Machine Vision and Applications (2001) 13: 25–37

Machine Vision and Applications
© Springer-Verlag 2001

Zoom tracking and its applications

Jeffrey A. Fayman, Oded Sudarsky, Ehud Rivlin, Michael Rudzsky

Computer Science Department, Technion – Israel Institute of Technology, 32 000 Haifa, Israel; e-mail: {jefff, sudar, ehudr, rudzsky}@cs.technion.ac.il

Received: 15 February 2000 / Accepted: 19 June 2000

Abstract. We present a new active vision technique called zoom tracking. Zoom tracking is the continuous adjustment of a camera's focal length in order to keep a constant-sized image of an object moving along the camera's optical axis. Two methods for performing zoom tracking are presented: a closed-loop visual feedback algorithm based on optical flow, and use of depth information obtained from an autofocus camera's range sensor. We explore two uses of zoom tracking: recovery of depth information and improving the performance of scale-variant algorithms. We show that the image stability provided by zoom tracking improves the performance of algorithms that are scale variant, such as correlation-based trackers. While zoom tracking cannot totally compensate for an object's motion, due to the effect of perspective distortion, an analysis of this distortion provides a quantitative estimate of the performance of zoom tracking. Zoom tracking can be used to reconstruct a depth map of the tracked object. We show that under normal circumstances this reconstruction is much more accurate than depth from zooming, and works over a greater range than depth from axial motion while providing, in the worst case, only slightly less accurate results. Finally, we show how zoom tracking can also be used in time-to-contact calculations.

Key words: Zoom tracking – Depth from zoom – Optical flow – Active vision – Autofocus

1 Introduction

Biological vision systems are remarkably adept at continuously delivering high-quality images of salient objects to underlying visual processes. This is due, in part, to the continuous update of visual system parameters. It would be desirable to provide machine vision systems with similar capabilities.

Over the past several years, the research field of active vision has been exploring the benefits and advantages of movable visual sensory systems, as possessed by biological systems, over passive systems. Moving systems have been shown to lead to improved robustness and the elimination of ill-posed conditions in several computer vision problems [1, 3, 4]. The degrees of freedom of interest in active vision systems include both *extrinsic* parameters (eye motion) and *intrinsic* parameters (eye configuration). We explore an interesting use of an intrinsic parameter unique to mechanical active vision systems: adjustable focal length, also known as *zoom*.

Applications of zoom include the ability to image a target with maximum resolution [29], determine depth [10, 17, 21], and minimize view degeneracies [32]. However, to our knowledge, no previous work dealt with *zoom tracking:* the use of zoom to stabilize the image of an object that moves along a camera's optical axis, i.e., to keep the image at a constant size.

An object moving towards a camera produces an expanding image, while an object moving away from the camera produces a contracting image. Zoom tracking compensates this expansion or contraction through focal length adjustments, thus stabilizing the object's imaged size. This is similar in nature to smooth pursuit tracking: smooth pursuit is used to stabilize targets in the image as they move perpendicular to the optical axis, while zoom tracking stabilizes targets as they move along the optical axis. Combined, these tracking methods allow more general object motions to be stabilized.

The image size stability resulting from zoom tracking enables the use of scale-variant algorithms, such as correlation techniques. Zoom tracking can also be used to construct a dense depth map of the tracked object; this technique is much more accurate than depth from zooming, and it is usable over a wider range of object distances than depth from axial motion. Additionally, by observing the changes in focal length, measurements such as the object's velocity and time-to-contact (TTC) can be computed.

We present two methods for computing the required focal length adjustments: a closed-loop visual feedback algorithm based on optical flow, and use of depth information from an autofocus camera's range sensor.

Zooming cannot totally compensate for an object's motion along the optical axis, because such motions produce perspective distortion. We analyze the effect of this distor-

Correspondence to: E. Rivlin

tion, and derive a bound on the performance of zoom tracking.

Finally, we suggest another application for zoom tracking: TTC calculation.

The remainder of the paper is organized as follows. Related work is reviewed in Sect. 2. An overview of our approach for controlling the focal length is presented in Sect. 3. Section 4 discusses the motion model, imaging model, and optical flow used in our work. The necessary equations for focal length control are derived in Sect. 5. The effect of perspective distortion is analyzed in Sect. 6, and used to derive a bound on the performance of zoom tracking. Section 7 shows how zoom tracking can yield a depth map of the tracked object, and compares the performance of depth from zoom tracking with alternative depth reconstruction techniques. Section 8 shows how zoom tracking can be used to compute TTC. Experiments are presented in Sect. 9, and conclusions in Sect. 10.

2 Related work

Since 1985, when active vision first appeared in the literature [1, 3, 4], it has received a dramatic increase in interest. Initial work focused on building active vision devices and on understanding and transferring to these devices capabilities possessed by biological vision systems, such as saccades [7], smooth pursuit [7, 9], fixation [23], attention [30], and prediction [8]. Some work has appeared in the literature which explores uses of the zoom mechanism, a visual parameter not shared by biological vision systems. However, to our knowledge, the ability of the zoom mechanism to stabilize moving objects in the image has not been explored.

Cahn von Seelen and Bajcsy [29] present an algorithm that allows an active vision system to track targets at a varying scale while decreasing the risk of template drift. Their algorithm is based on an adaptive correlation method that selectively updates the correlation template in response to zoom-induced scale changes. They assume an external agent controls zooming and update the correlation template to compensate the resulting change of scale. This improves the performance of a template-matching tracker. Our approach is to take control of the focal length and to adjust it to ensure that the scale of an object's image is minimally affected as the object moves.

2.1 Camera models

Research in computer vision (and computer graphics) generally makes use of one of four different camera models: pin-hole, thin-lens, thick-lens, and geometric.

The pin-hole model is the simplest: it assumes that all light rays coming from the object focus through a single point (the focal point) onto the image plane [31, Chap. 1.4]. This geometrically simple model is widely used in computer vision. However, it is only accurate if the focal length, i.e., the distance between the image plane and the focal point, is negligible compared to the object distance (e.g., in aerial photography).

The thin-lens model [6] assumes an infinitely thin lens, modeled by a plane. Light rays from the object hitting the plane at a particular angle continue after leaving the plane at a modified angle. This model handles aperture effects (such as depth-of-field) better than the pin-hole model. However, it is still inadequate for modeling zoom lenses at a close range.

The thick-lens model represents a lens by two planes called the *principle planes*. Light rays entering one plane at a specific angle travel parallel to the optical axis through to the second plane, from which they exit at another angle (similar to the thin-lens model). Ignoring aperture effects, this model is equivalent to the pin-hole model with the addition of a virtual axial motion [25], and it is sufficient for modeling zoom lenses, even at close range.

The geometric lens model [11, 17] is based on a complete geometric description of all of the glass elements of the lens. It is the most detailed and accurate of the models, and can be used to design lenses and to model effects such as optical abberations, lens flare, etc. However, this model is extremely complicated to implement; the thick-lens model provides a sufficient approximation for our purposes.

2.2 Depth reconstruction

Many researchers show how to obtain depth from axial motion [2, 14, 15, 27, 33, 36]. This axial motion, i.e., the change of object-to-camera distance, can be cause by the movement of the camera, the object, or both. However, depth can also be reconstructed when both the camera and the object are static, using the zooming mechanism. This is not due to the change of magnification, which is the main purpose of zooming, but because zooming has a side effect of slight axial motion of the focal point. Thus, depth from zooming is, in fact, a special case of depth from axial motion. Since the extent of this axial motion is usually quite small (on the order of a few millimeters), depth can accurately be recovered from zooming only for relatively close objects. This is corroborated by Subbarao [25], who develops formulas relating the depth of a static object to focusing distance, aperture, and focal length using the thick-lens model. He claims that depth can accurately be recovered from these intrinsic camera parameters only within a range of up to about one hundred times the focal length.

Ma and Olsen [21] develop two depth-from-zooming methods for the pin-hole camera model applicable to static objects: analysis of optical flow and feature matching. They conclude that feature matching is more accurate and reliable, because it is less sensitive to noise than optical flow analysis. However, they only present results for synthetic models.

Lavest et al. [10, 17, 18] develop a depth reconstruction method for a static object and camera using the thick lens model. They conclude that the simpler pin-hole model can be used (instead of the more accurate thick-lens model) only if the effective change of focal point during zooming is considered, enabling 3D information to be inferred by triangulation.

Our approach, depth from zoom tracking, is more accurate than depth from zooming because the object-to-camera distance is not fixed; therefore, the extent of axial motion is greater than when only the focal length is changed. Compared to depth from axial motion techniques, depth from

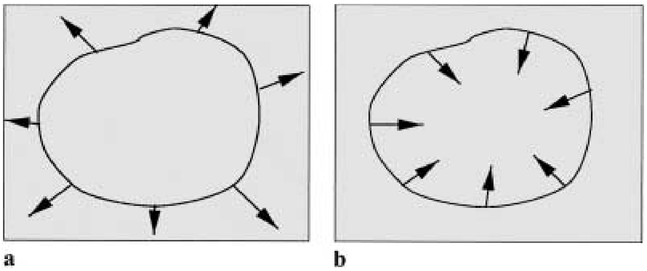


Fig. 1. **a** Divergent flow as an object approaches the camera; **b** Convergent flow as an object recedes from the camera

zoom tracking works over a greater range of object distances. Without zoom tracking, an object's axial motion can make its image too small, thus losing accuracy, or too large to fit within the image sensor.

3 Approach

Assume the zoom lens is calibrated [16, 19, 26, 34], and that the only motion the object is undergoing is translation along the optical axis of the camera. [1] By adjusting the focal length of the camera, the imaged size of an object undergoing translation along the optical axis can be kept constant. The required adjustment of focal length can be calculated in several ways, including optical flow analysis, feature matching, or observation of the focusing distance. (The latter zoom tracking technique is used in some autofocus cameras.)

Using the optical flow technique, the direction and magnitude of the flow vectors can be used to close a feedback loop controlling the focal length. The translatory motion described above gives rise to image flow which is either divergent or convergent, depending on whether the object approaches or moves away from the vision system respectively. This is illustrated in Fig. 1. If divergent flow is detected, the focal length is shortened (zoom out) to widen the field of view; if the flow is convergent, the focal length is extended (zoom in) to shrink the field of view.

Feature matching provides an inter-frame correspondence of features such as object corners and edges. This correspondence can be used in zoom tracking by determining the amount of convergence or divergence, in a similar manner to optical flow. The well known problems of optical flow [28] make feature matching an appealing alternative.

Depth information obtained from an autofocus sensor can also be used. Autofocus cameras usually measure the distance to the object at a point in the center of the image (although some cameras also have sensors at some other points in the image to handle off-center objects). If the distance to one point on the object is given, then zoom tracking can keep the object's image at a constant size based on the change of this distance over time.

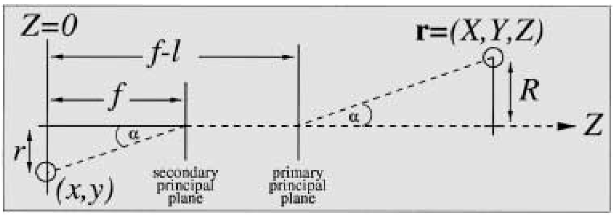


Fig. 2. The imaging model

4 Preliminaries

4.1 The imaging model

Following the notation of Horn [12, Chap. 17], denote the Cartesian coordinates of a point P on a rigid body by $\mathbf{r} = (X, Y, Z)^{\mathrm{T}}$. The body's motion is composed of a translational component $\mathbf{t} = (U, V, W)^{\mathrm{T}}$ and a rotational component $\omega = (A, B, C)^{\mathrm{T}}$. The velocity of point P is $\mathbf{V} = -\mathbf{t} - \omega \times \mathbf{r}$, or, in component form,

$$\begin{aligned} \dot{X} &= -U - BZ + CY, \\ \dot{Y} &= -V - CX + AZ, \\ \dot{Z} &= -W - AY + BX, \end{aligned} \tag{1}$$

where the dot denotes differentiation with respect to time.

Assume that the camera is static. Define the (X, Y, Z) coordinate frame such that the Z axis coincides with the camera's optical axis and the image plane is at $Z = 0$. As Lavest et al. [17] point out, the pin-hole camera model is inadequate for a zoom lens, and the thick-lens model must be used instead; however, the pin-hole model can be used if the object is virtually translated along the optical axis by the distance l between the zoom lens's principal planes. According to Subbarao [25], this distance is

$$l = f_a + f_b - \frac{f_a f_b}{f}, \tag{2}$$

where f is the current focal length of the zoom lens, and f_a, f_b are the focal lengths of the two lens groups corresponding to the two principal planes, determined during calibration. The sign of l is defined to be negative if the image plane is closer to the secondary principal plane than to the primary principal plane. See Fig. 2.

The perspective projection is

$$x = \frac{Xf}{Z + l - f}, \quad y = \frac{Yf}{Z + l - f} \tag{3}$$

where $\mathbf{r} = (X, Y, Z)^{\mathrm{T}}$ is an object point and $p = (x, y)$ is the corresponding image point. This implies

$$r = \frac{Rf}{Z + l - f}, \tag{4}$$

where $r = \sqrt{x^2 + y^2}$ and $R = \sqrt{X^2 + Y^2}$ are the distances of the image point and the object point, respectively, from the optical axis.

4.2 The motion field and the optical flow field

For optical-flow-based zoom tracking, let $(u, v) = (\dot{x}, \dot{y})$ denote the instantaneous velocity of the image point (x, y) under the perspective projection. This velocity can be obtained

[1] Discounting rotations that are not about the optical axis, this assumption is valid, because, in an active vision system, smooth pursuit tracking can stabilize motion perpendicular to the optical axis, and cyclotorsion can compensate for rotations about the optical axis.

by taking derivatives of Eq. (3) with respect to time:

$$
\begin{aligned}
u &= u_{\text{trans}} + u_{\text{rot}} + u_{\text{zoom}},\\
v &= v_{\text{trans}} + v_{\text{rot}} + v_{\text{zoom}},
\end{aligned}
\tag{5}
$$

where $(u_{\text{trans}}, v_{\text{trans}})$ is the translational component of the optical flow, $(u_{\text{rot}}, v_{\text{rot}})$ is the rotational component, and $(u_{\text{zoom}}, v_{\text{zoom}})$ is the zooming component:

$$
\begin{aligned}
u_{\text{trans}} &= \frac{-U + xW}{Z + l - f},\\
u_{\text{rot}} &= \frac{1}{f}\left(Axy - B\left(x^2 + \frac{Zf}{Z + l - f}\right) + Cy\right),\\
u_{\text{zoom}} &= \frac{\dot{f}x}{f}\left(1 + \frac{f^2 - f_a f_b}{f(Z + l - f)}\right),\\
v_{\text{trans}} &= \frac{-V + yW}{Z + l - f},\\
v_{\text{rot}} &= \frac{1}{f}\left(A\left(y^2 + \frac{Zf}{Z + l - f}\right) - Bxy - Cx\right),\\
v_{\text{zoom}} &= \frac{\dot{f}y}{f}\left(1 + \frac{f^2 - f_a f_b}{f(Z + l - f)}\right).
\end{aligned}
\tag{6}
$$

The differences between Eqs. (5), (6) and those derived by Horn are due to the different choice of the $Z = 0$ plane, f not necessarily being equal to 1, and the use of the thick-lens model rather than the pin-hole model.

Let $I(x, y, t)$ be the image intensity function, where t is time. The time derivative of I can be written as

$$
\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\partial I}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial I}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial I}{\partial t} = (I_x, I_y)\cdot \mathbf{u} + I_t = \nabla I \cdot \mathbf{u} + I_t, \tag{7}
$$

where ∇I is the image gradient, the subscripts denote partial derivatives, and $\mathbf{u} = (u, v)$ is the projected motion field (the optical flow) at the point (x, y). If we assume $\mathrm{d}I/\mathrm{d}t = 0$, i.e., the image intensity does not vary with time [13], then

$$
\nabla I \cdot \mathbf{u} + I_t = 0. \tag{8}
$$

Let $\mathbf{u} = u_\perp + u_\parallel$ where $u_\perp$ is the normal flow and $u_\parallel$ is perpendicular to $u_\perp$. Because the image gradient ∇I is parallel to $u_\perp$ and perpendicular to $u_\parallel$, only $u_\perp$ can be determined by observing ∇I locally. (This is known as the aperture problem [12, Chap. 12].) Therefore

$$
\begin{aligned}
\nabla I \cdot \mathbf{u} + I_t &= \nabla I \cdot (u_\perp + u_\parallel) + I_t = \nabla I \cdot u_\perp + \nabla I \cdot u_\parallel + I_t\\
&= \nabla I \cdot u_\perp + I_t = 0.
\end{aligned}
\tag{9}
$$

Consequently

$$
u_\perp = -\frac{\partial I}{\partial t}\frac{\nabla I}{\|\nabla I\|^2}. \tag{10}
$$

Various techniques [5] have been proposed to solve the aperture problem, that is, to recover $u_\parallel$ and integrate the measurements into a 2D flow field.

5 Adjusting the focal length

We explore two methods of finding the required adjustment of focal length to maintain a constant image size. These methods are based on optical flow and depth from an autofocus camera's range sensor, and are described in Sects. 5.1 and 5.2, respectively. We conclude that the optical-flow-based technique is not usable when employing the thick-lens model unless the depth of the object is known.

Zoom tracking keeps the image stabilized only for object points lying on one plane, called the *reference plane*, which is parallel to the image plane. All other points shift due to the effect of perspective distortion. This distortion is further explored in Sect. 6.

5.1 Optical-flow-based zoom tracking

In this section, we explore the use of the optical flow field $\mathbf{u}$ from Sect. 4 in a feedback loop algorithm which adjusts the focal length to keep the size of the object constant in the image.

The feedback loop has two phases. In phase 1, image capture, the focal length f is constant, and the approaching or receding object induces convergent or divergent optical flow, respectively. In phase 2, the focal length is adjusted in order to negate the changes in the projected object size relative to the size at the beginning of the tracking sequence. The resulting zooming action is jerky due to the suspension of zooming during phase 1. However, this is usually inevitable, as most zoom lenses are position controlled, rather than velocity controlled.

During phase 1, the focal length f is constant. The equations giving the change in projected object size are now given. Referring to the imaging model of Fig. 2, we have the relationship:

$$
r = \frac{fR}{Z + l - f}. \tag{11}
$$

Holding f constant and differentiating Eq. (11) we get:

$$
\frac{u}{r} = \frac{\dot{R}}{R} - \frac{\dot{Z}}{Z + l - f}. \tag{12}
$$

During phase 2, the focal length f changes in such a way as to compensate the change in projected object size. Keeping r constant and differentiating Eq. (11) we get:

$$
0 = \frac{\dot{R}}{R} - \frac{\dot{Z}}{Z + l - f} - \frac{\left(\frac{f_a f_b}{f^2} - 1\right)}{Z + l - f}\dot{f} + \frac{\dot{f}}{f}. \tag{13}
$$

Equations (12) and (13) imply:

$$
\dot{f} = -f\frac{u}{r}\phi(Z, f), \tag{14}
$$

where

$$
\phi(Z, f) = \frac{1}{1 + \frac{f}{Z + l - f}\left(1 - \frac{f_a f_b}{f^2}\right)}. \tag{15}
$$

The first part of Eq. (14) i.e., $\dot{f} = -fu/r$ is identical to the equation obtained for zoom tracking based on the pin-hole camera model. Using the thick-lens model, the factor $\phi(Z, f)$ is introduced. In Fig. 3, this factor is plotted as a function of object depth and focal length in order to illustrate

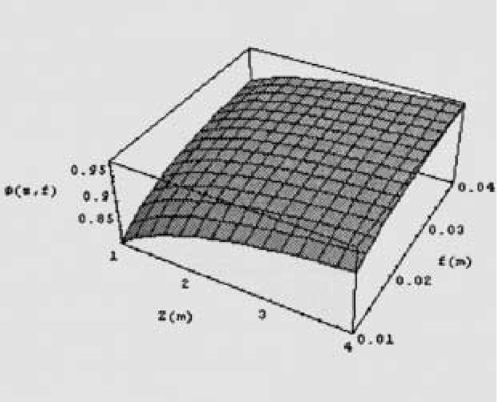


Fig. 3. Effect of the thick-lens model on zoom tracking accuracy

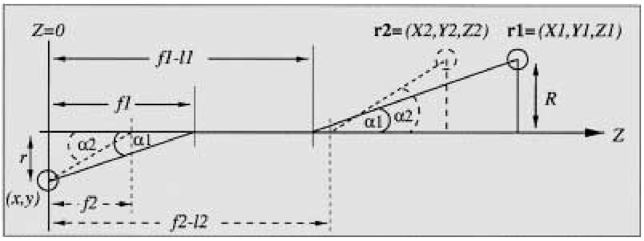


Fig. 4. Perspective geometry

its effect on zoom tracking computations.[2] The plot shows that the factor has more influence at closer distances and shorter focal lengths. While the pin-hole model cannot fully capture the behavior of a zoom lens, it can be used as a reasonable approximation when $Z >> f$.

5.2 Autofocus sensor-based zoom tracking

As mentioned in Sect. 3, the required change of focal length can also be calculated based on depth information from an autofocus sensor. Assume the only motion the object undergoes is translation along the optical axis. Suppose the object is tracked from an initial distance of Z_1 to a final distance of Z_2, as in Fig. 4. The distances Z_1 and Z_2 are given by the autofocus sensor. Let $\mathbf{r}_1$ be the initial position of a point P on the object, and let $\mathbf{r}_2$ be the final position of P. We would like to keep the point's image at a constant radial distance r from the image center as P moves from $\mathbf{r}_1$ to $\mathbf{r}_2$. This is accomplished by changing the distances between the image plane and the principal planes from given initial values of $f_1 - l_1$ and f_1 (for the primary and secondary principal planes, respectively) to final values of $f_2 - l_2$ and f_2.

Let R denote the radial distance of the point from the optical axis. The thick-lens model implies the following relationships:

$$r = \frac{f_1 R}{Z_1 + l_1 - f_1} \tag{16}$$

and

$$r = \frac{f_2 R}{Z_2 + l_2 - f_2} \tag{17}$$

where l_1 and l_2 are as in Eq. (2). Therefore

$$\frac{f_1 R}{Z_1 + l_1 - f_1} = \frac{f_2 R}{Z_2 + l_2 - f_2}. \tag{18}$$

Using Eq. (2) and solving Eq. (18) for f_2 and l_2, we obtain

$$f_2 = \frac{f_1(f_a + f_b + Z_2) + \xi}{2(l_1 + Z_1)} \tag{19}$$

$$l_2 = \frac{f_1(f_a + f_b)(f_a + f_b + Z_2) + f_b\xi + f_a(-2f_b l_1 - 2f_b Z_1 + \xi)}{f_1(f_a + f_b + Z_2) + \xi}$$

where

$$\xi = \sqrt{f_1(-4f_a f_b(l_1 + Z_1) + f_1(f_a + f_b + Z_2)^2)}. \tag{20}$$

Equation (19) gives the focal length required to keep the object at the same visual size as its distance from the camera changes. Also, given a camera with a fixed zoom range $f_{\text{wide}} \rightarrow f_{\text{tele}}$, where f_{wide} is the shortest focal length and f_{tele} is the longest, Eq. (19) gives the depth range over which an object can be tracked: given the initial distance Z_1 and the initial focal length f_1, the object can be tracked from a minimum distance of

$$d_{\text{min}} = (f_b/f_{\text{wide}} - 1)f_a - f_b + (l_1 + Z_1)f_{\text{wide}}/f_1) \tag{21}$$

to a maximum distance of

$$d_{\text{max}} = (f_b/f_{\text{tele}} - 1)f_a - f_b + (l_1 + Z_1)f_{\text{tele}}/f_1), \tag{22}$$

where l_1 is given by Eq. (2).

6 Perspective distortion

As mentioned in the previous section, zooming can accurately compensate for object translation only for object points lying on one plane, the reference plane. All object points not on this plane will undergo perspective distortion [22]. Let us begin with an intuitive description of this effect.

Suppose the projected image of an object that translates along the optical axis were to be stabilized by corresponding camera dollying, i.e., camera translation along the optical axis. Clearly the image would remain constant, as the distance between the camera and the object would not change. However, as is well known to photographers, zooming is not equivalent to dollying, because changes in focal length give rise to perspective distortion.

Strictly speaking, perspective distortion is not caused by the change of focal length, but by the change of camera-to-object distance; zooming by itself only changes overall magnification, not perspective. However, different view angles are usually associated with different viewpoints. A long focal length, which provides a narrow angle, is typically used to photograph distant objects, while nearby objects are normally taken with a wide-angle, short-focal-length lens. Therefore, it is convenient to regard zooming out as causing a deeper or more pronounced perspective and zooming in as causing a flatter or more compressed perspective.

[2] The focal lengths of the two lens groups used in the thick-lens model are taken as $f_a = -24.353$ mm and $f_b = 126.592$ mm. See Sect. 9 for an explanation.

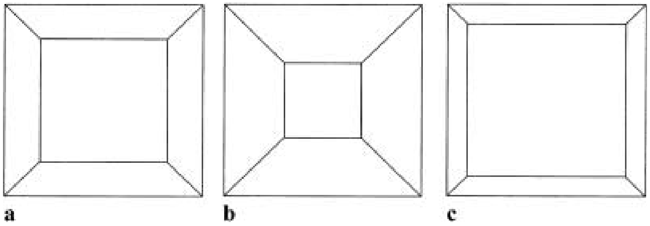


Fig. 5. **a** Reference view of a wire-frame cube; **b** Close-up view of the cube with a short focal length; **c** Distant view with a long focal length

Figure 5 illustrates this effect. Figure 5a shows a frontal view of a wire-frame cube. In Fig. 5b, the distance between the cube and the camera was decreased, and the front plane of the cube was kept the same size by zooming out. Clearly, other points on the cube (e.g., the back plane) did not remain the same size; in this case, they became smaller. In Fig. 5c, the distance between the cube and the camera was increased, and the front plane was kept the same size by zooming in. Again, other points did not remain the same size; in this case, they became larger than in Fig. 5a.

This effect has been put to artistic use in movies such as Alfred Hitchcock's *Vertigo*. In that movie, a man suffering from vertigo looks down a tall tower, which appears to shrink and stretch. This was achieved by zooming and dollying in opposite directions at the same time.

An analysis and quantification of perspective distortion provides an upper bound on the residual error of the zoom tracking process.

The thick-lens imaging model is given by Eq. (4). If point P is on the reference plane that is being tracked, then its image remains constant at $r = Rf_1/(Z_1+l_1-f_1) = Rf_2(Z_2+l_2-f_2)$. However, if P is not on the reference plane, but is at a distance z from this plane, then its image shifts from $r_1 = Rf_1/(Z_1+z+l_1-f_1)$ to $r_2 = Rf_2/(Z_2+z+l_2-f_2)$. The perspective distortion of point P is thus given by

$$r_2 - r_1 = \frac{Rf_2}{Z_2+z+l_2-f_2} - \frac{Rf_1}{Z_1+z+l_1-f_1}. \tag{23}$$

If our goal is to stabilize the object's image, e.g., for recognition purposes, then perspective distortion is error or "noise" in the stabilization process (in addition to any error induced by inaccuracies in the calculation of optical flow or autofocus sensing). This error is maximized at points of the object that are farthest away from the reference plane and from the optical axis.

If the object has a bounding box, then the maximum error e is attained at the corner of the box that is farthest away from the reference plane. Let z be the maximum distance of any point in the box from the reference plane, and let R be the maximum distance of any point in the box from the optical axis. By Eq. (23), the perspective distortion error e of any point on the object is bounded by

$$e \le R\left|\frac{f_2}{Z_2+z+l_2-f_2} - \frac{f_1}{Z_1+z+l_1-f_1}\right|. \tag{24}$$

Equation (24) specifies the maximum error that can be expected, given object size and depth difference between two object positions (ignoring zoom range constraints and inaccuracies of flow calculations or autofocus sensing). Alternatively, given a maximum tolerable error and an object size, initial object distance, and initial focal length, Eq. (24) can be used to calculate the range of object distances that may be tracked without exceeding this error.

7 Depth from zoom tracking

This section shows how depth can be recovered by zoom tracking, and compares depth from zoom tracking (DfZT) with two alternative techniques: depth from zooming (DfZ) and depth from axial motion (DfAM). We show that DfZT can provide better results than either of the others.

In DfZ, it is assumed that both the camera and the object are static. The zooming sequence from which depth is computed is generated by changing only the focal length [21]. In DfAM, the object-to-camera distance is assumed to change, either by camera motion or by object motion, but the lens's focal length is fixed. DfZT changes both object distance and focal length simultaneously.

In DfZ and DfZT, the extent of intrinsic zooming is assumed to be known, because the lens can be calibrated in advance. Since zooming is the only motion in DfZ, this technique recovers absolute depth. However, DfAM and DfZT also involve extrinsic (object) axial motion, whose extent may or may not be known, depending on the application. If it is known, then absolute depth can be recovered; otherwise, only relative depth can be reconstructed. We derive formulas for both cases, and compare the accuracy of absolute depth reconstruction by all three techniques.

DfAM normally provides better results than DfZ. As we will show, DfZT is significantly more accurate than DfZ under normal circumstances, and at worst is only slightly less accurate than DfAM. However, this is compensated for by the fact that DfZT is usable over a wider range of object distances than DfAM.

All three depth reconstruction techniques are based on correspondences between image points. Because of the discrete nature of the image sensor, correspondence can only be determined up to a finite error $\varepsilon > 0$. This error may be on the order of a sensor pixel size, or smaller if sub-pixel resolution techniques are used. For example, a video camera usually has a pixel size of about 10 μm (or slightly less). Sub-pixel resolution can determine correspondences with improved accuracy, but which is still linearly dependent on pixel size (e.g., $1/5$ pixel). Therefore, depth can also be reconstructed only up to a certain error.

The accuracies of the techniques are compared for typical working conditions: a correspondence error bound of $\varepsilon = 2\ \mu$m and an object point with radial distance $R = 12$ cm. The focal lengths of the two lens groups used in the thick-lens model are taken as $f_a = -24.353$ mm and $f_b = 126.592$ mm. These are the parameters of an actual Angenieux zoom lens, described by Lavest et al. [17]. They were obtained by solving Eq. (2) for f_a and f_b, using the values quoted in that paper: $f_1 = -21.93$ mm, $l_1 = -38.34$ mm, $f_2 = -80.39$ mm, $l_2 = -63.89$ mm.

7.1 DfZ

Let R be the radial distance of an object point. In DfZ, the point's depth Z is fixed, and given by

$$Z = \frac{(r_1 - r_2)f_1 f_2 + f_1 l_2 r_2 - f_2 l_1 r_1}{f_2 r_1 - f_1 r_2}, \tag{25}$$

where $r_1 = f_1 R/(Z + l_1 - f_1)$ is the radial distance of the point's image at focal length f_1 and $r_2 = f_2 R/(Z + l_2 - f_2)$ is the corresponding distance at focal length f_2 (see Fig. 2), $l_1 = f_a + f_b - f_a f_b/f_1$, and $l_2 = f_a + f_b - f_a f_b/f_2$. Both r_1 and r_2 are detected only up to ε. Therefore Z can be determined up to an error of

$$\varepsilon_{\mathrm{DfZ}} = \left|\frac{\partial Z}{\partial r_1}\right| \varepsilon + \left|\frac{\partial Z}{\partial r_2}\right| \varepsilon \tag{26}$$

$$= \left|\frac{(Z+l_1-f_1)(Z+l_2-f_2)(f_1 Z+f_2 Z+f_1 l_2+f_2 l_1-2f_1 f_2)}{R f_1 f_2 (l_1-l_2-f_1+f_2)}\right| \varepsilon.$$

Figure 6a shows $\varepsilon_{\mathrm{DfZ}}$ for $Z = 1\,\mathrm{m}$. The telephoto end of the focal length range f_1 is between 15 mm and 48 mm, and the wide-angle end f_2 varies between 6 mm and 12 mm. The longest focal length range (6 – 48 mm) was chosen because this is the zoom range of the Canon VC-C1 camera used in our experiments (see Sect. 9.1), and because we consider it to be representative of affordable video camera specifications. The shortest focal length range (12 – 15 mm) was chosen to keep f_2 always less than f_1, and sufficiently smaller than f_1 to keep $\varepsilon_{\mathrm{DfZ}}$ from getting too large. Obviously, $\varepsilon_{\mathrm{DfZ}}$ is better (smaller) the longer the zoom range, and worse (larger) the shorter the zoom range; if the zoom range shrinks to zero (i.e., $f_1 = f_2$) then $\varepsilon_{\mathrm{DfZ}}$ explodes to infinity.

7.2 DfAM

In DfAM, the focal length f is fixed and the point's depth varies from Z_1 to Z_2. If the extent of motion $d = Z_1 - Z_2$ is not known, then only the point's relative depth can be recovered. If d is known, then absolute depth can be recovered.

7.2.1 Relative depth

Let Z_1^a and Z_1^b be the depths of two object points, a and b, at the initial camera-to-object distance, and let Z_2^a and Z_2^b be these points' depths at the final distance. The object is assumed to be rigid and to undergo only axial motion, therefore $Z_1^a - Z_2^a = Z_1^b - Z_2^b$. Let r_1^a, r_1^b, r_2^a, and r_2^b be the radial distances of these points' images, respectively. Then the following ratio holds between the depth Z_1^a, Z_1^b of the two object points a, b:

$$\frac{Z_1^a + l - f}{Z_1^b + l - f} = \frac{r_2^a(r_2^b - r_1^b)}{r_2^b(r_2^a - r_1^a)}. \tag{27}$$

7.2.2 Absolute depth

If d is known, then a point's absolute depth can be recovered:

$$Z_1 = f - l - \frac{d r_2}{r_1 - r_2}, \tag{28}$$

where $r_1 = Rf/(Z_1 + l - f)$ and $r_2 = Rf/(Z_2 + l - f)$ are the radial distances of the point's image at depths Z_1 and Z_2, respectively.

Again, r_1 and r_2 are only known up to ε, so Z_1 can be calculated with error

$$\varepsilon_{\mathrm{DfAM}} = \left|\frac{\partial Z_1}{\partial r_1}\right| \varepsilon + \left|\frac{\partial Z_1}{\partial r_2}\right| \varepsilon$$

$$= \left|\frac{(Z_1 + l - f)(Z_2 + l - f)(Z_2 + Z_1 + 2l - 2f)}{Rf(Z_1 - Z_2)}\right| \varepsilon. \tag{29}$$

Figure 6b shows $\varepsilon_{\mathrm{DfAM}}$ for $f = 12\,\mathrm{mm}$. The far end of the object's axial motion range Z_1 varies between 1.25 m and 4 m, and the near end Z_2 is between 0.5 m and 1 m. These distances were chosen because they are comfortable working distances in a laboratory or a small studio setting, and for comparison with $\varepsilon_{\mathrm{DfZ}}$ (see Sect. 7.1). Notice that, for the purpose of comparing $\varepsilon_{\mathrm{DfZ}}$ and $\varepsilon_{\mathrm{DfAM}}$, we chose a constant ratio of 0.012 between the focal length f in DfZ and the depth Z in DfAM. For example, the shortest f for DfZ was 6 mm whereas the closest Z for DfAM was 0.5 m, and $6\,\mathrm{mm}/0.5\,\mathrm{m} = 0.012$.

7.3 DfZT

In DfZT, both the focal length and the camera-to-object distance change. As in DfAM, if the extent of axial motion is known then absolute depth can be determined; otherwise, only relative depth reconstruction can be attained.

7.3.1 Relative depth

Let Z_1^a, Z_1^b, Z_2^a, Z_2^b, r_1^a, r_1^b, r_2^a, and r_2^b be defined as in relative depth from axial motion. Then

$$\frac{Z_1^a(f_1 r_2^a - f_2 r_1^a) + (f_1 - l_1) r_1^a f_2}{Z_1^b(f_1 r_2^b - f_2 r_1^b) + (f_1 - l_1) r_1^b f_2} = \frac{r_2^a}{r_2^b}. \tag{30}$$

7.3.2 Absolute depth

If the axial motion $d = Z_1 - Z_2$ is known, then the absolute depth Z_1 can be recovered:

$$Z_1 = \frac{(r_1 - r_2) f_1 f_2 + f_1 l_2 r_2 - f_2 l_1 r_1 - d f_1 r_2}{f_2 r_1 - f_1 r_2}, \tag{31}$$

where the radial distances of the point's images are $r_1 = R f_1/(Z_1 + l_1 - f_1)$ and $r_2 = R f_2/(Z_2 + l_2 - f_2)$, respectively. ($d$ may be known because, in this case, zoom tracking is done in order to reconstruct depth rather than to stabilize a moving object's image.) r_1 and r_2 are only measured up to ε; therefore, Z_1 is calculated with error

$$\varepsilon_{\mathrm{DfZT}} = \left|\frac{\partial Z_1}{\partial r_1}\right| \varepsilon + \left|\frac{\partial Z_1}{\partial r_2}\right| \varepsilon$$

$$= \left|\frac{(Z_1+l_1-f_1)(Z_2+l_2-f_2)(f_1 Z_2+f_2 Z_1+f_1 l_2+f_2 l_1-2f_1 f_2)}{R f_1 f_2(l_1-l_2-f_1+f_2+Z_1-Z_2)}\right| \varepsilon. \tag{32}$$

Figure 7a shows the error of DfZT. For comparison with Fig. 6, the error $\varepsilon_{\mathrm{DfZT}}$ is plotted as a function of two parameters, r_f and r_Z, which control the range of focal lengths and object depths, respectively: $f_1 = (1.25 + 2.75\,r_f)f$, $f_2 = (1 - 0.5\,r_f)f$, and similarly for Z_1, Z_2. The range

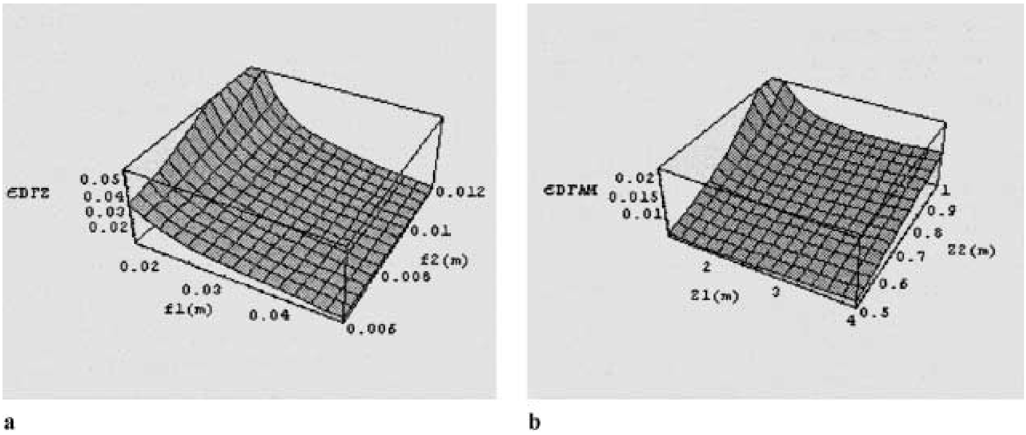


Fig. 6. a Error of depth from zooming; **b** Error of depth from axial motion

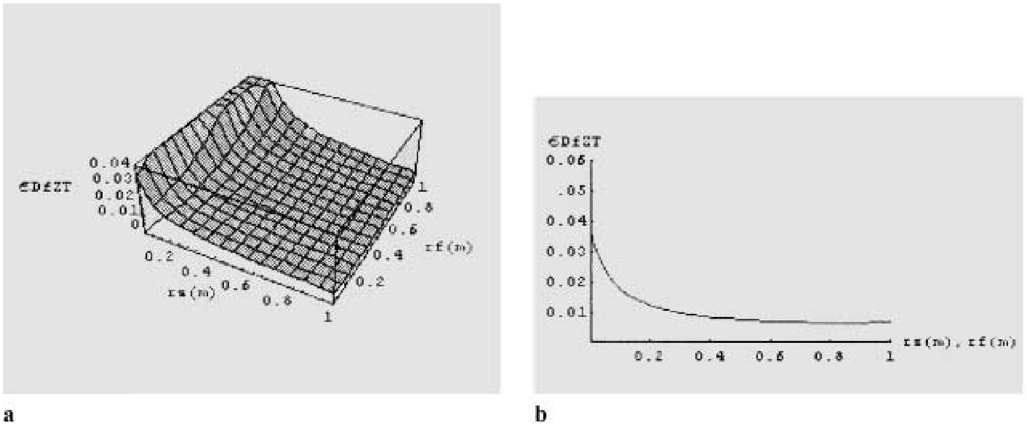


Fig. 7a,b. Error of depth from zoom tracking

is smallest when the parameter is equal to 0 and greatest when it is 1. The "normal" focal length is $f = 12\,\text{mm}$, and the "normal" distance is $Z = 1\,\text{m}$. This yields values comparable to those used in Fig. 6.

Under normal circumstances, zoom tracking uses a ratio of focal lengths similar to the ratio of depths. (The ratios are not necessarily identical, both due to the virtual translation caused by zooming and because the point at which the error is measured is not necessarily on the tracked reference plane.) To get an idea of the practical accuracy of DfZT, Fig. 7b shows a cross-section of the graph in Fig. 7a along the diagonal $r_Z = r_f$.

7.4 Comparison

In all cases, depth can be reconstructed as a result of the change of the distance between the object and the camera, whether this change is due to object motion, camera motion, or virtual translation caused by zooming. However, zooming usually produces only relatively small axial motion, because the focal lengths involved are quite short. A typical video camera with a 1/2-in. diagonal image sensor might have a zoom range of 6–48 mm. The maximum achievable virtual translation, attained by zooming all the way from the widest setting to the longest, is only about 49 cm (for $f_a = -24.353\,\text{mm}, f_b = 126.592\,\text{mm}$). This is usually far smaller than the axial motion range involved in DfAM and DfZT.

The errors of all three absolute depth reconstruction methods depend linearly on the error ε with which the image points can be detected. For example, if image points can only be detected up to 1 pixel resolution, rather than 1/5 pixel, then the resulting depth reconstruction error would be scaled by 5. Furthermore, the ratios between the techniques' accuracies, shown below, do not depend on ε.

Figure 8 shows the ratios of the error of DfZT to the errors of DfZ and DfAM. As in Fig. 7, the errors are computed as a function of parameters r_f and r_Z, and cross-sections along the diagonal $r_Z = r_f$ are also shown.[3] These graphs show that, for the chosen range of object depths and focal lengths, DfZT is three times more accurate than DfZ, and in the worst case only slightly less accurate than DfAM.

While DfAM may be a little more accurate than DfZT, it only works over a limited range of camera-to-object depths. It cannot cope with objects that are too close to the camera, because their image is too big to fit within the image sensor; neither does it deal very well with distant objects, whose small image size implies inaccurate depth reconstruction. DfZT can do better than DfAM by zooming out for near objects and zooming in for remote ones, thus it handles a greater range of object depths. In practical applications, this more than compensates for DfZT's slightly lower accuracy.

[3] Recall that, as mentioned in Sect. 7.3, under normal working conditions one would choose r_f to be equal to r_Z to compensate for the object's axial motion and keep its image constant-sized.

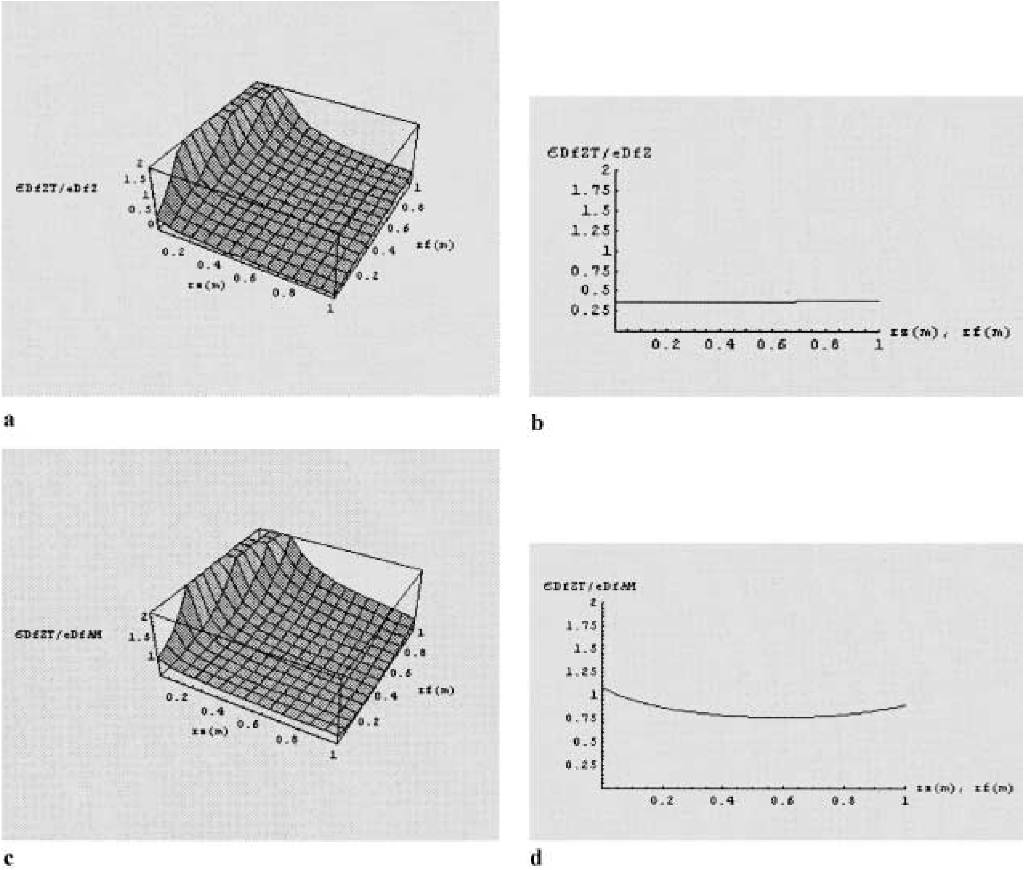


Fig. 8. **a**, **b** $\varepsilon_{\mathrm{DfZT}}/\varepsilon_{\mathrm{DfZ}}$; **c**, **d** $\varepsilon_{\mathrm{DfZT}}/\varepsilon_{\mathrm{DfAM}}$

8 Zoom tracking and time-to-contact

The TTC for a given scene point is the time until collision of the observer with the point. In this section, we show that TTC can be computed in a zoom tracking system.

According to Eq. (14) and assuming that $\phi(Z, f) \simeq 1$,

$$\dot{f} = -f\frac{u}{r} \tag{33}$$

which is equivalent to

$$\frac{\dot{f}}{f} = -\left(\frac{U}{R} - \frac{W}{Z}\right). \tag{34}$$

In zoom tracking, we have assumed that object motion consists only of translation along the optical axis (see footnote 1 in Sect. 3 for a justification of this assumption). The motion component $\frac{U}{R}$ is 0, and

$$\frac{\dot{f}}{f} = \frac{W}{Z}. \tag{35}$$

According to Eq. (35), there is a linear relationship between the change in focal length and change in object distance. If the object approaches at constant speed, the focal length will decrease at constant speed; if the velocity increases at an exponential rate, the focal length will decrease at an exponential rate.

Given some model of object motion, we can find the expected TTC. For example, assuming constant velocity, as shown in Fig. 9, there exists a linear relationship between distance and time. Therefore, a similar relationship also exists between focal length and time. The TTC can be computed by simple linear extrapolation.

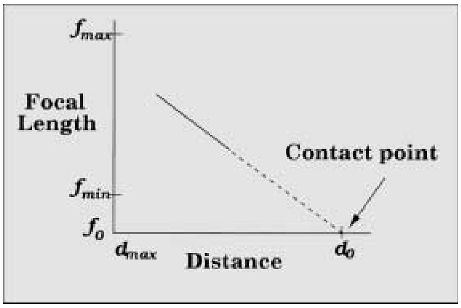


Fig. 9. Time-to-contact assuming constant velocity

9 Experiments

In this section, the usefulness of zoom tracking and its ability to reconstruct depth are demonstrated. The experiment shows the beneficial effect that zoom tracking has on a scale-variant algorithm, namely template matching. Finally, the accuracy of the depth equations is verified and depth reconstruction using zoom tracking on a synthetic image is performed. We begin with a description of the experimental platform.

Fig. 10. The Canon VC-C1 communication camera

9.1 Experimental platform

Our experiments were conducted in the Intelligent Systems Laboratory at the Technion – Israel Institute of Technology. The system consisted of a Silicon Graphics O2 workstation connected to a Canon VC-C1 communication camera. The Canon camera is a small pan-tilt unit with a serial interface allowing computer control of the extrinsic parameters of pan and tilt, as well as the intrinsic parameters of focus, zoom, and iris. The lens on the Canon camera provides 8× magnification, with focal length adjustments from $6 - -48$ mm. Figure 10 illustrates the VC-C1 camera used in our experiments. Although the VC-C1 camera allows for controlled adjustment of the focal length, it is not an ideal platform for performing zoom tracking. We encountered several difficulties in our experiments that were due to the VC-C1. In particular, accurate focal length adjustments on the VC-C1 cannot be made continuously. The VC-C1 uses a nonlinear quantization of its span of possible focal lengths which introduces a small oscillation when the focal length is changed.

9.2 Zoom tracking implemented

We began with optical-flow-based zoom tracking; however, we were unable to achieve reasonable performance due to unstable flow measurements resulting from noise in the image formation process. We therefore opted for a feature-based approach in which we used a detector to identify corners in subsequent images. The detector we used was provided by an image processing package called SUSAN (smallest univalue segment assimilating nucleus) which covers image noise filtering, edge finding, and corner finding. For more information on the SUSAN system, see Smith and Brady [24].

The corner detector returned a list of corners $L1$ and $L2$ for each image. Correspondence between the corners $L1_i$ and $L2_j$ from subsequent images was determined by exploiting the zoom-tracking-induced constraint that a corner moves radially from the center of the image. We identified the polar coordinates (r, θ) of each $L1_i$ and $L2_j$ with respect to the center of the image. Theoretically, $\theta_i = \theta_j$ and the matching corner will lie on the radial vector with angle θ. In practice, this is usually not the case, and we search for the matching corner using an SSD (sum of squared differences) template matcher within a radial segment with an angular spread of $\pm 1°$ as shown in Fig. 11. After the correspondences between $L1_i$ and $L2_j$ were found, we applied a discrete version of the proposed formula $\dot{f} = -fu/r$. The focal length was changed by the average focal length computed for each corner.

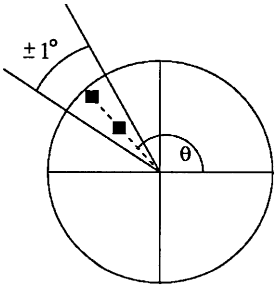


Fig. 11. Correspondence was searched for within a radial segment of $\theta \pm 1°$

Zoom tracking was performed on several images. Three of them are shown in Fig. 12. The results of the corner detector are displayed in the top left corner of Fig. 12b. The pictures were attached to a robotic manipulator which performed translation of the picture along the optical axis of the camera such that the distance from the camera to the picture ranged from 900 mm to 1400 mm (a span of 500 mm). Figure 13a, b shows the picture at 900 mm and 1400 mm, respectively, when zoom tracking was not employed. Figure 13c shows the picture at 1400 mm when zoom tracking was used.

9.3 Zoom tracking and template matching

In this experiment, we show the effects of zoom tracking on template matching, a scale-variant algorithm. A standard SSD tracker was implemented. The idea behind template matching is to find the location of a particular object in an image by searching the image for instances of a second, smaller image called a “template” which contains the object. The template-matching algorithm compares the template with the image at different image locations and finds the location in the image which best matches the template. Correlation provides the basis of template matching. For each image location, a similarity measure is computed indicating how well the template matches the image at that location. The image location that provides the maximal similarity measure is selected as the location of the object in the image.

The SSD tracking method is a classic example of an algorithm that is scale variant. Indeed, researchers have proposed a variety of techniques to handle this problem, such as updating the correlation template in response to scale changes in an image sequence [29] and providing templates at different scales and orientations [35].

The tracker that computes the similarity of patches P in an image $I(x, y)$ to a template $T(x, y)$, is defined as follows:

$$E(i, j) = \sum_{x,y \in P} (T(x, y) - I(x - i, y - j))^2 . \qquad (36)$$

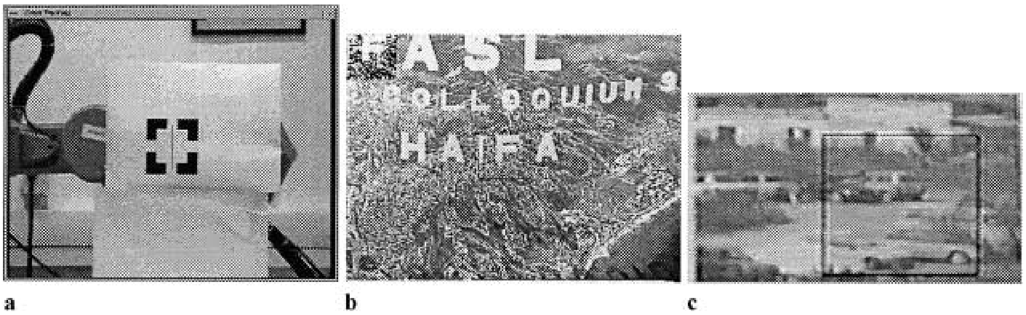


Fig. 12. Images used for zoom tracking: **a** corner image; **b** poster; **c** parking lot

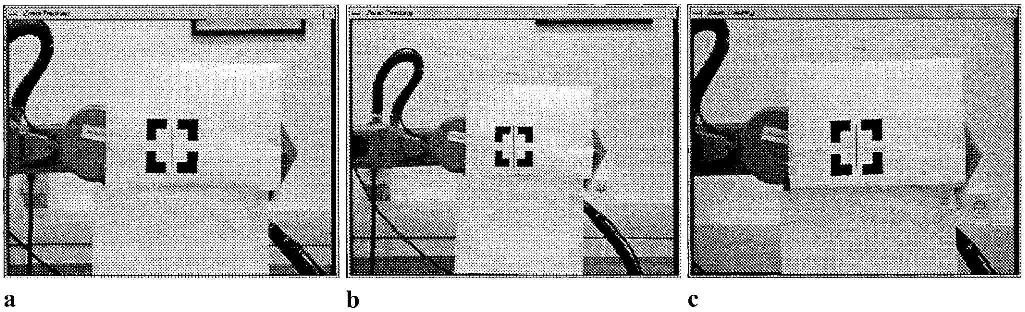


Fig. 13. a Corner image captured at 900 mm; **b** corner image captured at 1400 mm when zoom tracking was not used; **c** corner image captured at 1400 mm when zoom tracking was employed

9.3.1 Template matching performance with/without zoom tracking

The performance of the SSD tracker was measured as it tracked a moving picture of a light bulb. The picture was again attached to the robotic manipulator which translationed the picture along the optical axis of the camera. In this experiment, the distance from the camera to the picture ranged from 400 mm to 700 mm (a span of 300 mm). The template was captured at the halfway location of 550 mm with a focal length of 21 mm. Images of the light bulb picture at the near, middle, and far distances are shown in Fig. 14.

Figure 15a presents the results of the tracker when no focal length adjustments were made. Figure 15b shows the results of the tracker when zoom tracking was employed. In the case of no zoom tracking, we expect that the closer the target location is to the location that the template was captured, the lower the error will be. This is indeed the case, as can be seen in the graph. Lower errors occur when the distance between the current target location and the location of template capture is close to 550 mm. When we introduce zoom tracking to the scenario, we find that the performance of the SSD tracker improves, as shown in Fig. 15b. (The spikes in both graphs are due to the autofocus mechanism and to the quantized focal length of the Canon camera.)

9.4 Simulations

For this experiment the zoom tracking technique was implemented in C. OpenGL was used to generate and display the simulated coordinates of a wire frame cube with an edge length of 0.2 m, centered around $(0, 0, 0)$. OpenGL renders 2D images of 3D objects using the pin-hole camera model. In order to use the thick-lens model, the viewpoints given to OpenGL were virtually translated along the optical axis according to Eq. (2). The focal lengths of the two simulated lens groups were $f_a = -24$ mm, $f_b = 126$ mm, and the simulated sensor size was 15 mm square. The cube was zoom-tracked from the initial viewpoint $(0.3, 0.2, 0.1)$ m, looking toward the center of the cube, along an axial motion of 0.3 m. The results show the cube maintained a constant visual size, albeit affected by perspective distortion. See Fig. 16a.

To verify the correctness of Eq. (31), the depths of the eight corners of the cube were reconstructed by DfZT. The calculated depths were identical to the real distances of these points from the image plane. The real depths and computed depths are shown in Fig. 16b.

Equation (32) was verified by comparing its predicted upper bound on the depth error with the actual error, introduced by changing r_1 and r_2 by $\varepsilon = 4\ \mu$m in opposite directions. The results (Fig. 16c) show that the formula for $\varepsilon_{\mathrm{DfZT}}$ is indeed a tight upper bound on the depth error incurred by inaccurate image-space measurements.

To use more realistic measurements, we rendered the cube shaded, and used the SUSAN corner detector described earlier to identify the corners of the shaded cube before and after movement. The cube before movement with the identified corners is shown in Fig. 17a.

The reconstructed vs. real depths of the matched corners are shown in the Fig. 17b, and the depth reconstruction errors vs. the predicted upper bounds on these errors are shown in Fig. 17c. The correspondence error for Fig. 17c was taken as $\epsilon = 30\ \mu$m, which is the size of each pixel in a 500×500 pixel image simulating a 15 mm square sensor. Note that the corner matcher only found six pairs of the seven visible matching corners.

10 Conclusions

Zoom tracking exploits the adjustable focal length parameter in an active vision system to compensate for the scale changes of a target moving along a camera's optical axis. We presented two methods by which zoom tracking can be

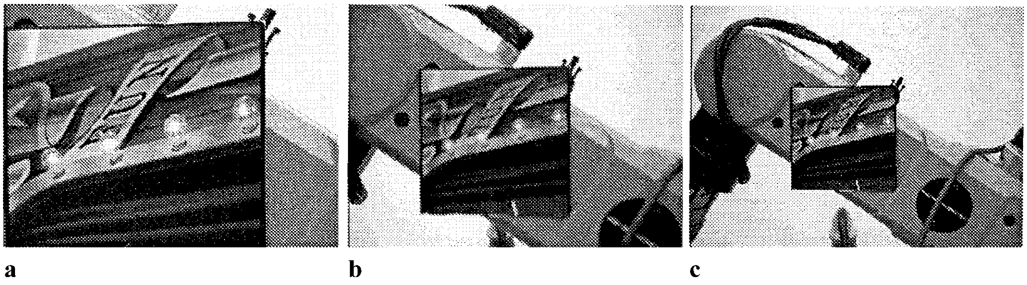


Fig. 14. **a** View of light bulb picture at 400 mm; **b** View of light bulb picture at 550 mm; **c** View of light bulb picture at 700 mm

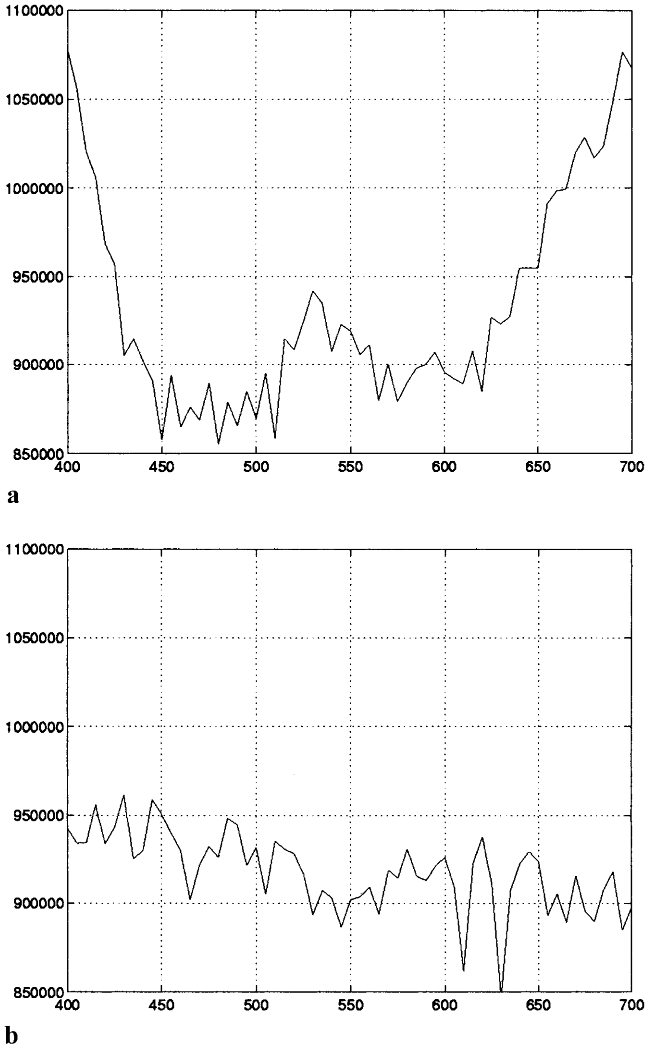
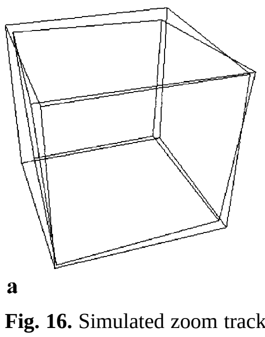


Fig. 15. **a** Template matching without zoom tracking; **b** Template matching with zoom tracking



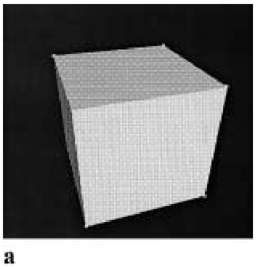
Depth	
Predicted	Actual
5.34523	5.34522
3.74166	3.74166
4.27618	4.27618
2.67261	2.67261
4.81070	4.81070
3.20713	3.20713
3.74166	3.74166
2.13809	2.13809

b

Depth Error	
Predicted	Actual
0.3500170	0.3339680
0.0627759	0.0620896
0.0864234	0.0852381
0.0426899	0.0422982
0.1338010	0.1312000
0.0490695	0.0486059
0.0627759	0.0620899
0.0613955	0.0604848

c

Fig. 16. Simulated zoom tracking: **a** visual results; **b** reconstructed and real depths; **c** predicted and real depth errors



Depth	
Predicted	Actual
4.24300	4.27618
3.89122	3.74166
2.67948	2.67261
2.12908	2.13809
3.73114	3.74166
3.16224	3.20713

b

Depth Error	
Upper Bound	Actual
0.648175	0.0.03317840
0.470820	0.14956300
0.320174	0.00686710
0.460466	0.00900603
0.470820	0.01052130
0.368021	0.04489250

c

Fig. 17. Zoom tracking with synthetic image and corner matching: **a** detected corners; **b** reconstructed and real depths; **c** upper bounds and real depth errors

performed: a two-step closed-loop feedback algorithm that measures the optical flow's radial component and then adjusts the focal length to negate it, and utilization of depth information from an autofocus sensor.

Zoom tracking yields exact results only for object points lying on a reference plane. Points not lying on this plane shift due to perspective distortion. A quantitative measure of this distortion was derived and used as an upper bound on the residual error of zoom tracking.

Zoom tracking can be used as a means to recover a dense depth map of the tracked object. This method was shown to be significantly more accurate than depth from zooming, and at the worst case only slightly less accurate than depth from axial motion, but useful over a greater range of object depths.

Experiments demonstrated that zoom tracking enables the use of scale-variant algorithms for objects whose motion is not parallel to the image plane. Specifically, a correlation-based tracker attempting to track a target moving along the optical axis fails without zoom tracking, but successfully tracks the target when zoom tracking is used.

Future work includes integration of zoom tracking with aperture and focus control, and utilization of the synergistic relationship between zoom tracking and smooth pursuit tracking.

References

1. Aloimonos Y, Weiss I, Bandyopadhyay A (1988) Active vision. Int J Comput Vision 2: 333–356
2. Alvertos N, Brzakovic D, Gonzalez RC (1989) Camera geometries for image matching in 3-d machine vision. IEEE Trans Pattern Anal Mach Intell 11(9): 897–915
3. Bajcsy R (1985) Active perception vs passive perception. In: Proc. of the 3rd IEEE Workshop on Computer Vision, Bellaire, Michigan, pp 55–59
4. Ballard DH (1991) Animate vision. Artif Intell 48: 57–86

5. Barron JL, Fleet DJ, Beauchemin SS (1981) Performance of optical flow techniques. In: Proc. DARPA Image Understanding Workshop, pp 121–130
6. Born M, Wolf E (1993) Principles of optics. Pergamon Press, Oxford
7. Bradshaw KJ, McLauchlan PF, Reid ID, Murray DW (1994) Saccade and pursuit on an active head/eye platform. Image Vision Comput 12(3): 155–163
8. Brown C (1990) Prediction and cooperation in gaze control. Biol Cybern 63(1): 61–70
9. Coombs D, Brown C (1992) Real-time smooth pursuit tracking for a moving binocular head. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Champaign, Illinois, pp 23–28, IEEE Computer Society Press, Los Alamitos, California
10. Delherm C, Lavest JM, Dhome M, Lapresté JT (1996) Dense reconstruction by zooming. In: 4th European Conference on Computer Vision, Cambridge, UK, pp 427–438
11. Heidrich W, Slusallek P, Seidel HP (1997) An image-based model for realistic lens systems in interactive computer graphics. In: Proc. Graphics Interface '97, Kelowna, B.C., pp 68–75
12. Horn BKP (1986) Robot vision MIT Press, Cambridge, Mass.
13. Horn BKP, Schunck BG (1981) Determining optical flow. Artif Intell 17: 189–203
14. Itoh H, Miyauchi A, Ozava S (1984) Distance measuring method using only simple vision constructed for moving robots. In: Proc. of the 7th Int. Conf. on Pattern Recognition, Montreal, Canada, pp 192–195
15. Jain R, Bartlett L, O'Brien N (1987) Motion stereo using egomotion complex logarithmic mappings. IEEE Trans Pattern Anal Mach Intell 9(3): 356–369
16. Kanatani K (1992) Statistical analysis of focal-length calibration using vanishing points. IEEE Trans Robotics Autom 8: 767–775
17. Lavest JM, Rives G, Dhome M (1993) Three-dimensional reconstruction by zooming. IEEE Trans Robotics Autom 9(2): 196–207
18. Lavest JM, Rives G, Dhome M (1995) Modeling an object of revolution by zooming. IEEE Trans Robotics Autom 11(2): 267–271
19. Li MX, Lavest JM (1996) Some aspects of zoom lens camera calibration. IEEE Transn Pattern Anal Mach Intell 18(11): 1110–1114
20. Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proc DARPA Image Understanding Workshop, pp 121–130
21. Ma J, Olsen SI (1990) Depth from zooming. J Am Opt Soc 7(10): 1883–1890
22. Mobasseri BG (1992) Focal length and compression of space. In: Proc. of the 11th International Conference on Pattern Recognition, pp 686–687
23. Pahlavan K, Uhlin T, Eklundh JO (1993) Dynamic fixation. In: 4th International Conference on Computer Vision, Berlin, Germany, pp 412–419
24. Smith SM, Brady JM (1997) SUSAN – a new approach to low level image processing. Int J Comput Vis 23(1): 45–78
25. Subbarao M (1988) Parallel depth recovery by changing camera parameters. In: Proc. of the International Conference on Computer Vision, Tampa, Florida, pp 149–155
26. Tarabanis K, Tsai RY, Goodman DS (1994) Calibration of a computer-controlled robotic vision sensor with a zoom lens. Comput Vis Graph Image Process 59(2): 226–241
27. Vernon D, Tistarelli M (1990) Using camera motion to estimate range for robotic parts manipulation. IEEE Trans Robotics Autom 6(5): 509–521
28. Verri A, Poggio TA (1987) Against quantitative optical flow. In: Proceedings First International Conference on Computer Vision (ICCV87), London, MAy 1987, pp 171–180
29. Cahn von Seelen UM, Bajcsy R (1996) Adaptive correlation tracking of targets with changing scale. GRASP Laboratory, Technical Report MS-CIS-96-22, June 1996, University of Pennsylvania, Department of Computer and Information Science
30. Wai WYK, Tsotsos JK (1994) Directing attention to onset and offset of image events for eye–head movement control. In: Proc. of the IEEE Workshop on Visual Behaviors, Seattle, Washington, pp 79–84
31. Watt A (1989) Fundamentals of three-dimensional computer graphics. Addison Wesley
32. Wilkes D, Dickinson S, Tsotsos J (1995) A quantitative analysis of view degeneracy and its use for active focal length control. In: Proc. of the International Conference on Computer Vision, Cambridge, Massachusetts
33. Williams TD (1980) Depth from camera motion in a real world scene. IEEE Trans Pattern Anal Mach Intell 2(6): 511–516
34. Willson RG, Shafer SA (1994) What is the center of the image? J Opt Soc Am 11(11): 2946–2955
35. Yoshimura S, Kanade T (1994) Fast template matching based on the normalized correlation by using multiresolution eigenimages. In: International Conference on Intelligent Robots and Systems, Munich, Germany, pp 2086–2093
36. Zemlyak BR (1997) Direct depth from axial camera motion. Master's thesis, Technion — Israel Institute of Technology, Haifa, Israel

Jeffrey A. Fayman received a BSc degree in business administration (information systems) and an MSc degree in computer science from San Diego State University. He received a PhD degree in computer science from the Technion – Israel Institute of Technology where his research interests included various areas of active vision.

Oded Sudarsky received a BA in 1985, an MSc in 1988 and a DSc in 1998, all in computer science from the Technion – Israel Institute of Technology. His doctoral research was in the field of computer graphics and focused on the generalization of occlusion culling techniques to dynamic scenes. In 1997 he cofounded Virtue 3D, a company that commercializes 3D graphics optimization techniques. Currently he is the chief software scientist of the MediGuide company, which deals with medical imaging and the accurate guidance of medical instruments in the human body

Ehud Rivlin received his PhD in computer science from the University of Maryland, USA in 1993. Since 1999, he has been an associate professor of the Technion – Israel Institute of Technology. His research interests include robotics, active vision, object recognition, artificial intelligence, and image processing.

Michael Rudzsky received his PhD in physics and mathematics in 1980. He worked at the Scientific and Industrial Association for Space Research in Baku, Azerbajan until 1990. Since 1991 he has worked in the Physics Department and since 1995 in the Computer Science Department of the Technion – Israel Institute of Technology. His research interests are in nuclear astrophysics, computer vision, and compression of images.