Proc. VIIth Digital Image Computing: Techniques and Applications, Sun C., Talbot H., Ourselin S. and Adriaansen T. (Eds.), 10-12 Dec. 2003, Sydney

Recognition of Motion in Depth by a Fixed Camera

Huynh Quang Huy Viet, Michio Miwa[2], Hidenori Maruta[3], and Makoto Sato

[1] Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503, Japan
[2] Development Center, Corporate Engineering Division, Panasonic System Solutions Company, Matsushita Electric Industrial Co., Ltd, 4-3-1, Tsunashima-higashi, Kohoku-ku, Yokohama, 223-8639, Japan
[3] Information Science Center, Nagasaki University, 1-14 Bunkyo, Nagasaki 852-8521, Japan

Abstract. The research on perception of motion has important applications for surveillance and autonomous robot navigation in dynamic environments. The issue of estimation of motion in depth is the crucial point of the problem of recognition 3D motion. In this paper, we propose a fixed monocular camera with focus changed cyclically to recognize the absolute translational motion in depth of a rigid object.

1 Introduction

By a fixed monocular camera, the motion of an object in a scene can be computed by using the so-called eight-point algorithm or linear algorithm [3], [8] after matching out at least eight point correspondences in the two images of that scene captured at varied points in time. Let t_1 and t_2 ($t_1 < t_2$) be two points in time corresponding to two captured images of a moving object. Let (x_1, y_1, z_1) be the coordinates of one point in the object at time t_1 and (x_2, y_2, z_2) be the coordinates of that point at time t_2. The motion of the object determined by using the eight-point algorithm or linear algorithm is the relative displacement $x_2 - x_1$, $y_2 - y_1$, $z_2 - z_1$; the coordinates in depth such as z_1, z_2 of the object at time t_1 and t_2 cannot be determined independently. This means that, to estimate the absolute motion in depth of an object, the initial position in depth of the moving object needs to be known previously.

The issue of estimation of motion in depth is the crucial point of the problem of recognition 3D motion utilizing a fixed monocular camera. In this paper, we propose a fixed monocular camera with the focus (the sensor plane position) changed cyclically, to recognize the absolute translational motion in depth of a rigid object. The cycle of the focus-change is an interval of time during which the value of focus changed from near value to far value and returning to initial value. The images captured in a half cycles of the focus-change form a multi-focus image sequence. The motion in depth or the focus-change of the camera causes defocused blur. We developed an operator in order to detect the in-focus frame

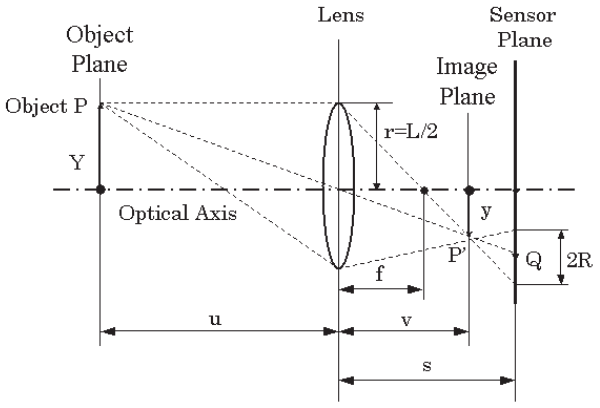


Fig. 1. Formation of focused and blurred image

in a multi-focus image sequence of a moving object. The in-focus frame gives an instantaneous depth position in the motion of the object at the time that the frame was captured in each half cycle. The trajectory of motion in depth of the object is computed from the in-focus frames extracted from the multi-focus image sequences captured through all cycles. In the proposed system, the condition of previously knowing the initial position in depth of the moving object is not necessary.

This paper is organized as follows. Section 2 reviews the camera model and the blur model, which are the basis of presentation in the subsequent sections. Section 3 describes the principle of the proposed system. Section 4 presents the experiment result and the discussion. Finally, section 5 is devoted for the conclusions.

2 Preliminaries

2.1 Camera Model

A simple camera consists of a thin lens and a sensor plane. Fig.1 shows the image formation of a thin lens model. Light rays from object point P, which are intercepted by lens, are refracted to converge at the point P' on the image plane. The relationship among the object distance u, the focal length of the lens f, and the image distance v is given by the lens formula:

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v} \tag{1}$$

In Fig.1, $L(= 2r)$ is the aperture diameter and s is the distance between the sensor plane and lens. The distance s, the focal length f, and aperture diameter L, are referred as the camera parameters.

Each point on the object plane, which is projected onto a single point on the image plane, causes a focused image on the image plane. When the sensor plane coincides with the image plane ($s = v$), the focused image gives a clear image or in-focus image of the object on the sensor plane.

In the image formed by a camera on the sensor plane, when sensor plane is fixed at a distance $s(s > f)$ from lens, only objects at a certain distance are in-focus; other objects are defocused in the varying degrees depending on their distance. Inversely, when the object is being at a distance u $(u > f)$ from lens, there is only one position behind the lens at which the sensor plane has the in-focus image of the object.

2.2 Blur Model

In Fig.1, if the object point P is not in-focus (defocused), then it gives rise to a blurred image Q on the sensor plane. The blurred image Q has the same shape as the lens aperture but is scaled by a factor. Because the aperture is circular, the blurred image Q is also a circle with uniform brightness inside the circle and zero outside, called blur circle. The radius R of the blur circle is given by [7]:

$$R = rs(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}) \tag{2}$$

where r is the radius of the lens.

Let the light energy incident on the lens from the point P during one exposure period be one unit. The blurred image Q of the object point P is the response of the camera to a unit point, hence it is the point spread function (PSF) of the camera, denoted by $h(x, y)$. The blurred or defocused image $g_d(x, y)$ formed on the sensor plane is the result of convoluting the focused image $g_f(x, y)$ with the point spread function $h(x, y)$ [7]:

$$g_d(x, y) = g_f(x, y) * h(x, y) \tag{3}$$

3 Fixed Monocular Camera System to Recognize the Absolute Translational Motion in Depth

3.1 Principle

The Fig.2 outlines the operating principle of the fixed monocular camera system to recognize motion in depth of an object. The camera has the sensor plane

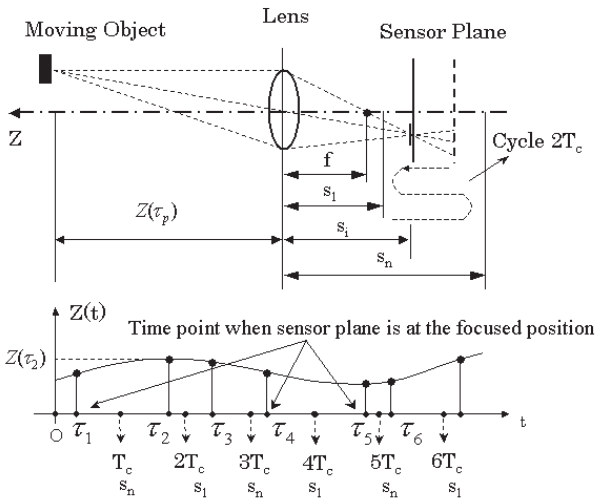


Fig. 2. Principle of the Proposed Camera

position (also called the focus) changed cyclically. Suppose that $s_1, s_2, ..s_i., s_n$ are the values of the distance s to the lens that the sensor plane can have. The cycle $2T_c$ of the focus-change is an interval of time during which the sensor plane of the camera changing its position versus the lens from s_1 to s_n and returning to s_1 or vice versa. The images of a moving object taken at the sensor plane in each half cycles T_c form a multi-focus image sequence (MFIS), which includes a in-focus image frame, called, for short, in-focus frame. In the in-focus frame taken at time point τ_p, the corresponding focus value s_i gives the instantaneous position in depth of the moving object in the following expression inferred from the lens formula (1):

$$Z(s_i) = \frac{s_i f}{s_i - f} \tag{4}$$

To automatically extract the in-focus frame in the sequence of the multi-focus images of a moving object, we developed an operator called the in-focus frame tracking operator, which the Sect.3.2 will present in detail.

Suppose that T is the time interval that the motion of object were observed by the camera and T/T_c is an integer N, then the number of the MFISs are N. The N in-focus frames extracted from N MFISs give the N instantaneous positions in depth $Z(s_i)$ of the moving object at N time points τ_p at which the in-focus frames were taken . The motion in depth of the object can be described by the function $Z(t)$ of position and time inferred from N relation pairs $(Z(s_i), \tau_p)$ through out the time of N half cycles T_c of the focus-change.

3.2 The In-focus Frame Tracking Operator

Depth from focus (DFF) [2], [4], [7] is a well-known method of estimating the 3D surface of a motionless scene from a sequence of images of that scene captured by camera at varied focuses. In DFF, the image is divided into regions and the sharpness of a particular region is computed by utilizing the so-called focus measure. The focus measure is an operator defined on the set of the fixed regions of the same position in the sequence of images that gets maximum at the one having best focused image. In the dynamic scene, due to motion, the image of an object in the scene is moved, magnified or reduced, and hence the division of image into fixed regions of DFF loses its significance and consequently the conventional focus measures don't work properly.

Based on the idea of DFF, we developed an operator for tracking the in-focus frame in a MFIS. For the convenience of presentation, we previously give some definitions:

The multi-focus background image sequence (MFBIS) is a sequence of the images of background scene captured for a monotone sequence of focus values s_i of camera.

Suppose that MFIS and MFBIS are two image sequences captured with the same focus values. The multi-focus subject image sequence (MFSIS) is the subtraction image sequence formed by frame difference of the images in the MFIS with the images of the same focus value in the MFBIS.

Let $g_i(x, y)(i = 1, 2...N)$ be a MFSIS where i is the order of the focus s_i, the in-focus frame tracking operator is an operator that is maximum for the best focused image among the images $g_i(x, y)$ of sequence and gradually decrease as the image blur increases.

The images taken at the sensor plane in a cycle of the focus-change when there is not an object in front of the camera form two MFBISs, in which the order of the focus in a MFBIS is inverse to the other one. When an object is moving in front of the camera, the images taken in a half cycles of the focus-change form a MFIS. This MFIS becomes a MFSIS, after being subtracted the corresponding MFBIS having the same orders of the focus. In the frames of a MFSIS of a moving object, the images of the object not only become defocused and then focused but also become reduced and then magnified. We defined an operator called the in-focus frame tracking operator, whose the region of limit of integration is variable, as shown in the following expression:

$$m(i) = \frac{1}{S(D_i)} \int \int_{D_i} g_i^2(x, y)\ dxdy \tag{5}$$

Here, i is the order of the focus, D_i is a variable region $\{(x, y) : g_i(x, y) \geq Th\}$, Th is a threshold value, and $S(D_i)$ is the area of D_i. The region of limit of integration D_i is varied in the frames of the MFSIS due to the magnification or the reduction of the image of the moving object. Intuitively, the above expression bears the significance of the power density of the portion of the image restricted by threshold value Th, of the object. The power density becomes maximum at

the frame having the best-focused image. The automatic extracting the in-focus frame in a MFSIS is carried out by detecting the frame in MFSIS, at which the in-focus frame tracking operator is being maximum. Since the MFSIS is corresponding to the MFIS in the orders of focus, the in-focus frame in MFSIS is the same order of focus with the in-focus frame in MFIS. On digital image computing, the area of D_i in the in-focus frame of a MFSIS is summation of all pixels which have luminance value greater than or equal to the threshold value Th.

4 Experiment Result and Discussion

4.1 Experiment

In this section, we present one of the experimental results that show the effectiveness of the proposed system.

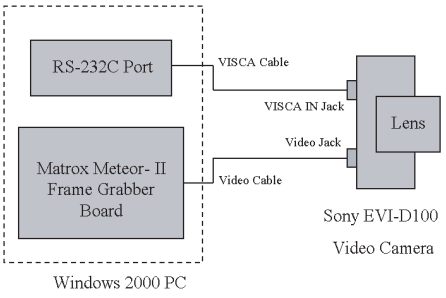


Fig. 3. Experiment System

A block diagram of the equipment system for the experiments is as shown in Fig.3. It consists of a Sony EVI-D100 CCD video camera and a PC with a frame grabber board Matrox Meteor-II installed inside it, under Windows 2000 environment. The EVI-D100 is connected to the PC through RS-232C port and the alteration of the lens focus or of the other parameters was performed via VISCA protocol. The images captured by the frame grabber board from the camera are the grayscale images in the size of 320 x 240 pixels. In every experiment, the diameter of the camera lens aperture was set to maximum value corresponding to f-number F1.8.

The aim of the experiment is to demonstrate the capability of recognition of the motion in depth of the proposed camera through estimating depth distance which a moving object passed through in time of two half cycles of the focus-change. In the experiment, the object was a small toy bus painted in orange

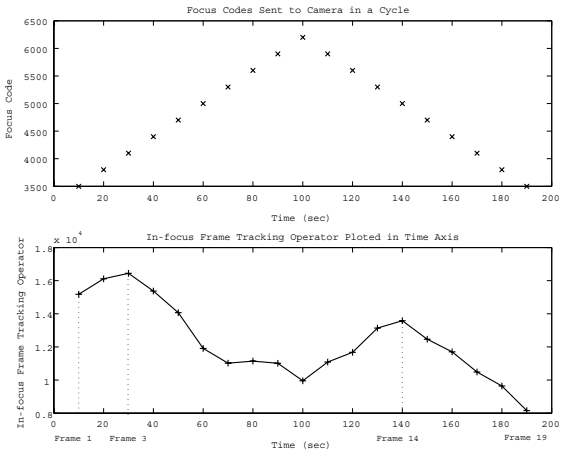


Fig. 4. Focus Code and In-focus Frame Tracking Operator

at below part and white at upper part. The experiment was performed in three stage: the first stage for taking two subsequent MFBISs, of which the focus values increase in one MFBIS and decrease in the other; the second stage for taking two subsequent MFISs of the moving toy bus with the same focus values as in the first stage; the third stage for calculating distance in depth of motion and comparing with the actual distance that the toy bus passed by.

In controlling of the focus of camera from the PC via VISCA, every actual focus value is symbolized by a focus code that takes value within the range of from 1000 to 8400. The relation of the focus values and the focus codes is monotone but not linear. Because the focus code is not a real focus value, the corresponding object distance needs to be calibrated previously. Upper part of the Fig.4 gives the chart of the focus codes in time, sent to camera by the PC for taking two subsequent MFBISs and two subsequent MFISs.

In the first stage of this experiment, background was arranged from the toy houses. The frame rate is a 0.1 frames/sec or 10 secs/frame. The number of frames captured in two subsequent MFBISs are 19 frames during the time 190 seconds of two half cycles as shown in the upper chart of Fig.4. The five frames of the first MFBIS are given in the first line of Fig.5

In the second stage, a toy bus was chosen as an object to estimate motion; the initial position of it is 94 cm in front of the camera. 19 frames of two subsequence MFIS were captured with the same focus values and the same frame rate as performed in the first stage. The frame rate was set to low speed of 0.1 frame/sec or 10 secs/frame in order that within the time 10 seconds for capturing the next frame the toy bus can be moved in 2.5 cm by hand toward the camera orthogonal to the lens surface. The second line of Fig.5 gives the five frames of the first MFIS.

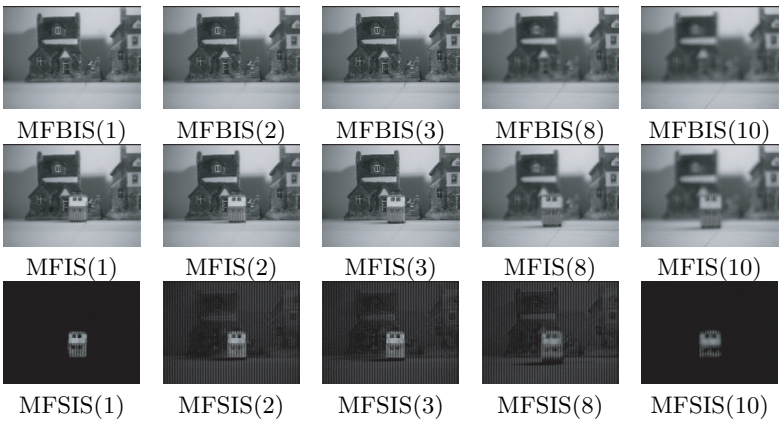


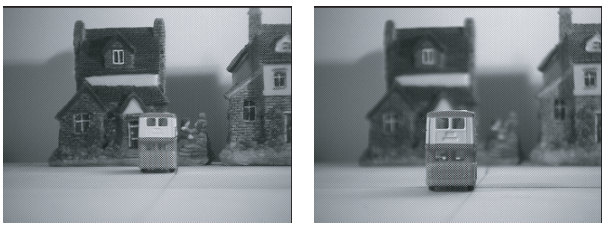
Fig. 5. The MFBIS, MFIS and MFIS of the Moving Object

In the third stage, by subtracting in frame by frame two MFISs with two MFBISs and applying the in-focus frame tracking operator with threshold of $Th = 70$ onto two acquired MFSISs, we obtained the result as shown in the lower part of Fig.4. In this figure, applying the in-focus frame tracking operator onto two subsequent MFSISs gave the result of two charts, which are located subsequently according to the orders in time that the frames were captured. Two local maximums in the charts are corresponding to two in-focus frames in two MFSISs and two in-focus frames of MFISs (see Fig.6). The points in time axis at two maximums give positions in front of the camera, which the toy bus was moved to. Here, position means the distance toward the lens of the camera. They are 89 cm and 61.5 cm corresponding to the time points at the first and the second maximum. The distance that the toy bus actually moved through is 27.5 cm. The focus codes sent to camera from the PC, of the frames corresponding to the first and the second maximum are 4100 and 5000. The calibrated object distance of the focus code 4100 is 87 cm and the its bias is about 3.5 cm; the calibrated object distance together with its bias of the focus code 5000 are 62 cm and 2.5 cm. The estimated distance from two maximums is 25 cm. The estimated distance is different to measured distance by 2.5 cm.

4.2 Discussion

In the above experiment, it is obvious that the estimated motion is absolute translational motion and the condition of previously knowing the initial position in depth of the moving object is not necessary as in conventional methods (cf. [3], [8]). The above experiment that if carried out with several cycles of focus-change can give the trajectory of motion of the toy bus instead of the distance.

In practice, due to aberration, all lenses cannot perfectly converge rays from an object point to form a true image point (i.e. an infinitely small dot with zero



Frame 3 Frame 14

Fig. 6. The In-focus Frames

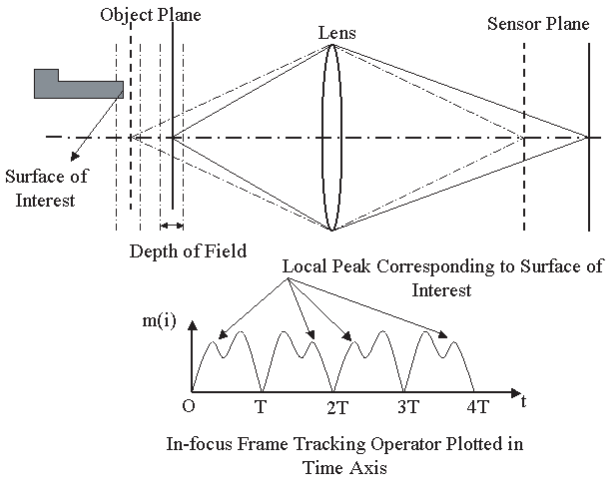


Fig. 7. Complex Surface Object

area). This means that the images are formed from a composite of dots having a certain area. As a result of this, the object at many positions in front of the lens will give the same sharp image. Depth of field (see [6]) is a concept defined as the depth of the region in front of and behind a focused object within which the lens provides the sharp images. It become evident that the value u of the position of the object in the motion calculated from in-focus frame in the MFIS contains a systematic error or bias which is identical to depth of field at that position. Therefore, the shallower the depth of field becomes, the smaller the error will be. In order that systematic error becomes minimized, the following principles inferred from the expression of function of depth of field should be considered in design stage:

1.The error becomes smaller at longer focal lengths.

2.The error becomes smaller at larger apertures.

3.The error becomes smaller at closer object distances.

In the case of the object with complex surface (several surfaces corresponding to multiple in-focus frame) as shown in Fig.7, if utilizing the camera with the high sampling rate and the narrow depth of field for all focuses, the MFIS will have multiple in-focus frames and consequently the in-focus frame tracking operator will have several local maximum peaks. In this case, it is necessary to consider the partial surface of the object that is nearest the lens of the camera as the surface of interest. Due to the fact that depending on the movement direction of the sensor plane, the surface of interest firstly (or lastly) causes the in-focus image in the MFIS, the in-focus frame of the surface of interest can be detected by tracking the frame corresponding to the first (or last) local peak of the in-focus frame tracking operator. In translational motion, the motion of all points belonging to the object take the same trajectory, tracking the surface of interest will give the motion of the object.

In the general case of recognition 3D motion, since the coordinate information of the feature points (e.g. corner, zero crossing, etc) [5] in the focused image of the moving object in the in-focus frame give horizontal and vertical position of object, the conventional methods of matching correspond points can be used for tracking the feature points in the in-focus frames of the subsequent MFSISs. The 3D positions time computed from focus value together coordinate information of the feature point in the in-focus frames through all MFSISs give the 3D motion of the object.

5 Conclusion

This paper has described a fixed monocular camera system to recognize the absolute translational motion in depth of a planar rigid object. The system can totally be extended further to recognize the absolute 3D translational motion of the object with complex surface.

References

1. T. Kaneko, T. Ohmi, N. Ohya, N. Kawahara, and T. Hattori, "A new, compact and quick-response dynamic focusing lens," in *Transducers'97*, 1997, pp. 63–66.
2. E. Krotkov, "Focusing," *Int. J. Comput. Vision*, vol. 1, pp. 223–237, 1987.
3. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, Sept., 198.
4. S.K. Nayar, "Shape from focus system," in *IEEE Conf. Comput. Vision & Pattern Recognition*, 1992, pp. 302–308.
5. S.M. Smith and J.M. Brady, "Susan - a new approach to low-level image processing," *Int. J. Comput. Vision*, vol. 23, pp. 45–78, 1997.
6. W.J. Smith, *Modern Optical Engineering*, McGraw-Hill, New York, 1990.
7. M.Subbarao, T.Choi, and A.Nikzad, "Focusing techniques," *J. Opt. Eng.*, vol. 32, no. 11, pp. 2,824–2836, 1993.
8. X.Zhuang, "A simplification to linear two-view motion algorithms," *Comput. Vision, Graphics & Image Process.*, vol. 46, pp. 175–178, 1989.