{aligned} &K[UWV^T \mid u_3], \quad K[UWV^T \mid -u_3] \\ &K[UW^TV^T \mid u_3], \quad K[UW^TV^T \mid -u_3] \end{aligned} \quad (11)$$

Because we selected the camera coordinate system of the first image as the world coordinate system, so the 3D coordinates of the feature points should be positive in the Z-axis, calculate the 3D structure of the object in four cases mentioned above and select R and t that right meet this condition, then we can obtain the final P_2.

As the projective matrices of the two images are obtained, we can calculate the corresponding 3D spatial coordinates of the matching points, Suppose P_{1i} and P_{2i}（i=1,2,3）are the three row vectors of P_1 and P_2 respectively, the projection of the 3D spatial point $\tilde{X}_w = (X, Y, Z, 1)$ in the form of homogeneous coordinate on the two images are $(u_1, v_1, 1)$ and $(u_2, v_2, 1)$, according to (7), for each image, two linear equations can be obtained independently, so that for each pair of matching points we can get [5]:

$$\begin{bmatrix} P_{13}u_1 - P_{11} \\ P_{13}v_1 - P_{12} \\ P_{23}u_2 - P_{21} \\ P_{23}v_2 - P_{22} \end{bmatrix} \tilde{X}_w = 0 \qquad (12)$$

As illustrated in figure 1, each point of the image can determine a ray that lets the optical centre of the camera as the endpoint, at least two rays intersecting with each other can determine the location of the 3D spatial point, due to the impact of noise, it is difficult for the two rays to intersect directly, so the least square method can be adopted from (12) to solve the corresponding spatial coordinate for each pair of the matching points lineally.

In order to obtain a genuine reconstruction of the object, this paper performed meshing and texture mapping by means of the triangulation technique after obtaining all the 3D coordinates of the discrete spatial points.

VI. Experimental Results

In order to verify the feasibility of the proposed algorithm, this paper conducted experiments with two images of a carton captured by a camera of UP-800 from the corporation of UNIQ, the resolution of the images are 1024×778, the two images used in the experiments are as follows:

Fig. 2. Original images.

We first calibrated the camera by Zhang's method, the intrinsic parameter matrix is:

$$K = \begin{bmatrix} 1138.81 & 0 & 535.107 \\ 0 & 1159.81 & 298.384 \\ 0 & 0 & 1 \end{bmatrix}$$

Matching points were extracted by SIFT, a total number of 672 pairs of matching points were obtained in the two images with the threshold of SIFT as 0.49, let the fundamental matrix calculated from the matching points as the initial value, optimized the fundamental matrix by RANSAC, the fundamental matrix after optimization is:

$$F = \begin{bmatrix} \text{-1.13446e-005} & \text{-0.000103091} & 0.0808436 \\ 0.000135 & \text{-2.29259e-006} & \text{-0.384779} \\ \text{-0.0635346} & 0.384512 & 1 \end{bmatrix}$$

The projective matrix of the image 2 calculated from the fundamental matrix is:

$$P_2 = \begin{bmatrix} \text{-1069.29} & 94.3755 & \text{-656.461} & 1240.33 \\ \text{-42.5842} & \text{-1147.84} & \text{-338.879} & 261.467 \\ 0.111941 & 0.0291443 & \text{-0.993287} & 0.334103 \end{bmatrix}$$

Optimized the initial matching points according to the theory of epipolar line restriction by the fundamental matrix calculated above, there were 624 pairs of matching points left after removing mismatches, the matching result of the SIFT feature points is shown in Fig. 3.

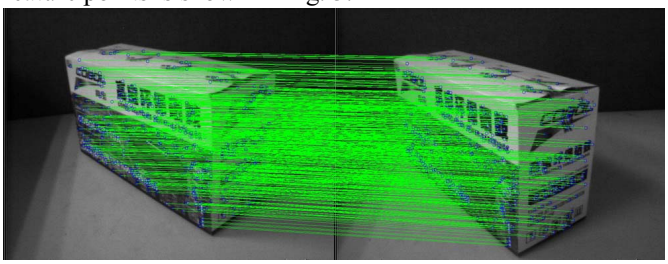

Fig. 3. SIFT matching points.

We can see from the corresponding matching points from Fig. 3 that the amount of the feature points from SIFT is relatively large and the points can match each other accurately, but many points which can express the key features of the carton, such as the points at the position of the corners and edges, are not detected by SIFT, this can lead to the details of the carton after reconstruction are not prominent. The results of the object reconstructed by SIFT matching points only are shown in Fig. 4.

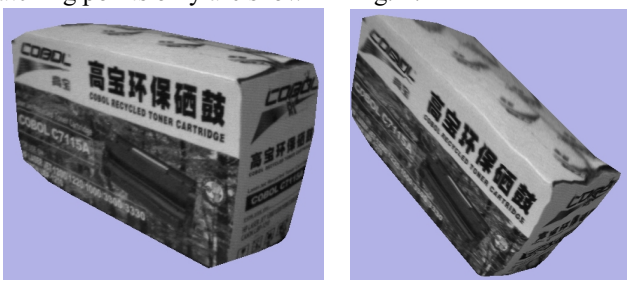

Fig. 4. Reconstruction results by SIFT.

As can be seen in Fig. 4, the edges and corners of the object are not well reflected and there is a big difference compared with the real one, this paper improved the reconstruction effect by Harris corners, the detection and matching results of the corners are shown in Fig. 5, all the pairs of matching corners in the figure have been optimized by fundamental matrix.

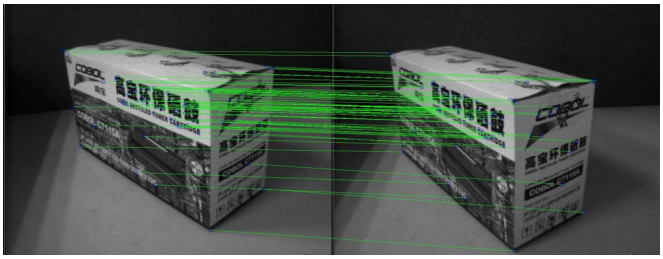

Fig. 5. Matching points of corner.

In the experiments, 66 pairs of corners were obtained in total, although we only got a small number of corners, but we can see that many key points of the object were detected, the final reconstruction results also proved that there was an obvious improvement of the reconstruction by combining the SIFT feature points with the Harris corners. The results of the experiments had proved the effectiveness of the proposed method.

Fig. 6. Reconstruction by SIFT and corners

VII. CONCLUSION

The previous literatures have referred to the idea of combining the SIFT feature points with corners, but no one has really given the solution on how to realize it, this paper combines the advantages of both of them and proposes a new 3D reconstruction method based on SIFT and Harris detector, we carried out experiments and also gave out the experimental results. Compared with the reconstruction using the SIFT feature points only, the proposed algorithm is more prominent in detail, the images in this paper are captured from a regular object, if the object is an irregular one, improvement of the reconstruction will be more obvious. Compared with the reconstruction using the corners only, the reconstruction results of the proposed algorithm are more accurate, if the texture of the image is rich enough, we could even obtain the effect of quasi-dense reconstruction, so the proposed algorithm makes full use of the advantages of both SIFT and Harris, the two algorithms complementary with each other, actual experimental results also confirmed the validity of the proposed algorithm.

It should be pointed out that how to match the corner more effectively still need the common work of a vast number of researchers as the question of matching has always been a difficulty in the field of computer vision. In this paper, we only got a small number of matching corners with ZNCC matching strategy, but the ultimate effect of the reconstruction had a great improvement even if constructed with few corners, so further work needs to be done on the accuracy of the corner matching. In addition, the experiments in this article were carried out based on two images, which were not enough for reconstructing the entire scene, we will do research on reconstruction based on image sequences in future.

REFERENCES

[1] C. Harris, M. Stephens. A combined corner and edge detector［A］. Proceedings of the 4th Alvey Vision Conference［C］. Manchester, UK：1998. 147-151.

[2] D. Lowe. Distinctive Image Features from Scale-Invariant Interest Points［J］. International Journal of Computer Vision，2004，60(2)：91-110.

[3] Fishler M A, Bolles R C. Random Sample Consensus, A paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of ACM, 1981, 24(6):381-395.

[4] K. Mikolajczyk and Cordelia Schmid, “A performance evaluation of local descriptors,” IEEE Trans. On PAMI, vol. 27. 2005, pp. 1615-1630.

[5] R. Hartley, A. Zisserman. Multiple view geometry in computer vision［M］.Cambridge University Press，2000.

[6] J. P. Lewis, Fast template matching, in Proc. Conf. on Vision Interface, May 1995, pp. 120-123.

[7] Zhang Z. A flexible camera calibration by viewing a plane from unknown orientations［A］. Proceedings of the 7th International Conference on Computer Vision［C］.Corfu, Greece：1999. 666-673.

[8] Luong Q-T, Faugeras O. The fundamental matrix: Theory, algorithms, and stability analysis. International Journal of Computer Vision, 1996.

[9] Xu G, Zhang Z. Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach. Kluwer Acamdemic Publisher, 1996.

[10] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. Numerical Recipes in C. Cambridge University Press, 1988.