I-061

Motion Detection Based On Active Stereo Camera Array

Yingdi Xie† Jun Ohya†

1. Introduction

Motion detection based on active camera, as one of the most popular research topics among robot vision, has been actively researched since decades ago. Generally, if camera's ego-motion can not be obtained from mechanical measurement, motion in the scene is detected based on the assumption that static objects occupy major area of the scene. Based on this assumption, a method that aims at relaxing a trade-off between detection speed and accuracy is proposed in this paper. And this system could scale up its performance by simply adding up more cameras. We assume the camera moves only in translational direction. In order to simplify the problem, a global shutter camera array is used.

In the next section, the proposed system is analyzed. In Section 3, the experiment result and discussion is described. Conclusion and future work are given in Section 4.

2. The Processing System

In this paper, motion detection is carried out based on a 3D (three dimensional) view of the scene, in order to obtain a more precise result. Thus, it is necessary to get depth information for feature points in the scene. From this motivation, every two cameras in a camera array are combined so as to be a stereo pair and controlled to acquire image at the same time. The whole processing system composes of three modules in general:

(1) Image Acquiring Module — camera array. In this camera array, each camera is positioned with a known distance B_x in the horizontal direction and B_y in the vertical direction, where the values B_x *and* B_y denote the baselines in the horizontal and vertical directions, respectively. Fig.1 Note that the "horizontal direction" is parallel to the x axis of the camera's focal plane. Each camera's coordinate axes are aligned so that each baseline is parallel to the camera's x coordinate axis. For a camera array which composes of $2m \times n$ cameras, in order to acquire image sequence with a desired frame rate, it is controlled to acquire image with m_t stereo camera pairs for each time interval t. Here, m_t is decided by the Image Acquiring Control Module. Assume each camera could obtain image at a maximum frame rate of 30fps (frame-per-second), thus $t = m_t/(m \times n \times 30)$, and image sequence at the frame rate of $(m \times n \times 30)/m_t$ fps could be obtained by this camera array.

(2) Processing Module. This system is designed for real-time processing. In order to achieve this goal, each stereo pair camera is assumed to have one micro processor to calculate feature extraction and stereo matching. A central processor carries out the other processing steps. Details of this module are described in 2.1.

(3) Image Acquiring Control Module. This module controls the number of stereo pair number m_t: i.e. the selection of camera pairs, as well as the timing for image acquisition for the next frame. The trade-off between detection speed and accuracy is optimized in this step. Details are described in 2.2.

The relation between these three modules is demonstrated as Fig.2.

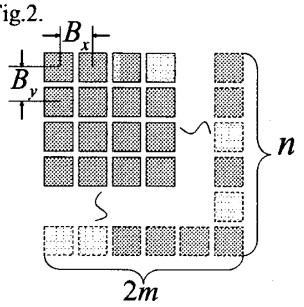


Fig.1 Camera Array

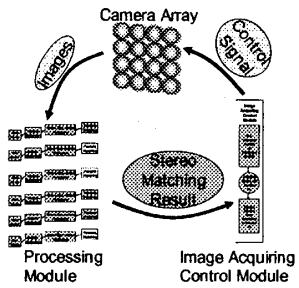


Fig2. System structure

2.1 Processing Module

Motion detection is carried out by comparing the translated image which is generated from the previous frame according to the camera's ego-motion. The detections of camera's ego-motion and depth information of the scene are two important processing steps. Depth information can be calculated from stereo matching. In this paper, we use the method proposed by Okutomi et al [1]. Although it has already been proposed for ten more years, the evaluation function in [1] is the key point for optimizing the trade-off between accuracy and detection speed, as described in 2.2.

The other precondition for motion detection is the camera's ego-motion, which is processed after stereo matching. Generally, the ego-motion can be expected from the flow vectors in a scene does not contain a moving object. However, in many cases there may be moving objects exist in the scene. To solve this problem, quasi-ego-motion, in stead of ego-motion, is used. The so called quasi-ego-motion is estimated from an ego-motion based on the assumption that moving objects in the scene can only move in depth direction. And it is computed by the following two steps.

(1) Because the depth information of the current frame and the previous frame have been obtained from stereo matching, the ego-motion in z direction $\triangle Z_{t,e}$ is firstly estimated by accounting the flow vector in z direction $z_{i,t}$ and $z_{i,t-1}$ (i indicates the point serial number in one image) for all the points to be processed, from which a probability function $P(\triangle Z_t)$ could be obtained. Given that our method is based on the assumption that static objects occupy major area of the scene, $\triangle Z_{t,e}$ is decided as the expectation value of $P(z_{i,t})$'s major part.

(2) The quasi-ego-motion in x and y direction ($\triangle X_t$, $\triangle Y_t$) is then computed according to the following equations.

$$\begin{aligned}\Delta X_t &= E[P(\Delta X_{i,t})] \\ &= \sum_{i=1}^{N} \Delta X_{t,i} P(\Delta X_{i,t}) \\ &= \sum_{i=1}^{N} \frac{1}{F}(z_{i,t-1}x_{i,t-1} - z_{i,t}x_{i,t}) P(\Delta X_{i,t}) \end{aligned} \quad (1)$$

† Waseda University GITS Ohya Lab

$$\Delta Y_t = \sum_{i=1}^{N} \frac{1}{F}(z_{i,t-1}y_{i,t-1} - z_{i,t}y_{i,t})P(\Delta Y_{i,t}) \qquad (2)$$

where, $(x_{i,t},y_{i,t},z_{i,t})$ and $(x_{i,t-1},y_{i,t-1},z_{i,t-1})$ indicates the coordinates of the ith point which lies in the vicinity of the peak position of $P(z_{i,t})$ for the current frame and the previous frame respectively. $\triangle X_{i,t}$, $\triangle Y_{i,t}$ are the elements of the quasi-ego-motion vector in the x and y directions respectively, which are decided by point i. N is the number of points described above, F is the focal length. $P(\triangle X_{i,t})$ and $P(\triangle Y_{i,t})$ are probability functions for quasi-ego-motion in x and y direction.

After quasi-ego-motion computation, motion is detected if point i satisfies any one of Eqs. (3) ~ (5):

$$|\Delta X_{i,t} - E[P(\Delta X_{i,t})]| > TH_x \qquad (3)$$

$$|\Delta Y_{i,t} - E[P(\Delta Y_{i,t})]| > TH_y \qquad (4)$$

$$|\Delta Z_{i,t} - E[P(\Delta Z_{i,t})]| > TH_z \qquad (5)$$

where, (TH_x,TH_y,TH_z) are pre-determined thresholds for detecting the motions in x, y and z directions respectively. The camera's ego-motion can now be computed by excluding the moving points, if necessary.

2.2 Image Acquiring Control Module

This module carries out the control of m_t. It is well known that while applying a template matching method the stereo matching, same pattern in one local area will result in ambiguity of matching result. The motivation for adjusting m_t is to eliminate the ambiguity with critical camera pair number. Here we define that the maximum number of same pattern in each local area as "Scene's Complexity", by which m_t is adjusted for the next frame. In the experiment given in [1], it is demonstrated that the more images are acquired by the cameras with different baselines, the more easily the ambiguity can be eliminated. Thus, this causes a trade-off between accuracy and detection speed, because although acquiring more images, in other words increasing m_t can eliminate ambiguity, it will result in decreasing frame rate of image sequence. This trade-off is optimized by minimizing the minimum number of evaluation function in Eq.(6), according to [1].

$$\begin{aligned} e_{\varsigma(12\ldots n)}(x,\varsigma) &= \sum_{i=1}^{n} e_{\varsigma(i)}(x,\varsigma) \\ &= \sum_{i=1}^{n}\sum_{j\in W}(f_0(x+j) \\ &\quad - f_i(x+B_iF\varsigma+j))^2 \end{aligned} \qquad (6)$$

where, $f_i(x)$ is the function of the ith image among the multiple baseline images. Because multiple baseline cameras are set in a straight line, the stereo matching problem can be solved by searching pixels in one dimension (searching is carried out in x direction here), W is the size of searching window. B_i indicates the ith baseline distance, ς is the inverse distance defined by 1/Z, and, this model is assumed based on independent Gaussian white noise model.

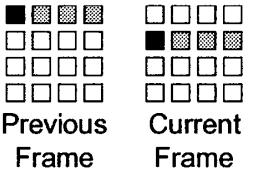
Ambiguity occurs when the number of local minima of Eq.(6) exceeds one. To avoid this, one more camera pair ($m_{t+1} = m_t + 1$) is controlled to acquire image for the next frame. When there is no ambiguity for all the points being matched, the system carries out computation of stereo matching with fewer baselines, adjust m_t to a number m, so that with $2 \times m$ images the ambiguity can be critically eliminated. The adaptation steps are demonstrated as the following pseudo-code:

ADAPTATION()

If (6)'S MINIMUN NUM > 1
  Then $m_{t+1} = m_t + 1$
    QUIT
Else $m_t = m_t - 1$
  If (6)'S MINIMUN NUM < 1
    Then ADAPTATION()

2.3 Expectation of the image of different baseline

By using this camera array, different cameras are used at each frame. As demonstrated in Fig.3, black block is the main image for each frame, denoted as f_0 in Eq.(6). In this example, every four cameras, blocks in gray or black, are controlled to acquire image for each frame. It is obvious that the two consecutive main images are not necessarily acquired by cameras at the same position in the array, while this proposal requires that consecutive main frame acquire by the same camera in the array. Thus, another step for transforming the previous frame to the image which is assumed to be acquired by the camera acquires the current frame's main image. Here the expected image is noted as I_{expect}. This step is carried out after applying stereo matching to the previous frame according to the equations (7) and (8):



$$\hat{x}_{i,t-1} = x_{i,t-1} - \frac{BF}{z_{i,t-1}} \qquad (7)$$

$$\hat{y}_{i,t-1} = y_{i,t-1} - \frac{BF}{z_{i,t-1}} \qquad (8)$$

Fig.3 Image Expectation

where, B is the baseline between the two main images. The left side of Eqs(7) and (8) is the coordinate of ith corresponding point in I_{expect} at time t.

3. Experimental Result and Discussion

We have done an experiment on the step of "Expectation of the image of different baseline". By setting 4 cameras in a straight line, images are acquired in the same time. First, distances of corresponding feature points within the image acquired from the left two cameras are measured. And by Eqs. (7) and (8), these feature points' corresponding positions within images acquired by the other two cameras are computed. Then, by comparing these computed positions with the real positions, The validity of Eqs.(7) and (8) is verified.

4. Conclusion

In this paper, a new approach of adaptive motion detection based on active camera array is presented. With this approach a good trade-off between detection speed and precision is achieved.

Reference [1] M.Okutomi and T.Kanade, "A Multiple-Baseline Stereo" IEEE Trans. on Pattern Analysis and Machine Intelligence, VOL.15, NO.4, April 1993.