NEW DEPTH FROM FOCUS METHOD FOR 3D PTZ CAMERA TARGET TRACKING

Tiago Gaspar, Paulo Oliveira

Instituto Superior Técnico
Institute for Systems and Robotics
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ABSTRACT

A new active depth from focus method is proposed for a generic moving target, based on real time images from a low cost single pan and tilt camera. Resorting to a (sub-)optimal Multiple Model Adaptive Extended Estimator, a nonlinear tracking system is obtained that provides estimates on the target 3D position, velocity, and acceleration, identifying its angular velocity. The proposed methods are robust to broad conditions of operation and to image disturbances. Results from experiments with a real target mounted on a robotic platform validate the proposed methods.

Index Terms— Tracking filters, Nonlinear filters, Active vision, Focusing

1. INTRODUCTION

Depth estimation plays a key role in a wide variety of domains, such as target tracking [2], 3D reconstruction [3], obstacle detection [5], and video surveillance [8]. In 3D image applications a common approach consists in using triangulation methods applied to the data collected by two or more cameras. However, there has been work on estimating depth resorting to a single camera [10], [6]. In addition to the main advantage of requiring just one camera, this technique reduces the impact of the image to image matching problem, as well as the impact of occlusion problems [13]. The ideia is to explore the relation between the depth of a point in the 3D world and the amount of blur that affects its projection into acquired images. This is done by modelling the influence that some of the camera intrinsic parameters have on images acquired with a small depth of field. Based upon this principle, there are three main strategies that have been explored: depth from blur by focusing [15], [12], zooming [1], and irising [6].

In this paper, we are mainly concerned with depth estimation from blur by focusing. Two different techniques based upon this approach can be found in the literature: depth from defocus [12], [6], and depth from focus [10], [11], [15]. This work is based on this later method, since this type of approach does not require a mathematical model for the blurring process of the camera, i.e. the point spread function responsible for the blurring does not need to be modeled.

This work was motivated by previous work on target tracking and positioning [7], where a low cost single pan and tilt camera based indoor positioning and tracking system was proposed. The problem of estimating the depth of a moving target is now tackled directly by proposing a novel method to estimate the target depth without having any information about its dimensions. The blur information present on the target boundary is used to infer depth based upon a depth from focus strategy. As a consequence, a tracking system capable of estimating the 3D position of a target in real time, using a single PTZ camera and without additional information about the target dimensions must be designed.

Partially funded with grant SFRH/BD/46860/2008, from Fundação para a Ciência e a Tecnologia. Authors' emails: {tgaspar, pjcro}@isr.ist.utl.pt.

2. BACKGROUND ON THEORY OF DEFOCUS

The ideia of inferring depth from focus and defocus is based on the concept of depth of field, which is a consequence of the inability of cameras to simultaneously focus planes on the scene at different depths. The depth of field of a camera with a given focus corresponds to the distance between the farthest and the nearest planes on the scene, in relation to the camera, whose points appear in acquired images with a satisfactory definition, according to a certain criterion.

At each instant, a lens can exactly focus points in only one plane, denominated object plane. Considering a thin model for the lens, it is possible to establish a relation between the distance u of this plane to the lens and the distance v between the lens and the image plane at which points appear sharply focused in the image. To complete the relation, the focal length f of the lens must be considered, as expressed in the Gaussian Lens Formula [9]:

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

Considering that the CCD sensor plane is located at a distance $v_0 < v$ from the lens, and using (1) and some trigonometric manipulations, it is possible to write the distance u from the lens to the object plane in the scene as

$$u = \frac{f v_0}{2RF + v_0 - f}, \tag{2}$$

see [6] and [12] for details, where F is the f-number of the lens and R is the effective radius of the point spread function. This expression is valid when $v > v_0$, i.e. when $u < u_0$. An expression similar to this would be easily derived for the case $u > u_0$.

In practical applications, usually all parameters in the right-hand side of equation (2) are known, except for R. Depth from focus methods consist in finding the sensor plane position that minimizes the amount of blur present in image points of interest. This corresponds to finding the camera focus parameter v_0 that leads to $R = 0$, which is solved by optimizing a cost function that depends on the amount of blur present in the image point of interest. Depth can be deduced from (2), resulting in

$$u = \frac{f v}{v - f}. \tag{3}$$

Depth from defocus strategies estimate R by measuring the amount of blur present in the image point of interest, and u follows directly from (2). In this type of methods, it is common to express the defocused image $I_d(x, y)$ formed on the sensor plane as the convolution of the focused image $I_f(x, y)$ with the

point spread function $h(x, y)$ of the lens system, i.e. $I_d(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_f(\alpha, \beta)h(x-\alpha, y-\beta)d\alpha d\beta$. In practice, only $I_d(x, y)$ is known. Thus, $h(x, y)$ must be recovered from $I_d(x, y)$, where the amount of blur can be a function of both the characteristics of the lens and the scene itself. Then, this function is used to infer depth, since R is intrinsically related to depth, according to (2). The solution of such inverse problem, usually ill-posed and difficult to solve, is avoided with the approach proposed in this paper.

3. MINIMUM BLUR FOCUS SETTING ESTIMATION

3.1. Cost function

The estimation of the camera focus setting that minimizes the amount of blur in an image discontinuity requires the definition of a metric that quantifies the sharpness of a transition in an image. Metrics related with high-frequency energy contents in the image, Fourier transform, image gradient, or Laplacian, are detailed in [10].

The cost function proposed in this work is based upon a gradient magnitude maximization strategy for a particular region of the image, thus departing from the classical approach in [14], where all image was considered. The aim of our system is to estimate the depth of a moving target. Thus, the cost function proposed is based on the image gradient magnitude across lines orthogonal to the target boundary. Clearly, it is a function of the camera focus parameter. The problem at hand can be formulated as

$$\min_{f_s} g(f_s),$$

where f_s is the camera focus setting that corresponds to the focus parameter that we want to estimate, and the cost function

$$g(f_s) = \frac{1}{\frac{1}{N_l}\sum_{i=1}^{N_l} \max|\nabla I_{f_s}(x, y)|}, \qquad (x, y) \in l_i, \quad (4)$$

is the inverse of the mean of the image gradient magnitude maximum values across each of the lines orthogonal to the target boundary. Moreover, N_l denotes the number of lines used, ∇ the gradient operator, $I_{f_s}(x, y)$ the intensity of the image acquired with the focus setting f_s at point (x, y), and l_i the i^{th} line. The formulation of this problem as the minimization of $g(f_s)$ is based on the model that will be proposed for this function in next section.

The implementation of this method requires the estimation of the image intensity gradient $\nabla I_{f_s}(x, y)$ resorting to any of the existing operators, e.g. Sobel operator. The target boundary is assumed as known, based on the results obtained from the use of active contours, see [7] and [4] for details.

3.2. Minimization of the cost function

The minimization of the cost function proposed in (4) is difficult. The data available is scarce and to get new information, the acquisition of a new image is required. The problem is even more difficult as we want to estimate parameters related with the depth of a moving target.

The model for the cost function depends on the imaging system of the camera used. Experimental results for the 215 PTZ camera from AXIS are depicted in Fig. 1, thus validating a parabolic model for the cost function when using this camera, i.e.

$$g(f_s) = a(f_s - f_s^*)^2 + b, \quad (5)$$

where a, b, and f_s^* are parameters to be estimated. In particular, f_s^* is the camera focus setting that minimizes the cost function. This expression can also be written as

$$g(f_s) = a' f_s^2 + b' f_s + c', \quad (6)$$

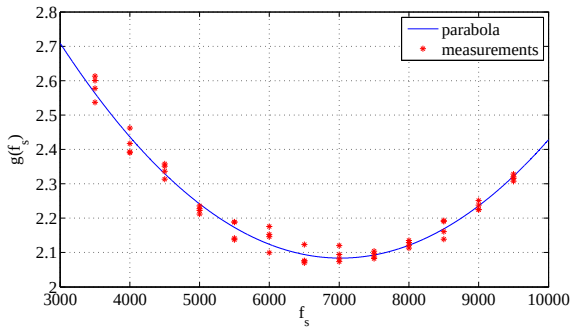

Fig. 1. Cost function (4) when the camera focal length is 29 mm and the target is 3 m away from the lens.

where $a' = a$, $b' = -2af_s^*$, and $c' = af_s^{*2} + b$. The linear dependency on the parameters simplifies significantly the fitting problem.

In this work, depth of a moving target must be estimated. Thus, the value that minimizes the cost function changes over time. The proposed method successively estimates the varying cost function parameters in real time.

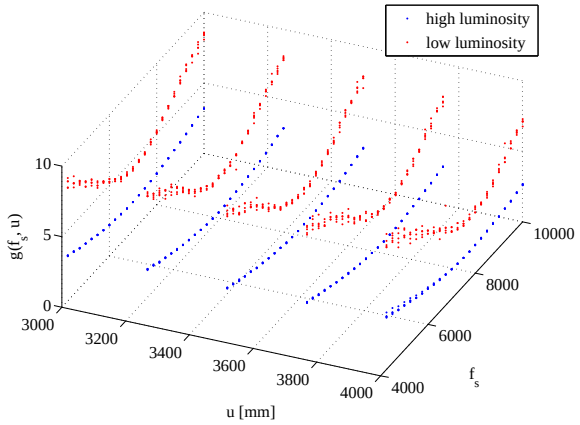

Fig. 2. Luminosity influence on the cost function, for several target depths.

Consider that at instant k we measure $g(f_{s_k})$ corrupted by additive white Gaussian noise. Stacking N of these measurements, a fitting problem can be formulated as

$$\min_{\mathbf{x}} ||\mathbf{Ax} - \mathbf{b}||_2, \quad (7)$$

where $\mathbf{A}$ is a matrix with N rows of the form $[f_{s_k}^2 \;\; f_{s_k} \;\; 1]$, $\mathbf{b}$ is a column vector that stacks the N measurements of $g(f_{s_k})$, and $\mathbf{x} = [a' \;\; b' \;\; c']^T$ is the vector of parameters. An optimal solution to this problem can be found using least squares method. At each iteration of the algorithm 3 images are acquired with different focus values: one corresponding to the target depth estimated by the tracking system in the previous time instant, and the other two a certain distance Δ away from this value.

This method has the advantage of resulting in a closed-form solution for the focus setting value that minimizes the cost function in each instant. Moreover, it is robust to variations in parameters such as scene illumination or camera zoom and aperture values, which may change the shape of the cost function, see Fig.2, since the estimation process implemented estimates new parabola coefficients in

each iteration of the algorithm, leading to the adaptation of the cost function model to those values.

4. DEPTH ESTIMATION

In this section, the process of obtaining depth from the camera focus setting that minimizes the proposed cost function is detailed.

Usually, the operator does not have access to the quantities v and f of the lens model defined in section 2, but the values of f_s and z_s, here called the camera focus and zoom settings, respectively, are known [1]. The relation between these quantities and the depth u of a target can be expressed as follows:

$$(f_s^*, z_s) \xrightarrow[v=h_2(f_s^*,z_s)]{f=h_1(f_s^*,z_s)} (f, v) \xrightarrow{u=s(f,v)} u,$$

or, equivalently,

$$(f_s^*, z_s) \xrightarrow{u=m(f_s^*,z_s)} u, \tag{8}$$

where $u = s(f, v)$ corresponds to relation (3), and $f = h_1(f_s^*, z_s)$ and $v = h_2(f_s^*, z_s)$ express lens model parameters as a function of the camera zoom setting and the camera focus setting f_s^* that minimizes the cost function. These three transformations can be merged into a single one, $u = m(f_s^*, z_s)$, that expresses directly depth as a function of f_s^* and z_s.

In practice, it is very difficult to estimate relations $f = h_1(f_s^*, z_s)$ and $v = h_2(f_s^*, z_s)$ between camera settings and lens model parameters, since we do not have access to the last ones. Therefore, in this work, the global expression $u = m(f_s^*, z_s)$ was used, which requires a previous step of calibration.

For the camera used, the 215 PTZ model from AXIS, it is possible to verify experimentally that the relation $u = m(f_s^*, z_s)$ is well described by a second-degree polynomial of the form

$$u = m(f_s^*, z_s) = a_1 z_s^2 + a_2 z_s + a_3 {f_s^*}^2 + a_4 f_s^* + a_5, \tag{9}$$

where coefficients $(a_1, a_2, a_3, a_4, a_5)$ must be estimated, see Fig. 3.

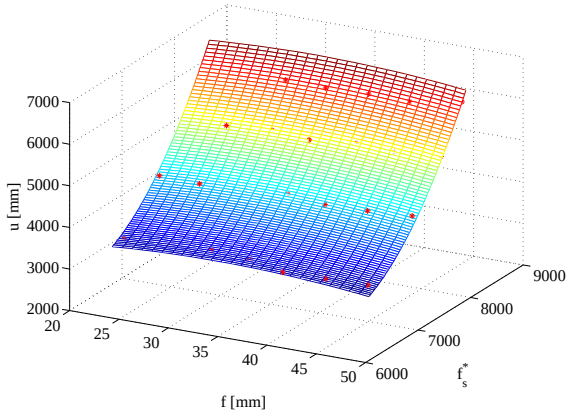

Fig. 3. Fitting of the target depth as function of the camera focal length and focus setting that minimizes the cost function. Red points correspond to experimental data.

5. EXPERIMENTAL RESULTS

In this section, experimental results that illustrate the performance of the proposed depth estimation method, when integrated in the tracking system described in [7], are presented.

Figure 4 depicts the overall target tracking system architecture, with emphasis on the depth estimation method proposed in this paper. A detailed description of the Multiple Model Adaptive Extended Estimator module (MMAE-EKF) can be found in [7], omitted here due to lack of space.

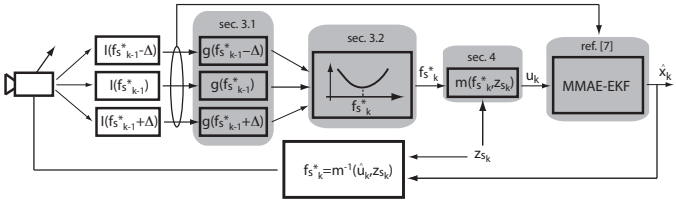

Fig. 4. Scheme of the target tracking system proposed.

Results presented in this section were obtained using the 215 PTZ camera model from AXIS and a target installed on a mobile platform. The focal length of this camera ranges from 3.8 mm to 45.6 mm and its focus setting assumes values in the interval $[1; 9999]$. Since shallower depths of field lead to smaller depth estimation errors, the maximum focal length was set. The aperture of the camera used in such conditions is F2.7. Images of size 704×576 were acquired and an initial step of lowpass filtering was performed, to reduce the influence of noise in the performance of the algorithm. For the sake of simplicity, only the red component of the images acquired by the camera was used, since the target in the experiment described in this section was red. However, the algorithm proposed in this paper is straightforward adapted to targets with other colours.

In the experiment reported in this section, the target moved along a straight line with a constant velocity of 5 mm/s between instants 62 s and 262 s. In time intervals $[0, 62]$ s and $[262, 340]$ s the target remained 3 m and 4 m away from the camera, respectively. Its 3D nominal and estimated trajectories are depicted in Fig. 5. The corresponding position, velocity, and acceleration errors are presented in Fig. 6. All errors converge approximately to 0, except in the vicinity of time instants 62 s and 262 s when the target started to move and stopped, respectively. This behaviour is related to the MMAE-EKF implemented, which requires time to converge when the target changes its type of movement abruptly. The standard deviation associated with the estimation error in the direction of z is always greater than the others since this is the direction in which the target depth is measured. In other words, the uncertainty in the target depth estimation (mainly related to coordinate z) is always greater than the uncertainty in the estimation of the target center in the image (mainly related to coordinates x and y), as expected. It is also possible to verify that the standard deviation of the position estimated by the whole target tracking system in the direction of z, approximately 13.6 mm, is smaller than the standard deviation of the measurements provided directly by the depth estimation algorithm proposed in this work, ap-

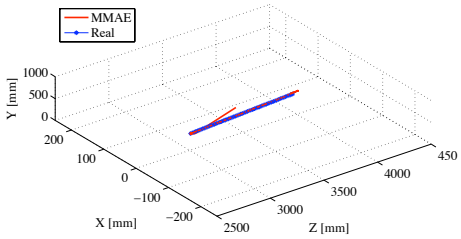

Fig. 5. 3D position estimate of a real target.

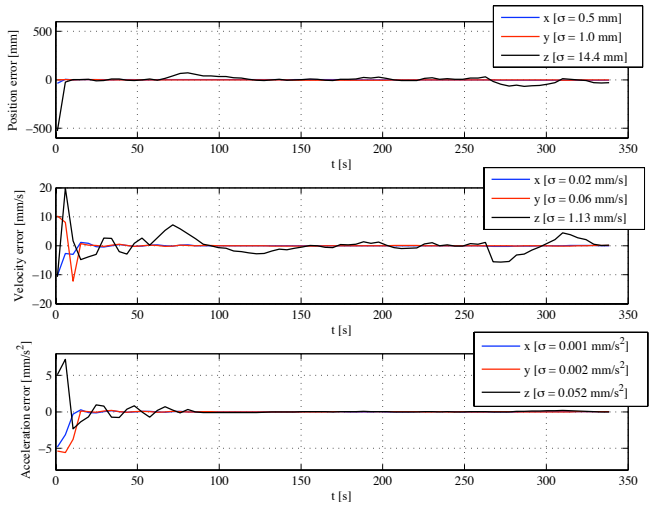

Fig. 6. Position, velocity, and acceleration estimation errors of the real target. The values of σ correspond to the standard deviation computed in the interval $[100, 250]$ s.

proximately 20.3 mm. Figure. 7 depicts the target depth evolution obtained directly from the algorithm proposed in this paper for this experiment.

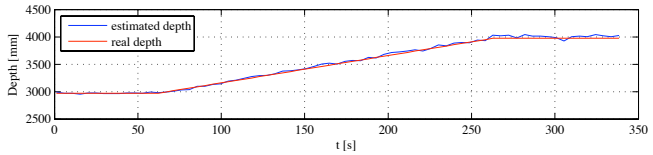

Fig. 7. Target depth estimation obtained directly from the algorithm proposed in this work.

In Fig. 8, the MMAE-EKF results on the identification of the target angular velocity are depicted. Before the target started to move, its estimated angular velocity could have converged to any value since, for stopped targets, the several models associated with different angular velocities are not distinguishable. Upon the beginning of the movement, the MMAE-EKF identifies correctly the model associated with the null angular velocity.

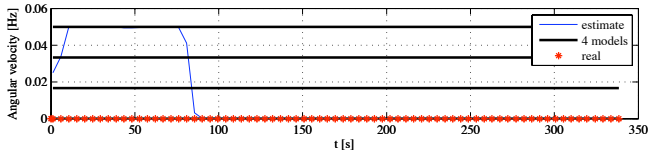

Fig. 8. MMAE-EKF target angular velocities identification.

There are several reasons that can justify the errors observed in this section: i) the uncertainty associated with the characterization of the real trajectory described by the target; ii) possible mismatches in the models considered for the camera, the target, and the lens, and iii) incomplete measurement and sensor noise characterization.

Tracking targets moving at higher velocities, when compared to the ones in this experiment, is also possible. However, the performance of the system degrades since, in such situations, the three measurements used in the fitting of the cost function are acquired with the target at very different depths. The influence of this limitation can be significantly minimized by using a camera that acquires images with different focus settings at higher rates.

6. CONCLUSIONS

A new active depth from focus method that estimates the depth of a moving target in real time using a single PTZ camera was proposed. Information present on the target boundary is used to infer depth, and combined with a (sub-)optimal MMAE results in a full 3D nonlinear tracking system. Tests and validation of the system were performed in a real scenario. Position estimates with accuracies on the order of few centimeters were obtained. The main limitation of the proposed method is the limited rate at which the target can change its depth, which is a consequence of the slow velocity at witch the camera used acquires images with different focus settings.

7. REFERENCES

[1] N. Asada, M. Baba, and A. Oda. Depth from blur by zooming. In *Proc. Vision Interface*, page 8, May 2001.

[2] Y. Bar-Shalom, X. Rong-Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, Inc., 2001.

[3] L. Bertelli, P. Ghosh, B. Manjunath, and F. Gibou. Robust depth estimation for efficient 3d face reconstruction. *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1516–1519, 2008.

[4] A. Blake and M. Isard. *Active Contours*. Springer, 1st ed. edition, 2000.

[5] A. Discant, A. Rogozan, C. Rusu, and A. Bensrhair. Sensors for obstacle detection - a survey. *Electronics Technology, 30th International Spring Seminar on*, pages 100–105, 2007.

[6] J. Ens and P. Lawrence. An investigation of methods for determining depth from focus. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(2):97–108, 1993.

[7] T. Gaspar and P. Oliveira. Single pan and tilt camera indoor positioning and tracking system. In *Proc. European Control Conference*, pages 2792–2797, 2009.

[8] I. Haritaoglu, D. Harwood, and L. Davis. W^4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809–830, 2000.

[9] E. Hecht. *Optics*. Addison-Wesley, 4th ed. edition, 2001.

[10] E. Krotkov. Focusing. *Intl. Journal of Computer Vision*, 1:223–237, Oct 1987.

[11] S. Nayar and Y. Nakagawa. Shape from focus. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(8):824–831, 1994.

[12] A. P. Pentland. A new sense for depth of field. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-9(4):523 – 531, Jul 1987.

[13] Y. Y. Schechner and N. Kiryati. Depth from defocus vs. stereo: How different really are they? In *Proc. International Conference on Pattern Recognition*, pages 1784–1786, Nov 1998.

[14] J. Schlag, A. Sanderson, C. Neuman, and F. Wimberly. Implementation of automatic focusing algorithms for a computer vision system with camera control. Technical Report CMU-RI-TR-83-14, Carnegie Mellon University, August 1983.

[15] H. Q. H. Viet, M. Miwa, H. Maruta, and M. Sato. Recognition of motion in depth by a fixed camera. In *Proc. VIIth Digital Image Computing: Techniques and Applications*, page 10, Dec 2003.