Registration of Depth Image and Color Image Based on Harris-SIFT

Simin Zhao,Xiangming Xu,Weilong Zheng,Jianwen Ling
School of Electronic and Information Engineering, South China University of Technology
Guangzhou, China, 510641
zhao.sm@mail.scut.edu.cn *xmxu@scut.edu.cn* *zheng.wl@mail.scut.edu.cn* *jw.ling@mail.scut.edu.cn*

***Abstract*—In the recent past, it has become a widely accepted and applied method to recognize and localize objects based on local point features. However, the existing approaches are based on matching color images. With the widely use of Kinect, the registration between depth image and color image becomes necessary and important. To effectively match 3D depth images and 2D color images and estimate the transformation homography, we present a type of features, which combines Harris corner detector with SIFT descriptor. In the experiments, the corresponding feature points between depth image and color image can be well located by the descriptor, which is invariant under blurring, rotation, shift, scaling and moderate changes in viewpoint.**

Keywords-Harris-SIFT; depth image; registration

I. INTRODUCTION

Recent developments in gaming technology, such as the Nintendo Wii, the SONY PlayStation Move and Microsoft Xbox360 Kinect, focus on robust motion tracking for compelling real-time interaction, while the geometric accuracy and appearance are of secondary importance. Kinect simultaneously captures 3D depth images and 2D color images at the rate of 30 frames per second. Essential benefits of this low-cost acquisition device include ease of deployment and guaranteed operability in natural environment. The user is not required to wear any physical markers or specialized makeup, and the performance is not adversely affected by intrusive light projections or clumsy hardware contraptions, either. However, these key advantages are achieved at the cost of a substantial degradation in the data quality comparing to the state-of-the-art performance capture systems based on markers or active lighting. As for further processing of depth image and color image, the registration between them becomes necessary and important. We thus propose to register depth image and color image to alleviate the problem of misalignment.

As we know, the registration and alignment of images taken by multi-source cameras at different spatial locations and orientations in the same environment is a vital task to many applications in computer vision. Feature extracting and feature matching are two difficult tasks in automatic image registration. Feature points in scale space, which are affine-invariant, provide a good way for feature extracting and matching. David G. Lowe [1] proposed a new feature extracting method named scale-invariance feature transform (SIFT), which have been applied to feature extracting and matching widely and proved to be steadily invariant to image scaling and rotation [2]. Jian Gao [3] applied a normalized Laplace detector to deal with image color information and perform real-time feature extracting and matching. Krystian Mikolajczyk [4, 5] proposed the Harris-Laplace detector and proved that it has a better performance in repeatability, localization error and scale changing than other detectors in scale space. Pedram Azad [6] combined Harris interest points and the SIFT descriptor for fast scale-invariant object recognition between two color images. All these efforts have been proved to work well in registrating intensity images. However, it is to be further investigated whether these features can also work well on the registration of depth image and color image.

In this paper, we present one type of features, which combines Harris corner detector with SIFT descriptor. In order to achieve scale-invariance in spite of omitting the scale space analysis step of the SIFT features, the features are explicitly computed at several predefined spatial scales. The main orientation of each feature point is built in its neighborhood and a feature descriptor is also constructed in the direction of its main orientation. The scale space projection makes feature space resolution-invariant and the feature's main orientation makes it rotation-invariant. In addition, the feature matching combines Euclidean distance with RANSAC to gain a better matching. From the experiments, the corresponding feature points between depth image and color image can be well located by the descriptor, which is invariant under blurring, rotation, shift, scale, and moderate changes in viewpoint and works perfectly in different kinds of transformation between images. In 2009, Wagner et al. had developed a similar approach based on the same idea, combining SIFT descriptor and Ferns descriptor [7] together with FAST detector [8], as presented in [9]. In this paper, the original SIFT descriptor is combined with Harris corner detector, and all parameters are derived from a thorough analysis of the scale coverage of SIFT descriptor.

This paper is organized as follows. In Section II, the feature points of the depth image and the color image are detected. The details of feature point detection based on Harris corner detection are introduced. In Section III, we develop a SIFT-like descriptor to describe the feature points of the depth image and the color image. In Section IV, the nearest neighbor algorithm and RANSAC algorithm are used to match the corresponding feature points. Experimental results and analysis are presented in Section V.

Finally, we draw our conclusions in Section VI.

II. FEATURE POINT DETECTION

A crucial component of our registration algorithm is how to find the robust feature that can properly describe the image. Stable feature points can be detected in both depth image and color image with very diverse methodologies. We use a local feature point descriptor to match the depth image and the color image in this paper.

Considering the huge difference between 3D depth images and 2D color images, we should firstly preprocess both images respectively. The low resolution and high noise levels of the input data are the primary challenges that we address in this paper. Then image preprocessing is necessary. As for depth image, we smooth them among frames. We achieve depth pixel value as the median pixel values among the successive 11 frames. In practice, to enhance the contrast of depth image and find more stable feature points, we also apply the histogram equalization to depth image. As for color image, we blur it through the median filter with a 5×5 window.

In order to localize feature points, we use multi-scale Harris-Laplace corner detector. The Harris-Laplace detector relies heavily on both the Harris measure and a Gaussian scale-space representation.

The Harris-Laplace detector combines the traditional 2D Harris corner detector with the idea of a Gaussian scale-space representation in order to create a scale-invariant detector. Harris-corner points are good starting points as they have been proven to have good rotational and illumination invariance as well as the ability to identify the interesting points of the image [10]. However, the points are not scale-invariant and thus the second-moment matrix must be modified to reflect a scale-invariant property. Let us denote M as the scale adapted second-moment matrix used in the Harris-Laplace detector [11].

$$M=\mu(\omega,\sigma_I,\sigma_D)=\sigma_D^2 g(\sigma_I)\otimes\begin{bmatrix} L_x^2(\omega,\sigma_D) & L_xL_y(\omega,\sigma_D) \\ L_xL_y(\omega,\sigma_D) & L_y^2(\omega,\sigma_D)\end{bmatrix} \quad (1)$$

Where $g(\sigma_I)$ is the Gaussian kernel of scale σ_I and $\omega=(x,y)$. Similar to the Gaussian-scale space, L_x is the Gaussian-smoothed image. The $\otimes$ operator denotes convolution. $L_x(\omega,\sigma_D)$ and $L_y(\omega,\sigma_D)$ are the derivatives in their respective directions applied to the smoothed image and calculated using a Gaussian kernel with scale σ_D. In terms of our Gaussian scale-space framework, the parameter σ_I determines the current scale at which the Harris corner points are detected. Building upon this scale-adapted second-moment matrix, the Harris-Laplace detector is a two-fold process: applying the Harris corner detector at multiple scales and automatically choosing the characteristic scale.

The algorithm searches over a fixed number of predefined scales. This set of scales is defined as:

$$\sigma_n=k^n\sigma_0(n=1,2,3...) \quad (2)$$

For each integration scale σ_I, chosen from this set, the appropriate differentiation scale is chosen to be a constant factor of the integration scale: $\sigma_D=s\sigma_I$. In our project we use $k=1.5$ and $s=0.7$. Using these scales, the interest points are detected through a Harris measure on the $\mu(\omega,\sigma_I,\sigma_D)$ matrix. The HrF, like the typical Harris measure, is defined as:

$$HrF=\det(\mu(\omega,\sigma_I,\sigma_D))-\alpha\,\mathrm{trace}^2(\mu(\omega,\sigma_I,\sigma_D)) \quad (3)$$

Like the traditional Harris detector, corner points are those local maxima of the HrF that are above a specified threshold. For the depth image and the color image, we use different α and the specified threshold.

We then verify for each of the initial points whether Laplacian-Gauss response attains a maximum at the scale of the point. We reject points for either the Laplacian attains no maximal or the response is below a threshold. If the Harris feature makes the normalized Laplace function $\mu(\omega,\sigma_n)=\sigma_n^2\left|L_{xx}(\omega,\sigma)\right|+\left|L_{yy}(\omega,\sigma)\right|$ satisfies:

$$\begin{aligned}&\mu(\omega,\sigma_n)>\mu(w,\sigma_{n-1})\\&\mu(\omega,\sigma_n)>\mu(w,\sigma_{n+1})\\&\mu(\omega,\sigma_n)>\mathrm{threshold}_L\end{aligned} \quad (4)$$

Then, this point is chosen as a Harris-Laplace feature point. In this way we obtain a set of characteristic points with associated scales. For some points the scale peak might not correspond to the selected detection scales of an image. Either these points are rejected due to the lack of a maximum, or their location and scale are not very accurate. Thus it is necessary to have a small scale interval between two successive levels in order to find the location and scale of an interest point with high accuracy. With Normalized Laplace operator we find the characteristic scale for each feature point. If the feature point has no characteristic scale, it's rejected. Similarly, the localization of the feature points is calculated in sub-pixel accuracy, with Taylor approximation.

III. FEATURE POINT DESCRIPTOR

SIFT [12] is first presented by David G Lowe in 1999. As already stated, the SIFT descriptor is a very robust and reliable representation for the local neighborhood of an image point. However, the scale-space analysis requiring for the calculation of the SIFT feature point positions is too slow for real-time applications. We combine the Harris

corner detector with the SIFT descriptor in our project. In order to omit the scale space analysis step of the SIFT features and reduce the computational complexity, the features are computed at several predefined spatial scales explicitly.

Through the previous process, we have detected the positions of the feature points and their characteristic scale. Similarly we can use the SIFT descriptor to describe the detected feature points for further registration. For each feature point in those neighborhoods calculated derivatives and with weighted histogram defined Main Orientation. Each key point is assigned with consistent orientation based on local image properties so that the descriptor has a character of rotation invariance. This process is similar to Lowe's SIFT descriptor. This step can be described by two equations below:

$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^2+(L(x,y+1)-L(x,y-1))^2} \quad (5)$$

$$\theta(x,y)=\tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \quad (6)$$

It's possible to influence this process: define the size of the angle bins, if to take only those feature points that have one very supported direction or to allow for the feature points to have more than one main orientation, that are bigger than some thresh. The highest peak in the histogram is detected and then any other local peaks with 80% of the highest peak value can be assigned as the main orientation for option. In practice, we find that the feature points with more than one main orientation are preferable.

After the main orientation is calculated, SIFT-like descriptor is calculated for each feature point with main orientation. Its difference from SIFT is that when calculating derivatives and their orientations in neighborhoods of the feature points, SIFT-like descriptors are taken in the direction of main orientation near each feature point. The description is taken samples of them and the step of sampling is taken dependently on characteristic scale, which is $\sqrt{\sigma^2+1}$. It is a function that specifies step, as function of scale. If the specified function is constant 1, it's possible to take regular SIFT descriptor. Also possible tunable options are specified number of windows for the description and size of angle bins.

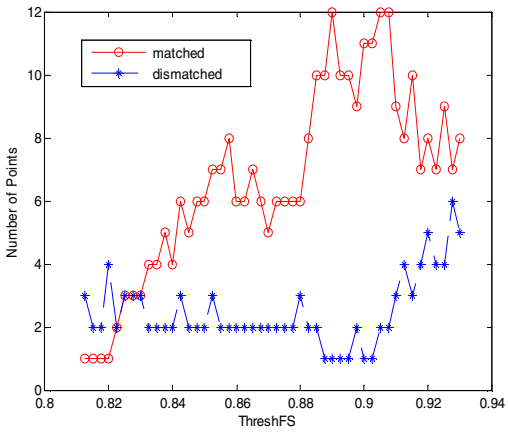

Figure 1. Matched and mismatched points with different ThreshFS

Each feature point is described by a 4×4 region around, each of which is further divided into 4×4 sub-regions consisting of 16 pixels. Then in each of the 4×4 sub-region, calculate the histograms with 8 orientation bins. After accumulation, the gradient magnitudes of the 4×4 region to the orientation histograms, we can create a seed point, an 8-demensional vector. Therefore, a descriptor contains $4\times4\times8$ elements in total. This vector is then normalized in order to enhance invariance to changes in illumination.

IV. CORRESPONDING POINTS REGISTRATION

As feature matching requires high precision, we combine Euclidean distance initial match with RANSAC instead of one single method.

In order to match the corresponding points, we firstly use Euclidean distance between two feature vectors as the similar criteria of two key points and use the k-nearest neighbor algorithm for registration. By comparing the distance of the closest neighbor with that of the second-closest neighbor we can obtain a more effective method to achieve greater accuracy. Given a threshold, if the ratio of the distance between the closest neighbor and the second-closest neighbor is less than the threshold, we then obtain a correct match.

As for the filtered set of feature correspondences resulting from the RANSAC algorithm, now a full homography is estimated with a least squares approach.

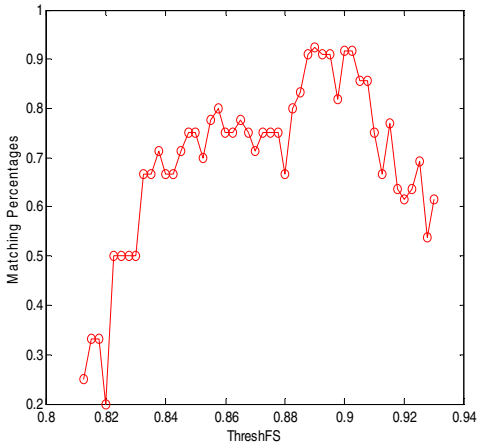

Figure 2. Percentages of correct matches at different ThreshFS

V. EXPERIMENTAL RESULTS

A set of depth images and color images are used to test

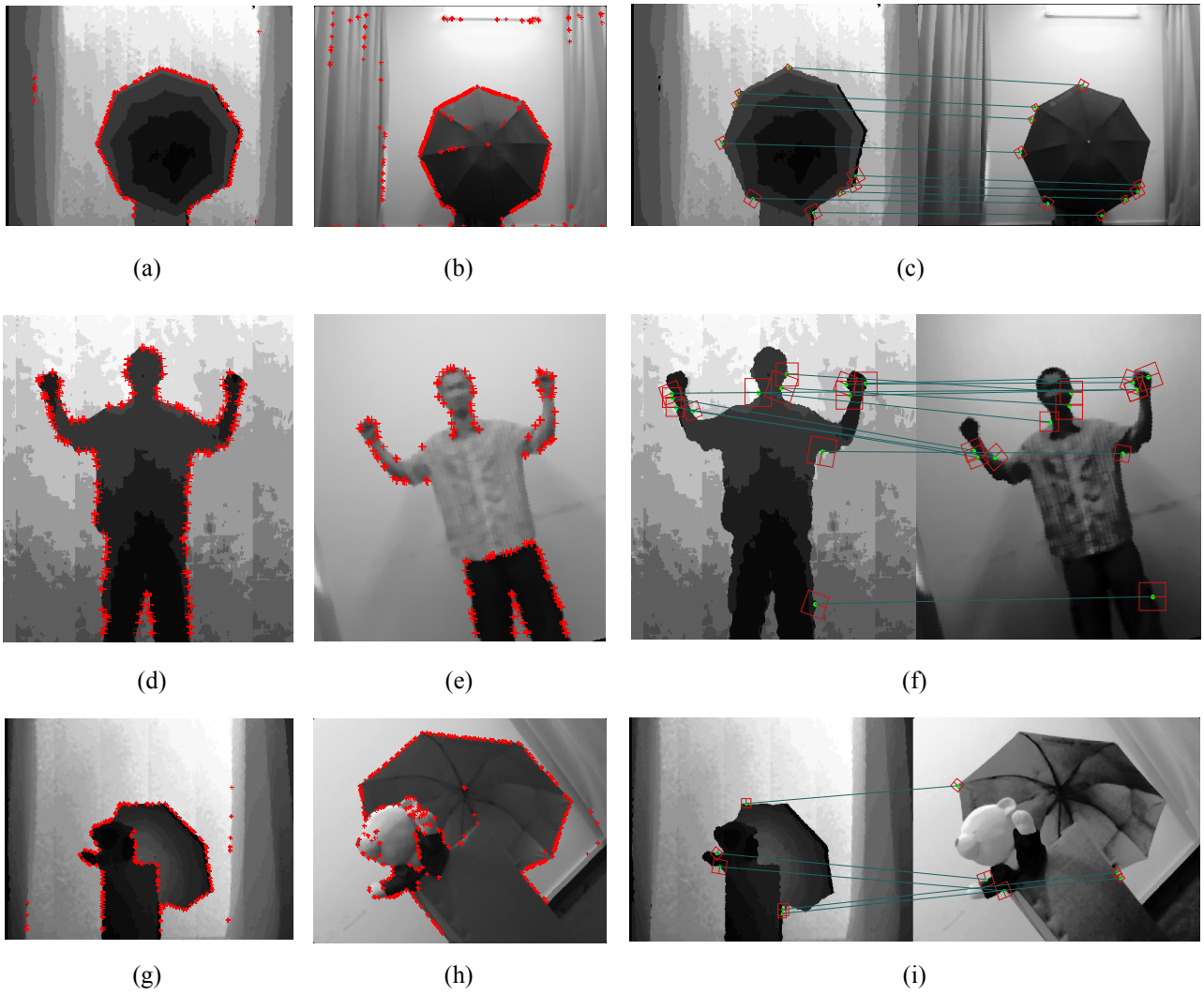

Figure 3. (a)(d)(g) feature points of depth image (b)(e)(h) feature points of color image (c)(f)(i) corresponding feature points between depth image and color image. The color images of first row and second row are captured by Kinect.

the performance of our algorithm in this section. Except for the specified ones, all depth images are captured by Kinect and their corresponding color images are captured by high definition camera. Depth images are specifically processed for display. In matching features, an effective measure is obtained by comparing the distance of the nearest neighbour (NN) to that of the second-nearest neighbour (SNN) and omitting the correspondences in which the ratio NN/SNN is bigger than ThreshFS. ThreshFS is the parameter to be determined. Figure 1 plots the results of matched and mismatched points with different ThreshFS. In Figure 2, the percentages of correct matches at different ThreshFS are plotted. Throughout all experiments, the scale number is 6 and Harris thresh is set to 0.01 of the maximum Harris measure HrF for depth image and 0.005 for color image. In Harris corner detection, the minimal distance between two feature points is 2 pixels. The quality threshold was set to 0.05 for depth image and 0.005 for color image in order to produce enough feature points. From the results, we can find that percentages of correct matches reach maximal when ThreshFS is approximately 0.9.

We firstly test the ability of our algorithm on highly symmetric and similar objects. Figure 3 (a)~(c) show a sample of two nearby views of an object (an umbrella). Each square shows one matched feature, with the location, size, and orientation of the square indicating the corresponding parameters for the Harris-SIFT feature. As can be seen, there are about 9 matched points. The algorithm has been proven to work well in those highly symmetric and similar images. Secondly, we would like to test the ability of our algorithm to symmetric and more complex objects. Figure 3 (d)~(f) show a sample of another target, a person. The color image is captured by a rotated camera. As we can see, when the person is standing in front

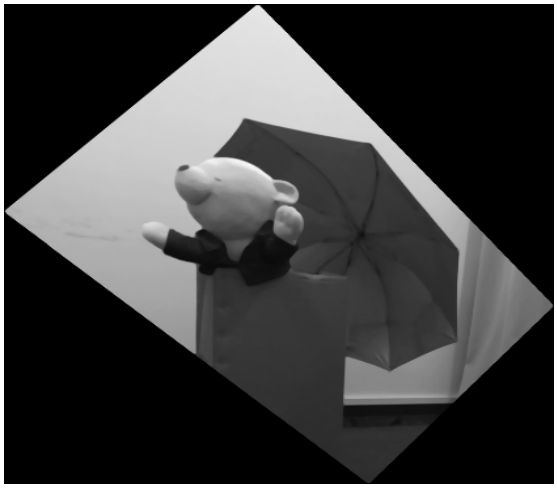

Figure 4. The result of the transformation of color image

of the Kinect, the descriptor can still match 11 corresponding feature points. Experimental results show that the algorithm can effectively reduce the influence of camera rotation.

To evaluate the accuracy of our algorithm, we get the homography estimation from the registration and show the result of transformation. Figure 3 (g)~(i) show an example of two different viewpoints images and the result of the registration. Figure 4 shows the transformation from color image to the view of the depth image using the homography estimation. From the result, we can see that the depth image captured by Kinect and the color image captured by common camera can also be well matched. The transformation of the color image can well match the original depth image captured by Kinect, which shows the homography estimation is accurate enough. The descriptor we use is invariant under blurring, rotation, shift, scaling and moderate changes in viewpoint.

VI. CONCLUSIONS

In this paper, we present a type of features, which combines Harris corner detector with SIFT descriptor. The practical applicability of our approach is verified by extensive experimental evaluation. In the experiments, the descriptor is invariant under blurring, rotation, shift, scale and moderate changes in viewpoint and works perfectly in different transformation between two types of images. The corresponding feature points between depth image and color image can be well located. The robustness of the method is sufficient for a range of purposes. For example, with the registration between depth image and color image, we can conveniently use the color image information to improve the depth accuracy and help three-dimensional reconstruction. Since registration can also combine information from aligned depth image and color image and severely reduce the noise in the data, it helps to turn depth cameras and color cameras into a viable tool for 3D shape scanning.

ACKNOWLEDGMENT

This work is supported by the Fundamental Research Funds for the Central Universities, the Science and Technology Planning Project of Guangdong Province under the grand No.2009B080701060 and No.2010A080402015.

REFERENCES

[1] Lowe, D.G., "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[2] Juan Geng, Yan Li and Tao Chian., "SIFT based Iris feature extraction and matching," Proceedings of SPIE, 2007, Vol. 6753, 67532F.

[3] Jian Gao, Xinhan Huang and Bo Liu, "A quick scale-invariant interest point detecting approach," Machine Vision and Applications, 2008.

[4] Krystian Mikolajczyk and Cordelia Schmid, "Indexing Based on Scale Invariant Interest points," In Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, pp.525-531,2001.

[5] Krystian Mikolajczyk and Cordelia Schmid, "An Scale and Affine Invariant Interest Point Detectors," International Journal of Computer Viseion, vol. 60, no. 1,pp. 63-86,2004.

[6] Azad, P., Asfour, T. & Dillmann, R., "Combining Harris Interest Points and the SIFT Descriptor for Fast Scale-Invariant Object Recognition," In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, USA, pp.4275–4280, 2009.

[7] Mustafa Ozuysal, Pascal Fua and Vincent Lepetit, "Fast Keypoint Recognition in Ten Lines of Code," IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[8] E. Rosten and T. Drummond, "Machine Learning for High-Speed Corner Detection," European Conference on Computer Vision (ECCV), Graz, Austria, pp. 430–443, 2006.

[9] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond and D. Schmalstieg,"Pose Tracking from Natural Features on Mobile Phones," International Symposium on Mixed and Augmented Reality (ISMAR),Cambridge, UK, pp.125–134, 2008.

[10] Schmid, C., Mohr, R., and Bauckhage, C. "Evaluation of interest point detectors," International Journal of Computer Vision, vol. 37, no. 2, pp.151-172, 2000.

[11] Mikolajczyk, K. and Schmid, C. "Scale & affine invariant interest point detectors," International Journal on Computer Vision, vol. 60, no. 1, pp.63-86,2004.

[12] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, pp.1150–1157,1999.

[13] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24, no. 6, pp.381–395, 1981.