Zoom Tracking and its Applications

by

Jeffrey Fayman, Oded Sudarsky, Ehud Rivlin

CIS Report #9717
December 1997
Zoom Tracking and its Applications

Jeffrey A. Fayman
Oded Sudarsky
Ehud Rivlin

Computer Science Department
Technion—Israel Institute of Technology
32000 Haifa, Israel

{jefff, sudar, ehudr}@cs.technion.ac.il

Abstract

In this paper we present a new active vision technique called zoom tracking. Zoom tracking is the continuous adjustment of a camera's focal length, 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: improving the performance of scale variant algorithms and recovery of depth information.

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 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 only slightly less accurate results.

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. In order to provide machine vision systems with similar capabilities, it is desirable to provide them 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, 18, 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, measures such as the object’s velocity and time-to-contact 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 distortion, and derive a bound on the performance of zoom tracking for image stabilization. Furthermore, we show how this distortion can be utilized to reconstruct the objects depth map.

The remainder of the paper is organized as follows. Related work is reviewed in section 2. An overview of our approach for controlling the focal length is presented in section 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 section 5. The effect of perspective distortion is analyzed in section 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. Experiments are presented in section 8, and conclusions in section 9.

2 Related Work

Since 1985, when active vision first appeared in the literature [1, 3, 4], the topic 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 parameters
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 et al. [29] present an algorithm that allows an active vision system to track targets at 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: pinhole, thin-lens, thick-lens, and geometric.

The pinhole 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, Chapter 1.4]. This geometrically simple model is widely used 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 pinhole 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 principal 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, 18] 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 such effects as optical aberrations, 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, 15, 16, 27, 33, 36]. This axial motion, i.e. the change of object-to-camera distance, can be due to 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 visible effect 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 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 concludes 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 pinhole 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, 18, 19] develop a depth reconstruction method for a static object and camera using the thick lens model. They conclude that the simpler pinhole 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 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.

2.3 Other uses of zoom

Wilkes et al. [32] use zoom to reduce the probability of view degeneracies. Degenerate views occupy a significant fraction of the viewing sphere surrounding an object. Furthermore, these degeneracies cannot be detected from a single viewpoint. Wilkes et al. choose a focal length that reduces the probability of view degeneracies, improving the performance of systems designed to recognize objects from a single arbitrary viewpoint.

Hosoda et al. [14] develop an interesting controller for robotic systems, which enhances visual servoing using zoom and a manipulator. Their controller is based on the observation that the vision system and the manipulator have complementary characteristics. Zoom can provide a wide range of image changes, but it cannot realize quick motions, and it has only one degree of freedom. On the other hand, the manipulator can move fast ("the hand is quicker than the eye"), but it cannot realize a wide range of image changes. They propose a complementary visual servoing controller of zoom and arm mechanisms.

3 Approach

Assume the zoom lens is calibrated [17, 20, 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

\(^1\)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.
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 recedes from the camera respectively. This is illustrated in figure 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 a correspondence of features such as object corners and edges between frames. 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 (AF) sensor can also be used. Autofocus cameras usually measure the distance to the object at one point in the center of the image (although some cameras use several sensors at a few points within 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.

4 Preliminaries

4.1 The Imaging Model

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

\[
\begin{align*}
\dot{X} &= -U - BZ + CY \\
\dot{Y} &= -V - CX + AZ
\end{align*}
\] (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. [18] point out, the pinhole camera model is inadequate for a zoom lens, and the thick-lens model must be used instead; however, the pinhole 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. The result is equivalent to the thick-lens model for light rays travelling through the lens’s optical axis, and is an adequate approximation if one is willing to neglect aperture effects, e.g. depth-of-field. According to Subbarao [25], the distance \(l\) is

\[
l = f_a + f_b - \frac{f_a f_b}{f}
\]

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 Figure 2.

The perspective projection is

\[
x = \frac{Xf}{Z + l - f}, \quad y = \frac{Yf}{Z + l - f}
\]

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

\[
r = \frac{Rf}{Z + l - f}
\]

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 by taking derivatives of equation 3 with respect to time:

\[
\begin{align*}
    u &= u_{\text{trans}} + u_{\text{rot}} + u_{\text{zoom}} \\
    v &= v_{\text{trans}} + v_{\text{rot}} + v_{\text{zoom}}
\end{align*}
\]

(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:

\[
\begin{align*}
    u_{\text{trans}} &= \frac{-Uf + xW}{Z + l - f} \\
    u_{\text{rot}} &= \frac{1}{f} \left( Axy - B \left( \frac{x^2 + \frac{Zf^2}{Z + l - f}}{Z + l - f} \right) + Cyf \right) \\
    u_{\text{zoom}} &= \frac{\dot{f} x}{f} \left( 1 + \frac{f^2 - f_0^2}{f(Z + l - f)} \right)
\end{align*}
\]

(6)

5.1 Optics

In this section, the algorithm derived by Horn is known as the thick lens model rather than the pinhole model. Let \(I(x, y, t)\) be the image intensity function, where \(t\) is time. The time derivative of \(I\) can be written as:

\[
\frac{dI}{dt} = \frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t} = (I_x, I_y) \cdot \vec{u} + I_t = \nabla I \cdot \vec{u} + I_t
\]

(7)

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

\[
\nabla I \cdot \vec{u} + I_t = 0.
\]

(8)

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

\[
\nabla I \cdot (\vec{u} + u_\parallel) + I_t = \nabla I \cdot u_\perp + \nabla I \cdot u_\parallel + I_t = \nabla I \cdot u_\perp + I_t = 0.
\]

(9)
Consequently

\[ u_\perp = -\frac{\partial I}{\partial t} \frac{\nabla I}{\|\nabla I\|^2}. \]  

(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 to find 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 sections 5.1 and 5.2, respectively. We conclude that the optical flow based technique is not useable 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 section 6.

5.1 Optical-flow based zoom tracking

In this section, we explore the use of the optical flow field \( \vec{u} \) from section 4 in a feedback loop algorithm which adjusts the focal length to keep the size of the object constant in the image. For the remainder of the section, \( u_{\text{rad}} \) will refer to the radial component of the vector \( \vec{u} \).

Refering to the imaging model of figure 2, equation 4 can be written in the following way:

\[ \ln r = \ln R + \ln f - \ln(Z + l - f). \]  

(11)

Holding \( r \) constant and differentiating equation 11, we get:

\[ 0 = \frac{\dot{R}}{R} - \frac{\dot{Z}}{Z + l - f} - \frac{(\frac{f\dot{f}}{f} - 1)}{Z + l - f}\hat{f} + \frac{\dot{f}}{f}, \]  

(12)

and letting

\[ \frac{u}{r} = \frac{\dot{R}}{R} - \frac{\dot{Z}}{Z + l - f} \]  

(13)

we get

\[ 0 = \frac{u}{r} - \frac{(\frac{f\dot{f}}{f} - 1)}{Z + l - f}\hat{f} + \frac{\dot{f}}{f}, \]  

(14)

which gives:

\[ \hat{f} = -f\frac{\frac{u}{r}}{1 + f\frac{1}{Z + l - f}\left(1 - \frac{f\dot{f}}{f^2}\right)} \]  

(15)

The first part of equation 15, i.e. \( \hat{f} = -fu/r \), is identical to the equation obtained for zoom tracking based on the pinhole camera model. Using the thick-lens model, the factor
Figure 3: (a) Effect of the thick-lens model on zoom tracking accuracy; (b) Valid tracking distances and focal lengths with a 10% accepted error

\[ \frac{1}{1 + f(Z + l - f) \left( 1 - f_a f_b f^2 \right)} \]

is introduced. In figure 3(a), this factor is plotted as a function of object depth and focal length in order to illustrate its effect on zoom tracking computations. The plot shows that the factor has more influence at closer distances and shorter focal lengths. While the pinhole model cannot fully capture the behavior of a zoom lens, it can be used as a reasonable approximation when the object depth Z is unknown. If one were willing to accept a 10% error in zoom tracking accuracy, objects can be tracked starting from a distance of approximately 1 meter, as shown in figure 3(b).

### 5.2 Autofocus-sensor based zoom tracking

As mentioned in section 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 is undergoing 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 figure 4. The distances \( Z_1 \) and \( Z_2 \) are given by the autofocus sensor. Let \( r_1 \) be the initial position of a point \( P \) on the object, and let \( r_2 \) 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 \( r_1 \) to \( r_2 \). This is accomplished by changing the distances between the image plane and the principal planes from 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:

\[ r = \frac{f_1 R}{Z_1 + l_1 - f_1} \]  

\[ \text{(16)} \]

The focal lengths of the two lens groups used in the thick-lens model are taken as \( f_a = -24.353 \text{ mm} \) and \( f_b = 126.592 \text{ mm} \). See section 8 for an explanation.
and

\[ r = \frac{f_2 R}{Z_2 + l_2 - f_2} \]  \hspace{1cm} (17)

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

\[ \frac{f_1 R}{Z_1 + l_1 - f_1} = \frac{f_2 R}{Z_2 + l_2 - f_2}. \]  \hspace{1cm} (18)

Using equation 2 and solving equation 18 for \(f_2\) and \(l_2\), we obtain

\[ f_2 = \frac{f_1 (f_a + f_b + Z_2) + \xi}{2(l_1 + Z_1)} \]  \hspace{1cm} (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

\[ \xi = \sqrt{f_1 (-4f_a f_b l_1 + Z_1) + f_1 (f_a + f_b + Z_2)^2}. \]  \hspace{1cm} (20)

Equation 20 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 \(f_{\text{wide}}\) is the shortest focal length and \(f_{\text{tele}}\) is the longest, equation 20 gives the depth range in 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

\[ d_{\text{min}} = (f_b f_{\text{wide}} - 1)f_a - f_b + (l_1 + Z_1)f_{\text{wide}} f_1) \]  \hspace{1cm} (21)

to a maximum distance of

\[ d_{\text{max}} = (f_b f_{\text{tele}} - 1)f_a - f_b + (l_1 + Z_1)f_{\text{tele}} f_1), \]  \hspace{1cm} (22)

where \(l_1\) is given by equation 2.
Figure 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.

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.

Figure 5 illustrates this effect. Figure 5(a) shows a frontal view of a wire-frame cube. In (b), 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 (c), 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 (a).

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 equation 4. If point $P$ is on the reference plane that is
being tracked, then its image remains constant at \( r = R f_i / (Z_1 + l_1 - f_i) = R f_2 / (Z_2 + l_2 - f_2) \). However, if \( P \) is not on the reference plane, but is at a distance \( z \) from this plane, then its image shifts from \( r_1 = \frac{R f_i}{Z_1 + z + l_1 - f_i} \) to \( r_2 = \frac{R f_2}{Z_2 + z + l_2 - f_2} \). The perspective distortion of point \( P \) is thus given by

\[
\frac{r_2 - r_1}{Z_2 + z + l_2 - f_2} = \frac{R f_1}{Z_1 + z + l_1 - f_1}
\]

(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 equation 23, the perspective distortion error \( e \) of any point on the object is bounded by

\[
e \leq R \left| \frac{f_2}{Z_2 + z + l_2 - f_2} - \frac{f_1}{Z_1 + z + l_1 - f_1} \right|
\]

(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, equation 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.

In DfZ, it is assumed that both the camera and 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 is only slightly less accurate than DfAM. However, this is compensated 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 accuracy $\varepsilon > 0$. This accuracy 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 $\mu$m (or slightly less). Sub-pixel resolution can determine correspondences with accuracy which is better, but still linearly dependent on pixel size (e.g. 1/5 pixel). Therefore, depth can also be reconstructed only up to a certain accuracy.

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. [18]. They were obtained by solving equation 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

$$Z = \frac{(r_1 - r_2)f_2l_2 + f_1l_2r_2 - f_2l_1r_1}{f_2r_1 - f_1r_2}$$

(25)

where $r_1 = f_1R/(Z + l_1 - f_1)$ is the radial distance of the point’s image at focal length $f_1$ and $r_2 = f_2R/(Z + l_2 - f_2)$ is the corresponding distance at focal length $f_2$ (see figure 2), $l_1 = f_a + f_b - f_3f_1f_3$ and $l_2 = f_a + f_b - f_3f_2f_3$. Both $r_1$ and $r_2$ are detected only up to $\varepsilon$. Therefore $Z$ can be determined up to an accuracy of

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

$$= \frac{(Z + l_1 - f_1)(Z + l_2 - f_2)(f_1Z + f_2Z + f_1l_2 + f_2l_1 - 2f_2f_2)}{Rf_1f_2(l_1 - l_2 - f_1 + f_2)} \varepsilon.$$ 

(26)

Figure 6(a) shows $\varepsilon_{DfZ}$ for $Z = 1$ 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. (This is the zoom range of the Canon VC-C1 camera used in our experiments.)

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.

Relative depth

Let $Z_a^b$ and $Z_b^a$ 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 $Z_a^1 - Z_a^2 = Z_b^1 - Z_b^2$. 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 $Z_a^1, Z_b^1$ of the two object points $a, b$:

$$\frac{Z_a^1 + l - f}{Z_b^1 + l - f} = \frac{r_2^b(r_2^a - r_1^a)}{r_2^a(r_2^b - r_1^b)}.$$  \hspace{1cm} (27)

**Absolute depth**

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

$$Z_1 = f - l - \frac{dr_2}{r_1 - r_2}$$ \hspace{1cm} (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 $\varepsilon$, so $Z_1$ can be calculated with accuracy

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

$$= \frac{(Z_1 + l - f)(Z_2 + l - f)(Z_2 + 2l - 2f)}{Rf(Z_1 - Z_2)} \varepsilon. \hspace{1cm} (29)$$

Figure 6(b) shows $\varepsilon_{\text{DIAM}}$ for $f = 12$ 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.
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.

Relative depth

Let \( Z_1, Z_2, Z_1^r, Z_2^r, r_1^r, r_2^r \) and \( r_2^r \) be defined as in relative depth from axial motion. Then

\[
\frac{Z_1^r(f_1r_2^r - f_2r_1^r) + (f_1 - l_1)r_2^r f_2}{Z_2^r(f_1r_2^r - f_2r_1^r) + (f_1 - l_1)r_2^r f_2} = \frac{r_2^r}{r_2^r} \tag{30}
\]

Absolute depth

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

\[
Z_1 = \frac{(r_1 - r_2)f_1 f_2 + f_1 l_2 r_2 - f_2 l_1 r_1 - df_2 r_2}{f_2 r_1 - f_1 r_2} \tag{31}
\]

where the radial distances of the point’s images are \( r_1 = f_1/(Z_1 + l_1 - f_1) \) and \( r_2 = f_2/(Z_2 + l_2 - f_2) \), 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 \( \varepsilon \), therefore \( Z_1 \) is calculated with accuracy

\[
\varepsilon_{DfZT} = \left| \frac{\partial Z_1}{\partial r_1} \varepsilon + \frac{\partial Z_1}{\partial r_2} \varepsilon \right| = \left| \frac{(Z_1 + l_1 - f_1)(Z_2 + l_2 - f_2)(f_1 Z_2 + f_2 Z_1 + f_1 l_2 + f_2 l_1 - 2 f_2)}{R f_2 (l_1 - l_2 - f_1 + f_2 + Z_1 - Z_2)} \right| \varepsilon. \tag{32}
\]

Figure 7(a) shows the error of DfZT. For comparison with figure 6, the error \( \varepsilon_{DfZT} \) is plotted as a function of two parameters, \( r_f \) and \( r_Z \), that 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 is smallest when the parameter is equal to 0 and greatest when it is 1. The “normal” focal length is \( f = 12 \) mm, and the “normal” distance is \( Z = 1 \) m. This yields values comparable to those used in figure 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, figure 7(b) shows a cross-section of the graph in figure 7(a) along the diagonal \( r_Z = r_f \).

7.4 Comparison

In all cases, depth can be reconstructed as a result of the change of 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 inch 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 accuracies of all three absolute depth reconstruction methods depend linearly on the accuracy \( \varepsilon \) 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 \( \varepsilon \).

Figure 8 shows the ratios of the error of DfZT to the errors of DfZ and DfAM. As in figure 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. These graphs show that, for the chosen range of object depths and focal lengths, DfZT is three times more accurate than DfZ, and only slightly less accurate than DfAM.

While DfAM is 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.
Figure 8: (a), (b) $\epsilon_DZI/\epsilon_DZI^Z$, (c), (d) $\epsilon_DZI/\epsilon_DIA$
8 Experiments

In this section, the usefulness of zoom tracking and its ability to reconstruct depth are demonstrated. In the first experiment, zoom tracking is implemented. Next, the beneficial effect that zoom tracking has on template matching, an algorithm that is scale variant, is shown. 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.

8.1 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 the intrinsic parameters of focus, zoom and iris. The lens on the Canon camera provides $8 \times$ magnification with focal length adjustments from 6–48mm. Figure 9 illustrates the VC-C1 camera.

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 which 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 non-linear quantization of its span of possible focal lengths which introduces a small oscillation when the focal length is changed.

8.2 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
Figure 10: Correspondence was searched for within a radial segment of $\theta \pm 1^\circ$

Figure 11: (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.

and Brady [24].

The corner detector returned a list of corners $L_1$ and $L_2$ for each image. Correspondence between the corners $L_{1i}$ and $L_{2j}$ 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 $(r, \theta)$ of each $L_{1i}$ and $L_{2j}$ 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 $\theta$. In practice, this is usually not the case, and we search for the matching corner using an SSD template matcher within a radial segment with an angular spread of $\pm 1^\circ$ as illustrated in figure 10. After the correspondences between $L_{1i}$ and $L_{2j}$ were found, we applied a discrete version of the proposed formula $f = -f_0/r$. The focal length was changed by the average focal length computed for each corner.

Zoom tracking was performed on the picture illustrated in figure 11(a). The picture was attached to a robotic manipulator which effected 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). Figures 11(a,b) show the picture at 900 mm and 1400 mm, respectively, when zoom tracking was not employed. Figure 11(c) shows the picture at 1400 mm when zoom tracking was used. The results of the corner detector are displayed in the top left corner of figure 11(c).
8.3 Zoom tracking and template matching

In this experiment, we show the effects of zoom tracking on template matching, a scale variant algorithm. A standard sum of squared differences (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 which 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, which computes the similarity of patches $P$ in an image $I(x,y)$ to a template $T(x,y)$, is defined as follows:

$$E(i,j) = \sum_{x,y \in P} (T(x,y) - I(x - i, y - j))^2. \quad (33)$$

Template matching performance with/without zoom tracking

The performance of the sum of squared difference tracker was measured as it tracked a moving picture of a light bulb. The picture was again attached to the robotic manipulator which effected translation of 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 illustrated in figure 12.

Figure 13(a) presents the results of the tracker when no focal length adjustments were made. Figure 13(b) 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 sum of squared difference tracker improves, as shown in figure 13(b). (The spikes in both graphs are due to the autofocus mechanism and to the quantized focal length of the Canon camera.)

8.4 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 pinhole camera model. In order to use the thick-lens model, the viewpoints given to OpenGL were virtually translated along the optical axis according to equation 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 figure 14(a).

To verify the correctness of equation 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 figure 14(b).

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 (figure 14(c)) show that the formula for \(\varepsilon_{DfZT}\) 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 figure 15(a). The reconstructed
Figure 14: Simulated zoom tracking: (a) visual results; (b) reconstructed and real depths; (c) predicted and real depth errors vs. real depths of the matched corners are shown in figure 15(b), and the depth reconstruction errors vs. the predicted upper bounds on these errors in 15(c). The correspondence error for figure 15(c) was taken as $\varepsilon = 30 \, \mu \text{m}$, which is the size of each pixel in a $500 \times 500$ pixel image simulating a 15 mm square image sensor. Note that the corner matcher only found 6 pairs of the 7 visible matching corners.

9 Conclusions

Zoom tracking exploits the adjustable focal length parameter in an active vision system to compensate the scale changes of a target moving along a camera’s optical axis. We presented two methods by which zoom tracking can be performed: a optical flow based feedback algorithm that measures the 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 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 only slightly less accurate than depth from axial motion, but useful over a greater range of object depths.

In our experiments we implemented zoom tracking, and demonstrated that it improves the performance of template matching, a scale-variant algorithm, when it is used on objects whose motion is not parallel to the image plane and hence transformed by scale. Additionally, we demonstrated the depth reconstruction capabilities provided by zoom tracking.

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.
Figure 15: Zoom tracking with synthetic image and corner matching: (a) detected corners; (b) reconstructed and real depths; (c) upper bounds and real depth errors.

Acknowledgement

The authors would like to acknowledge the assistance of Benny Yashinovsky and Michael Rudzsky for their contributions to this paper.

References


