An Improved Gesture Tracking Algorithm Based On Depth Image Information

YANG Quan

Software Institute, Xi'an University of Arts and Science, Xi'an, 710065, China
yangquan1110@yeah.net

Abstract - An improved Depth Image CamShift (DI_CamShift) algorithm is proposed to realize the accurate tracking of gestures in the sign language video. First, it used Kinect to obtain the depth image information of sign language gestures. Then it adjusted the search window by calculated spindle direction angle and mass center position of the depth images. Finally, the calculation of minimum depth information value in the search window was used to determine the target gesture area. Experiments results show that the algorithm has good robustness and can effectively track the sign language gestures.

Index Terms - Depth image, Kinect, DI_CamShift, sign language gesture tracking.

I. Introduction

Gesture tracking in the video is the key premise relative to other studies, such as sign language recognition, gesture control etc., especially under the condition of complicated background and unqualified circumstance. In complex cases, there may not appear or have more than one object in video, so the accurate detection and tracking can facilitate further purpose. Because color image contains more information and characteristics, so the tracking and segmentation methods based on skin color are widely used. Although the skin color can be easily used to distinguish the hand from other things, it may disturb by objects which have the similar color with skin and couldn't determine which is the target when many hands showing in the video at the same time. As depth image contains the distance information between camera and hands, it can define the forefront hand as the objective hand to solve this problem. While the values of skin pixel color in motion tracking area play a dominant role, if the depth image information can be combined with it, the skin color based gesture tracking will be more accurate.

Kinect, short for the Kinect for Xbox360, is a motion sensing input device by Microsoft for the Xbox 360 video game console and Windows PCs. Based around a webcam-style add-on peripheral for the Xbox 360 console, it enables users to control and interact with the Xbox 360 without the need to touch a game controller. The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot and is designed to be positioned lengthwise above or below the video display. The device features an "RGB camera, depth sensor and multi-array microphone running proprietary software". The depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. So Kinect is a 3D multifunction camera and can get color images and 3D depth information at the same time.

Using Kinect as video acquisition device in the research, it can get the depth image information of gesture corresponding with color gesture video. This paper improved classic CamShift algorithm by uses the Kinect depth image information, and verified the accuracy and robustness of the new algorithm through gesture tracking experiment.

II. CamShift

CamShift (Continuously Adaptive Mean Shift) algorithm is a non-parameter iteration algorithm searching the probability distribution with the core of MeanShift algorithm and based on objective color features. Applied MeanShift in the continuous image sequence is the basic ideas of CamShift, and its tracking of moving object in video through the following ways: (1) Detect probability distribution image by MeanShift. (2) Calculate initial window of the next video frame from the result of previous frame. (3) Iterative procedure of described above. CamShift algorithm can implement adaptive adjustment in view of the object in the video size, thus greatly improve the tracking performance. It makes full use of advantages of MeanShift algorithm, simple and easy to calculate, and realized the adaptive window size control without any increase in computational complexity at the same time. After MeanShift iteration is completed, it can adjust the window size[1, 2, 3] .

Window adjustment principle is as follows: for a gray scale continuously image, if it is projected to 3D plane, x-axis and y-axis corresponding to the image dimensions respectively, z-axis scope is [0, 255] and its numerical value is the gray value of pixel corresponding to the point. Then a gray scale image size can be defined as:

$$V = \iint I(x, y)\,dx\,dy \tag{1}$$

Area S of the image can be obtained from using gray volume divided by the average grey value of image:

$$S = \frac{V}{\bar{I}} \tag{2}$$

where $\bar{I}$ is the average grey value of image area. Because the image is a discrete one, so its volume is:

$$V = \sum\sum I(x, y) \tag{3}$$

In order to keep the image size, it uses the maximum value of pixels (255) to replace the average grey value to make the tracking window as small as possible that can prevent irrespective object come into the window. But that value

cannot be too small, otherwise it may easy to cause the tracking window size is too small to make the algorithm converge to local maximum. Therefore, the window width should be set as:

$$s = 2\sqrt{\frac{M_{00}}{256}} \qquad (4)$$

Square root is for consistent to the length of dimension.

Steps of the algorithm are as follows: give an image and the target histogram, window size h, precision known as ε.

(1) Set the initial search point as the initial target position, for a given image and the target histogram, carry on global reverse projection.

(2) Do the iterative computation by MeanShift algorithm, after convergence, return the zeroth moment M_{00}.

(3) Use (4) to calculate and update the window size according to the formula.

(4) Take the new size of window as the initial window of next frame and set the center of iterative window as the initial position of target object, do (2), if it is the end of video, then return.

III. DI_CamShift algorithm based on depth image information

When tracking the hand, the global operation of CamShift algorithm in the process of reverse projection increases unnecessary computation burden, thus may reduce the tracking performance. The main reason is that brightness of background region is not strong, so the reverse projection will produce noise. When converse from RGB space to HSV space, the pixel with lower brightness and saturation shows the low stability, and make some irrelevant points to be the target area. It's also possible to include irrelevant objects in the tracking window while tracking model is based on skin color. Some further research based on single hand only detect and tracking the right hand as the target object[4, 5, 6], but when other moving object appearing in the video under the complex background, CamShift algorithm may have false tracking problem, and the misjudgment can be very obvious especially in the situation of other hand moving simultaneously. As shown in figure 1, the CamShift algorithm produces tracking mistakes when another signer appearing in the video and playing her hand with some gestures which misguide the tracking area including non-goals.

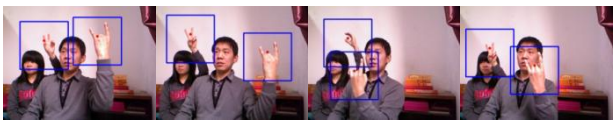


Fig.1 CamShift tracking the interference gesture

As shown in figure 2, the performance of tracking effect of manual alphabet T by CamShift shows it has losing target hand in some frames.

frame20 frame435 frame550 frame725

Fig.2 CamShift lost tracking gesture

Because the depth image took by Kinect contains information related to distance data between Kinect and its scene object's surface, therefore it is more intuitive to show 3D features of the object's surface without Interferences of colors, brightness of shadow problems compared with color image. According to the definition of depth image, it was characterized as follows:

(1) Color-blind, depth image is different from color image in which it barely be affected by light, shadow and changes of surrounding environment.

(2) The change' s direction of the depth image grey value is in accordance with the Z direction of view, that means 3D space can be reconstructed within the feasible region by the depth image. It can also be used to solve the problem of shade or overlap on the basis of properties belonging to the depth image. If two objects are blocked, the stratified condition of grey value generated by their different distance from the camera can be used to distinguish them. As long as set a threshold value to separate two objects with the relationship of front and behind, occlusion problem can be resolved contrast with optical image.

In consideration of insufficient transformation and tracking results given by the CamShift algorithm in color space, this paper uses an improved CamShift algorithm based on depth image information, which named Depth Image CamShift (DI_CamShift) algorithm.

D(x, y) is the depth image, and its (p + q) order 2D origin moment is defined as:

$$M_{pq} = \sum_x \sum_y x^p y^q D(x,y) \quad p,q = 0,1,2,\ldots\ldots \qquad (5)$$

where D(x, y) indicates the depth value of pixel in the location of (x, y).

μ_{pq} is the (p + q) order central moment of D(x, y). It's defined as:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q D(x,y) \qquad (6)$$

Its second central moment can be used as the spindle of sign language gesture in the frame, and the spindle direction is determined by the directions of the maximum and minimum second moment, that are major axis and minor axis. According to the theory of moment, the angle of spindle direction θ is calculated by:

$$\theta = \frac{1}{2}\tan^{-1}\frac{2\mu_{11}}{\mu_{20}-\mu_{02}} \qquad (7)$$

In formula (7), θ is the angle between spindle and coordinate axis, its scope ranged in $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ which shown in the table below.

TABLE I Angle between spindle of sign gesture and axis

μ_{11}	$\mu_{20}-\mu_{02}$	θ
0	-	0
+	-	$-\pi/4<\theta<0$
+	0	0
+	+	$0<\theta<\pi/4$
0	0	0
-	+	$-\pi/4<\theta<0$
-	0	0
-	-	$0<\theta<\pi/4$

If θ is the spindle direction of sign language gesture S, then

$$S^2(\theta)=\frac{1}{n}[(S_1(\theta)-m)^2+(S_2(\theta)-m)^2+\cdots+(S_n(\theta)-m)^2] \quad (8)$$

where $m=\frac{1}{n}[S_1(\theta)+S_2(\theta)+\cdots+S_n(\theta)]$, $S_i(1,2,3,\ldots,n)$ is the spindle direction of object extract from the average frames with same sign language gesture.

Specific steps for DI_CamShift are:

(1) Set the entire depth image as search area.

(2) Use frame difference method to detect the area of moving hand and initialize the search window, locate its size and position.

(3) Calculate the probability distribution of depth histogram in the Search Window area.

(4) Calculate θ_1 and θ_2 separately, they are the major axis and minor axis directions of gesture in the depth image.

(5) Apply MeanShift algorithm to calculate the mass center of depth gesture image in the window, adjust the size of Search Window according to position of the mass center and the spindle direction θ_1 and θ_2.

(6) For the next sign language video frame, use the mass center and size of the Search Window generated by (3) and jump to (3) to continue.

(7) If multiple moving targets are detected, the real sign gesture is Hand Gesture = Min {$M_{00}(Obj_1)$, $M_{00}(Obj_2)$, …, $M_{00}(Obj_n)$}. The closer to camera, the greater depth value can be obtained by object. As the sign language gesture in front of a signer's body is considered to be the target hand and the closest object to Kinect camera, Search Window with minimum value of zero order moment contains the minimum sum of depth value, which comes from the pixels of gesture. Thus the window can be identified as the top target area.

Once confirmed the tracking window in depth video, it will be drawn to the corresponding color video in same location simultaneously for tracking.

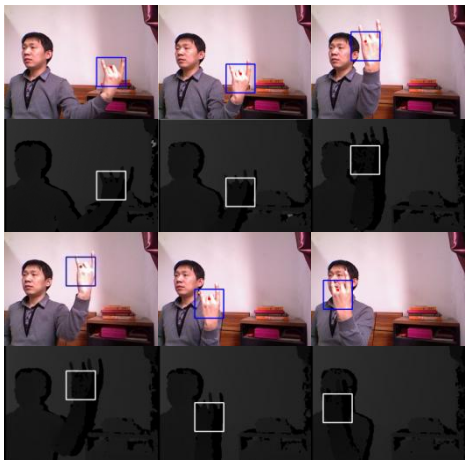
Under the same scene, DI_CamShift algorithm has better performance of tracking, not only avoid lost tracking but also correct the inaccurate tracking of other similar color area farther away from the camera.

IV. Sing language gesture tracking experiments based on DI_CamShift

The experiments set various of interference factors in the same scene to compare the tracking effects between DI_CamShift and CamShift algorithm.

As shown in figure 3 (a), when using the CamShift, target may be lost sometimes. While (b) shows the tracking effect by DI_CamShift in the same situation (color images are frames from the Kinect color video, the corresponding depth image below are frames from the simultaneous depth video).

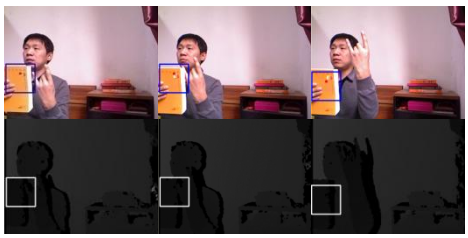
(a) Camshift lost tracking gesture



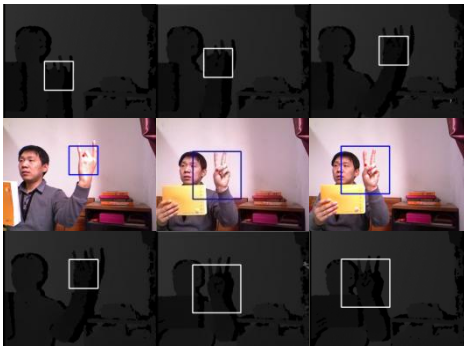
(b) DI_Camshift robust tracking

Fig.3 Compare of DI_CamShift and CamShift

In figure 4 (a), a book whose cover is yellow disturb the tracking because the color is similar to skin. When the hand closing to the yellow cover, Camshift has misjudging the cover to be the target hand and moving it's tracking window on the book. After the hand moving away from the book, it still stayed in place and keeping the false tracking. Affected by light and shooting, the book cover changed its color from canary yellow to bold yellow, but the DI_CamShift shows accurate tracking in (b) while the hand arriving closer and farther apart to the book.



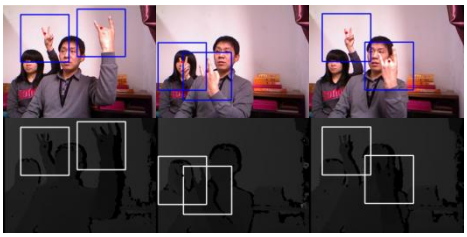
(a) Misjudgement of Camshift



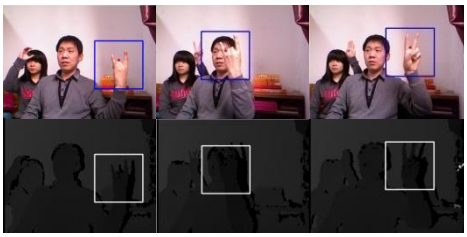
(b) Anti-interference tracking of DI_Camshift

Fig.4 Compare of DI_CamShift and CamShift under interference

At the time of tracking, sign gesture is considered to be the forefront part of human body, even there are more moving hands in video, only the closest to camera is the target for tracking. Camshift shows its tracking in figure 5 (a) when there are two moving gestures. Although located in the different distances and the posterior hand is an interference factor, the algorithm cannot distinguish and treat both of them as the tracking objects. (b) is a correct tracking by DI_Camshift which recognize the target by depth information.



(a) Camshift can't distinguish multiple moving hands



(b) DI_Camshift can recognize and tracking the real object

Fig.5 Compare of identification ability between two algorithms

DI_Camshift presents the exact tracking under the circumstance of indoor and darker in figure 6. The last group of images in it shows the effective tracking in multi-hands video frame.

As experiments results showed, DI_CamShift algorithm uses the depth information of gestures returned by Kinect to reduce a lot of calculations in transform from RGB to HSV space and process of reverse projection. It also avoid misjudgment situation caused by the pixel mapping error, and achieved accurate tracking effect of gestures under interference information and low light conditions.

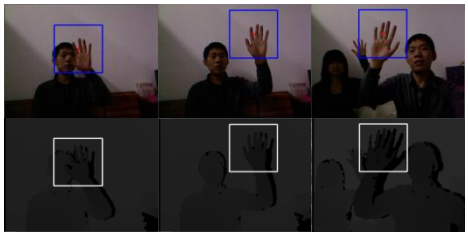


Fig.6 Tracking effect of DI_Camshift under darker indoor circumstance

V. Discussion and Future Work

An improved CamShift algorithm based on depth image information from Kinect sign language video shown the robust and effective gesture tracking effect in this paper. The new algorithm adjusts the size of Search Window by calculating the spindle direction and mass center of sign gestures in depth images. It can make the steady gesture tracking continuously and improved disadvantages of CamShift in lost tracking target and misjudgment with similar color. Depends on the limitation of distance value between Kinect and gestures, the correct hand can be isolated from the depth image information, which solved problem of object recognition for the original algorithm in case of multi-hands tracking. Due to the data calculation of different Search Windows in depth image, the progress of gesture tracking is freedom from interference of color, light and shadow. Comparative experiments show that the algorithm has a strong adaptive ability for varieties of scene, be able to adjust search scope timely, improved speed of tracking and achieved high efficiency, robustness and real-time tracking.

These improvements are the first step for the further research, if the search window could be shaped closer to contour of hand, the tracking effect and gesture segmentation for next step will be more precise. The future work will focus on the search window calculation and accurate gesture segmentation.

Acknowledgment

This work is supported by Xi'an University of Arts and Science, young and Middle-aged professional scientific research project (No. 90158).

References

[1] P.Dreuw, J.Forster and H.Ney. Tracking Benchmark Databases for Video-Based Sign Language Recognition. In ECCV International Workshop on Sign, Gesture and Activity(SGA). Crete, Greece, September 2010: 335-341

[2] Zhaowen Wang, Xiaokang Yang, Yi Xu, et a1. CamShift guided particle filter for visual tracking. Pattern Recognition Letters. 30(2009): 407-413

[3] J.Kovacevic, S.Juric-Kavelj, I.Petrovic, et a1. An Improved CamShift Algorithm Using Stereo Vision For Object Tracking. MIPRO 2011. May 23-27, 2011, Opatija, Croatia: 707-710

[4] Juan Pablo Wachs, Mathias Kolsch, Helman Stern, et al.Vision-Based Hand-Gesture Applications. Communications of the ACM, February 2011, Vol. 54 No.2

[5] RAHEJA J.L, CHAUDHARV A, SIGNAL K. Tracking of Fingertips and Centres of Palm using Kinect[A]. CIMSiM, Langkawi, IEEE, 2011:248-252

[6] OpenKinect Organization. Imaging Information for Kinect[EB/OL]. http://openkinect.org/wiki/Imaging_Information, 2013.