A Method of Cylindrical Panoramic Video Stitching

Jie Chen¹, Xiaoyu Li² and Xiao Yu³

¹Science and Technology on Information Systems Engineering Laboratory, 99th Houbiaoying Road, Nanjing, China.
²The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing, China.
³Computer Science and Technology Department, Nanjing University, Nanjing, China.
Email: chenjie600006@126.com.

Abstract. In this paper, we propose a cylindrical stitching method which based on template frame image to realize 360° panoramic video stitching. Firstly, we selected template frame to determine the camera parameters and distortion coefficient. And, removing the image distortion. After that, extracting the template frame feature and making rough matching with the feature points, and then eliminating the feature points of mismatching by the improved feature point screening method. Then, feature points are projected on the cylinder in advance for subsequent video image fusion. Finally, image fusion is carried out after the cylinder projection of the video image. The speed was greatly improved in the subsequent video splicing by using camera parameters, distortion coefficient and transformation matrix of template frame image. In addition, the CUDA platform was adopted for GPU parallel processing in order to further improve the speed to meet the real-time requirements. Finally, we realize real-time video stitching with 30 frames per second.

1. Introduction

Video stitching technology splices multiple small-field video with overlapping areas into a complete large-field scene. It effectively solves the limitation of single camera field of vision, and has a wide range of applications in virtual reality, intelligent monitoring and other fields. Video stitching usually has two video stitching method as following [1]: by moving or rotating video acquisition device, such as dynamic observation by rotating cloud terrace around the scene, but this can only be at a certain moment in a local vision; The second method is to collect video images at the same time through multiple video acquisition devices, and then conduct feature matching and image fusion for video image frames to complete mosaic. This paper adopts the second method.

Video is made up of images arranged in chronological order, so the essence of video stitching is image stitching and image matching is the core step in image stitching. David G Lowe proposed the SIFT (Scale-invariant feature transform) [2, 3] which is often used in image matching, but the SIFT algorithm calculation speed is slow which cannot satisfy the requirement of real-time video. So it is not suitable for application in real-time video stitching. Based on SIFT algorithm, Bay H et al. proposed the SURF (Speeded Up Robust Features) algorithm [4]. The research of Bauer J, et al. shows that: SURF algorithm improves the speed of feature detection, and has basically the same performance as SIFT algorithm [5]. But video stitching and image stitching and there is a big difference, the video stitching needs to have good real-time performance, although SURF is faster, but it still needs improvement in the video stitching. Lum H et al. improved SURF algorithm, but the stitch speed still fails to meet the real-time requirements [6]. Therefore, in this paper, we proposed an improved SURF algorithm for image registration, in order to apply SURF to real-time video stitch.
In this paper, the video stitching process is: collecting the video frames real-time, and then extracting video frames at a rate of 30 frames per second. Firstly, the template frame is extracted. Then, the transformation matrix is determined between multi-channel video template frame using improved SURF algorithm. The subsequent frame image is fused by using the transformation matrix to forming a panorama image, finally the panorama image output in the form of video.

2. Video Frame Acquisition and Distortion Correction

N-channel (N is even) high definition camera consists of a ring hardware module. Two adjacent cameras in the module have the same angle. It collects the 360° video sources in the real environment. We use the distortion correction algorithm to correct the distortion of each camera. Firstly, each camera according to the difference of angle, distance to collection images which contain checkerboard used for calibration. Every frame of video needs to correct in theory, but this algorithm to calibrate the template frame for the first time, then we save the camera’s internal and distortion coefficient of correction for subsequent video frames to reduce the distortion correction time each frame of video. It can significantly increase the speed of algorithm.

2.1. Camera Internal Parameter Calibration

By camera calibration, the model parameters of camera can be obtained from the image coordinates and world coordinates by the known characteristic points. The model parameters which used for calibration include internal parameters and distortion parameters. We use the model parameters to correct the distortion [7].

Using linear pinhole model to represent the relationship between the image coordinates and the world coordinates. The points on the world coordinate are recorded as: \( M = [x, y, z]^T \). The 2D point on the image plane is denoted as \( m = [u, v]^T \). Then the mapping relation between spatial point \( M \) and image point \( m \) is: \( sm = A[Rt]M \), Where \( S \) is an arbitrary non-scale factor, \( R \) and \( t \) respectively represent the rotation matrix and translation vector of the external parameters of the camera, and \( A \) is the internal parameter matrix of camera which is defined as:

\[
A = \begin{bmatrix}
\alpha_x & 0 & u_0 \\
0 & \alpha_y & v_0 \\
0 & 0 & 1
\end{bmatrix}
\]  

(1)

Where \((u_0, v_0)\) is the optical center, \(\alpha_x \) and \(\alpha_y \) are scaling factors of \(X\) axis and \(Y\) axis respectively. \(\alpha_x = f / d_x \), \(\alpha_y = f / d_y \). The \(d_x\) and \(d_y\) are the pixel spacing of horizontal and vertical direction. The camera calibration method based on 2D planar checkerboard proposed by Zhang Z Y is used in this Paper. In this method, a camera is used to take a picture of a planar checkerboard [8]. The mapping matrix between the target plane and its image plane is calculated. The essence of the process is to minimize the residual difference between the actual coordinates \(h_i\) and the image coordinates \(\tilde{h}_i\) which calculated by the single-response matrix. The objective function is:

\[
\min \sum_i (h_i - \tilde{h}_i)^2
\]  

(2)

The inner parameter matrix and optical center coordinates are obtained.

2.2. Image Distortion Correction

In this paper, we use the calibration method based on 2D planar template proposed by Zhang Z Y, and the method just requires two images of the same calibrating template from different angles to find the internal and external parameters of the camera. It doesn't need to know the exact location and the displacement information of the plane template, and it's easy to make the plane template, so it's simpler and more flexible than any other method.

We use the camera internal parameters and distortion coefficient matrix to correct the distortion of video frames. The effect is as follows:
3. Cylinder Image Panoramic Stitching Algorithm

The next step after distortion correction is image registration. Image registration aligns overlapping parts of images which is the core part of the image stitching. The image to be stitched is converted to the coordinate system of the reference image to form a complete image. The image registration includes three aspects: feature extraction, feature matching and determination of transformation matrix.

Firstly, the feature of template frame image is extracted and matched by the extracted feature points. This algorithm uses SURF, which is improved on the basis of SIFT algorithm, to extract feature points. The image feature extraction includes two steps: feature detection and feature description. Feature detection includes two processes: constructing scale space and determining extreme point.

3.1. SURF Feature Points Matching

In order to speed up the matching process, SURF in the feature vector has added a new variable, namely the characteristic points of Laplace corresponding plus or minus, the two types of feature points with different attributes, in only when the matching of the same kind of feature point matching, this will greatly high matching speed and accuracy. Between two feature points in the image similarity measure of equidistance function commonly used Euclidean distance measure, through similarity metric can be stay registration the potential matching points between images on. This is actually a distance retrieval problem, if you use the exhaustive rule to waste too much time. As the standard $k$-$d$ tree is used, the dimension of the data set should not exceed 20 [9], while the improved SURF dimension is 16, so $k$-$d$ tree is adopted for feature matching.

Ratio matching method are often used: One of the images will be taken as a sample, and the feature points of the sample will be searched for the nearest feature points and the sub-near feature points in the other image. Then calculated the two feature points and the ratio of Euclidean distance between sample points. For feature points whose ratio is less than a certain threshold range, they are considered to be correctly matched feature points.

$$\frac{d_{Nearest}}{d_{NearestNearest}} \leq \text{threshold}$$

(3)

Original SURF algorithm is to extract the feature points of the whole image area, and then within the scope of the whole image feature matching, because the real environment images of two cameras there will always be the existence of the overlapping area, so just on a feature point extraction can overlap, such benefits is can reduce the extraction time, thus reduce the matching time and reduce false match.

Due to the mismatching of the extracted feature points, the matching feature points are efficiently filtrate according to the size of the overlapping area between adjacent cameras.

The filtering algorithm is as follows: set the Width of the image as the $W$, overlap ratio of factor, then the Width of the overlapping area as the $factor \times W$, the left image feature points as the Left-Points, on the right side of the image feature points for Right-Points. then filtering method is: if Left-Points is greater than the $factor \times W$, and is less than the Width, the matching points, otherwise discarded; If Right-Points is less than $(1-factor) \times Width$ and greater than 0, the match point remains, otherwise
discarded. The algorithm can eliminate the matching feature point pairs outside the overlapping area and improve the precision of feature matching. The error matching feature point pairs outside the overlapping areas are avoided and the program time is reduced. Figure 2 shows the comparison effect of matching feature point filtering.

![Figure 2. Comparison effect of matching feature point filtering.](image)

3.2. Image Cylinder Projection

We need to perform a cylindrical projection transformation on the extracted selected height matching feature points in the plane, then we can accurately calculate the amount of transformation required for the two fusing images later.

![Figure 3. Cylinder projection.](image)

As is shown in figure 3, two cameras circular arrangement, the image captured by each camera indicated by a solid line in front of it, image stitching is to fuse the two images according to the overlap. However, due to the different orientation of the two cameras in the figure 3 (a), the image does not meet the visual consistency. Therefore, the image is projected onto the cylinder surface, that is, on the arc in front of the camera. The overlapping areas of the projected image overlap with the cylinder surface, which has good visual consistency. In addition, cylindrical images can be easily expanded into rectangular images, which can be directly used in image processing. In figure 3 (b), the middle point $O$ is the position of the camera, and the rectangle in front is the image taken, and the cylinder projection is to project it onto the red cylinder.

Since image after cylinder projected has good visual consistency, so we first project the source image onto a cylinder and then stitched together.

3.3. Determine the Transformation of Two Fusing Images

The set of highly matched feature points after cylinder projection is used for image alignment fusion. The image alignment and fusion operation of our algorithm is very simple. It only needs to use the matching feature point set to calculate the translation between images and then shift the source image to the target image. There are many ways to calculate the translation amount of image stitching. We took the average translation amount of all matching feature points as the translation amount of the image, and the calculation formula is (4):
\[
\begin{align*}
\text{diff}_x &= \frac{\sum_{i=1}^{N} (W_i - P_{i,x} + P_{i,x})}{N} \\
\text{diff}_y &= \frac{\sum_{i=1}^{N} (P_{i,y} - P'_{i,y})}{N}
\end{align*}
\] (4)

Where \( W_i \) is the target image of ammonia, \( P_{i,x} \) is the feature points on the target image, \( P'_{i,x} \) is the corresponding feature points on the source image, \( N \) is the number of feature points matching, \( \text{diff}_x \) is sliding on the \( X \) direction, \( \text{diff}_y \) is sliding on the \( Y \) direction.

The image stitching result \( I_{res} \) can be obtained from source image \( I'_{src} \) and \( I'_{tar} \) by translation according to formula (5):

\[
I_{res}(x, y) =
\begin{cases}
I_{res}(x, y), & x \in [0, left) \\
I'_{res}(x - left, y - \text{diff}_x) \times X_{factor} \\+ I_{res}(x, y) \times (1 - X_{factor}), & x \in [left, W'_{tar}] \\
I'_{res}(x - left, y - \text{diff}_x), & x \in [W'_{tar}, W]
\end{cases}
\] (5)

Where, \( \text{left} = I'_{tar}(x, y) - \text{diff}_x \), \( X_{factor} = (x - \text{left}) / |\text{diff}_x| \). The process of image fusion is simply stated that the left part of the image is taken from the left, the right part is taken from the right, and the middle overlap area is the linear representation of the two images. Above is the translation calculation of 1&1 image stitching. Before multiple cameras video stitching, it used the formula (4) to calculate the transformation matrix (translation of 1&1 camera video) between the pairs of camera video frame in advance. we took 8-channel cameras for example, following 2&2, 4&4 image stitching translational quantity calculation should be based on the change of the previous splicing matrix accumulative. The translation of each time is kept for the subsequent video frame image stitching fusion.

3.4. Template Frame Fusion
When the translation between images is calculated, it can be fused. Prior to the fusion of two images, the cylinder projection of the image based on plane coordinates is needed.

In cylindrical panorama stitching process, in order to keep the actual scene space constraint relations and practical visual consistency of the scene, the overlapping image sequence must be mapped to a standard cylindrical coordinates space on the cylindrical projection, using the above cylindrical projection to get the cylindrical projection image of source image. The image of each camera is projected on the cylinder, and the image sequence of the cylinder is obtained.

The image after cylindrical projection can be used for image alignment and fusion. The specific method is: take left image as the target image, the right image as the source image, and the source image can perfectly fuse with the target image under the action of the translation quantity. Figure 4 (a) are the source images. Figure 4 (b) shows the images after cylindrical projection. Figure 4 (c) shows the fusion image.
4. Subsequent Video Frame Stitching

The camera parameters, distortion system and image translation required in the subsequent video frame image fusion process have been obtained in template frame processing, so subsequent video frame fusion no longer need camera calibration, feature detection and matching and translation calculation, which plays a significant role in the improvement of algorithm speed.

In addition, the cylindrical projection transformation and image fusion process can make use of GPU CUDA platform parallel acceleration. Therefore, the processing speed of subsequent video frames is faster, and real-time stitching can be achieved. Experimental results show that running the algorithm on NVIDIA 1080Ti GPU can generate seamless and clear panorama video in real time. See figure 5.

5. Acknowledgements

This work is supported by the pre-research project in 13th five-year(31511040202).

6. References

[1] Zhaoyu Shou, Ning OuYang, Huajun Zhang. Real-time video stitching based on SURF and dynamic ROI [J]. Computer Engineering and Design[J], 2013, 34 (3): 998-1003

[2] Lowe D G. Object recognition from local scale invariant features[C]//Proceedings of the Seventh International Conference on Computer Vision (ICCV’99), 1999

[3] Lowe D G. Distinctive image features from scale invariant key points [J]. International Journal of Computer Vision, 2004

[4] Bay H, Ess A, Tuytelaars T, et al. Speeded-Up Robust Features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359
[5] Bauer J, Sunderhauf N, Protzel P. Comparing several implementations of two recently published feature detectors [C]//Proc of the international conference on intelligent and autonomous systems. Toulouse, France: [s.n.], 2007: 348-350
[6] MinHao Lu, Wenjia Qiu, Xiaoshu Dong. Real-time video stitching algorithm based on feature point selection [J]. Command Information System and Technology[J],2018,9(2):85-88
[7] Yuyan Zhu. A research of camera calibration[D]. Kunming: Kunming University of Science and Technology,2010
[8] Zhang Z Y. A Flexible new technique for camera calibration[J]. Transactions on Pattern Analysis and Machine Intelligence,2000,22(11):1330-1334
[9] Xiaozhu Xie. Real-time video stitching based on platform control[D]. Xiamen: Xiamen University, 2009