Zoom Techniques for Achieving Scale Invariant Object Tracking in Real-Time Active Vision Systems

Eric D. Nelson and Juan C. Cockburn

Rochester Institute of Technology, 83 Lomb Memorial Drive, Rochester, NY, USA

ABSTRACT

In a manually operated visual tracking system, a camera operator follows an object of interest by moving the camera, then gains additional details about the object by zooming in. As the active vision field progresses, the ability to automate such a system is nearing fruition. One hurdle limiting the deployment of real-time visual tracking systems is in the object recognition algorithms that often have restrictive scale and pose requirements. If those conditions are not met, the performance of the system rapidly degrades to failure. The ability of an automatic fixation system to capture quality video of a non cooperative moving target is strongly related to the response time of the mechanical pan, tilt, and zoom platform. However, the price of such a platform rises with its performance. The goal of this work is to investigate the feasibility and issues that arise when using inexpensive off-the-shelf components in the development of a visual tracking system that provides scale-invariant tracking. One of the main challenges is in the zooming action. Optical zoom acts as a measurement gain, amplifying both resolution and tracking error. Previous work has shown that adding a second camera with fixed focal length can assist the zooming camera if it loses fixation, effectively bounding the error. Furthermore, optical zoom has a longer time–constant than digital zoom. This work proposes a dual camera hybrid zoom configuration where digital zoom is combined with optical zoom to achieve a behavior closer to an ideal zooming action.

Keywords: Active vision, Object tracking, Optical zoom, Digital zoom, Hybrid zoom, Real time systems

1. INTRODUCTION

In the early 1960s, researchers viewed computer vision as a relatively simple problem—if humans and multitudes of other organisms can so effortlessly see, then how difficult can it be to design a man-made system with similar attributes? The perception was that it would be only a few decades before we were able to surpass the capabilities of a natural vision system. Nearly half a century later research has revealed that human vision is considerably more complex than imagined, not to say that advances have not been made. It is fair to say that the human eye is similar in function to a digital camera, inasmuch as they both capture an image for further processing by either the brain or a computer. It would diminish the significance of this work to say that the sensing problem has been solved. However, it is the interpretation of the captured image—the task performed by the brain—that continues to frustrate researchers. If and when we figure out the perception problem, potential applications include an automated cameramen for sporting events, or—for these days of heightened security awareness—a system that can recognize known terrorists in a crowd. This work was driven by our interest in automatic recognition and interpretation of sign language, specifically American Sign Language (ASL). Used as the primary mode of communication for millions worldwide, ASL is a language that communicates not only words but also the full gamut of human emotions, similarly to how inflexions in the spoken word can convey joy, sarcasm, or indifference. Therefore, to understand ASL as a human would, a computer vision system must look at the hands, face, and body language of the signer. The starting point of this daunting task is to interpret single hand gestures in real-time. We have already some encouraging results on tracking hands[1] and recognizing basic letter signs.[2]

A key feature of a robust visual tracking system is its ability to recognize and track objects of different sizes, poses, and degrees of occlusion. Most existing object recognition algorithms amenable to real-time implementation often have strict scale and pose requirements where if the requirements are not met, the performance

Further information about the authors:
J.C.C.: E-mail: Juan.Cockburn@rit.edu
E.D.N.: E-mail: EricNelsonCE@gmail.com

Acquisition, Tracking, Pointing, and Laser Systems Technologies XXI, edited by Steven L. Chodos, William E. Thompson,
Proc. of SPIE Vol. 6569, 65690F, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.720634

rapidly degrades to failure. For example, the work of Yang *et al.* requires a 60×60 pixels image[3] while the work of Heisele *et al.* requires 58×58.[4] Other algorithms are more robust to scale and pose variation, but still have a preferred size, such as Rupe's work which performs well between 50 pixels $\times$ 50 pixels and 250 pixels $\times$ 250 pixels, but works best with smaller images in that range.[2] The main goal of this work is to develop a system that can reliably capture images of a specific size and position determined by a higher-level process.

2. STABLE VISUAL TRACKING

A key component of our approach to scale invariance is the coordination of zoom and pan/tilt motion for target tracking. It is assumed that there is a robust segmentation algorithm that captures the object of interest. The focus of this work is in achieving scale invariance by simultaneously controlling zoom and pan/tilt position.

2.1. Optical zoom, tracking and stability

Proper use of *optical zoom* can improve the perceptibility of an acquired image by either increasing the details (resolution) or broadening the field of view (context). Magnification can be defined as the ratio of the height of the image to the height of the object. Since increasing magnification reduces the field of view, both improvements cannot occur simultaneously. A trade-off between these competing objectives must be optimized. This problem was solved for a single-camera system, by maximizing resolution while bounding fixation error.[5]

In any practical visual tracking system it is likely that the target will be lost due to mechanical limitations of the camera system such as slew rates or hard limits on pan/tilt motion, external disturbances such as illumination changes, target occlusion, camera ego motion, or due to lack of robustness of the segmentation algorithm. We will say that a tracking system is operating in an *unstable mode* when the target is lost, or equivalently, when fixation is lost. From a practical point of view it is important to design a system that can recover from instability. We will call a tracking system *practically stable* is the time spent in unstable mode is bounded. To design practically stable visual tracking system we will capitalize on known strategies commonly used by human camera operators. Tordoff elaborates on this saying "consider a camera operator viewing a stationary object (it might be a golf ball on the fairway, or a gnu on the veld). While stationary, the operator's instinct is to zoom in. However, as soon as the object starts to move, the cameraman will react both by attempting to track and by zooming out. As tracking is restored, say at constant angular velocity, the operator may have sufficient confidence to zoom in again, but if subsequently the tracked object moves unexpectedly he will surely once more zoom out. It appears that the camera operator is reducing tracking error to an acceptable distance in the image, where 'acceptable' means better than half the image dimension—at worst he wishes to retain sight of the object on the image plane."[5] Now consider that the camera operator does indeed lose sight of the object. Recognizing that, he pulls his eyes away from the camera, finds the object, and then adjusts the camera pose and zoom accordingly. This is the desired behavior that our visual tracking system will try to emulate.

2.2. Dual camera visual tracking

One way to implement a practically stable a visual tracking system is to use two cameras: one for fixation on the target and one for stabilization. If in addition the fixating camera has zoom control, is also possible to achieve scale invariance. This will be our visual tracking approach.[6] In this work the fixating camera will be referred as *zooming camera* and the second camera used for stabilization as *panoramic camera.* The design of a dual camera system requires a careful coordination of the two cameras. The fixation system needs information about how the target should "look" before a determination of the quality of fixation could be asserted. This information is provided by a second camera: the panoramic camera. The dual-camera system must be designed so that 1) visual tracking is practically stable and 2) scale invariance with certain tolerance is maintained. The control and coordination of the cameras is handled by a multi-mode feedback control system were a *control arbitrator* switches between *assisted control* and *autonomous control* mode. Assisted control occurs when one camera is controlled using another camera's sensors and autonomous control when a camera has enough information to control itself. When a camera loses the target, *i.e.* fixation, it must switch to assisted control to prevent system instability. The control arbitrator must decide when to switch, that is, which type of control is appropriate. Experiments showed that adding a second camera with fixed focal length produced a system with improved practical stability comparable to a system without zoom.[6] In summary, the use of a panoramic camera takes cares of the practical stability of the tracking system.

3. HYBRID ZOOM

The long time constants of typical zoom lenses have proven to be a hindrance not only to the stability of the system, but also to scale invariant tracking. Our approach to scale invariance is based on controlling the zoom to render the apparent size of the object in the image constant. At the same time a tracking feedback loop is used to control the camera pan/tilt position to keep the target close to the image center at all times.

A zoom lens must follow the motion of the target towards and away from the camera to maintain image size; so a slow-moving zoom lens is limited to all but slow-moving objects in directions orthogonal to the image plane. Instead of using only optical zoom to change the scale we could use *digital zoom*—a concept integral to digital imaging—which magnifies the digital image via pixel interpolation and decimation. Digital zoom is typically significantly faster than optical zoom, but however lacks the ability of optical zoom to adjust image detail and context. The basic idea is to combine the strengths of digital and optical zoom to offset their individual weaknesses.

With visual tracking stability no longer a concern, we can focus on optimizing zoom performance to adjust the size and resolution of the target under tracking. As mentioned above, the slow zoom response of optical zoom imposes severe limitations on the ability to produce scale invariant target images in real-time. To speed up zoom action without changing the mechanical design of the optical zoom hardware, we propose to use digital zoom on top of optical zoom. Digital zoom is used to maintain scale invariance while optical zoom is used to gain resolution and details. This new proposed zoom scheme is called *hybrid zoom*.

4. ACHIEVING SCALE INVARIANCE

An object under tracking is *scale invariant* when its apparent size in the image remains constant. As mentioned earlier digital zooming will be exploited to maintain the scale of the moving target as constant as possible.

4.1. Choosing zoom magnification

The magnification necessary to provide exact invariance can in principle be obtained by instantaneously changing zoom as the object moves in the scene. To maintain constant the size of the target image, it is necessary that the magnification, m, be proportional to the distance from the object to the camera, z_0. In a (passive) visual tracking system, this distance cannot be measured directly; therefore, it must be estimated from the image itself. A naive approach is to use the approximate *area* of the object to estimate z_0. The rationale behind this strategy is that if the area of the object appears to be constant then the object is probably not getting closer or moving away from the camera. While this estimate can handle effectively translations of a moving object in any direction, it can not always handle rotations properly. Indeed, as the *spinning coin problem* shows, this estimate will cause unwanted zoom oscillations, *i.e.* zoom instabilities. As illustrated in Figure 1, the area of the coin becomes smaller when only the edge of the coin is visible, this causes the zoom to increase, and vice versa. Clearly, in this case, the zoom should not change at all. A solution to this problem, that applies only to planar objects, was proposed by Tordoff.[7]

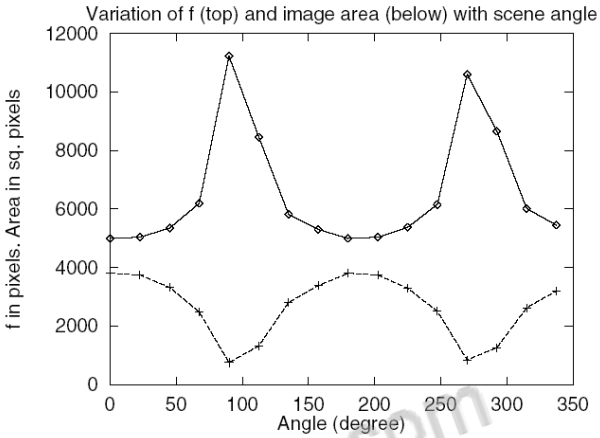
In general, the problem of designing a robust estimator for the distance from the object to the camera is crucial for a successful scale invariant visual tracking system. An alternative approach to magnification estimation that avoids unwanted zoom oscillations is to consider not the area of the object but a robust estimate of the measure of its length and width taken separately. If height (width) alone is used to estimate z_0, then zoom oscillations do not occur when the width (height) oscillates. Therefore both height and width must be considered.

One possible estimate of the magnification factor based on height and width is

$$m_k = \frac{h'_0}{\max\left(h'_k, w'_k\right)} \tag{1}$$

where h'_0 is the desired image dimension, h'_k is the measured image height (in pixels), and w'_k is the measured image width (in pixels) at frame (*i.e.* time) k. Note that the denominator is nothing more than the ℓ_1-norm of the feature vector describing the size of the bounding box of the object at frame k. Experiments showed that this choice of magnification improves the robustness of the estimate.

(a) Spinning coin. http://www.startcreative.co.uk/images/Coin rotate.jpg



(b) Adjusting focal length with image area[7]

Figure 1. Spinning coin problem

4.2. Measurement noise

Any practical measurements inherently have noise. Therefore, it is necessary to model the effects of noise on the measurements. Here a standard linear model with additive noise will be used. The measurement equation for the height of an object in an image is

$$h'_{m,k} = h'_k + \mathrm{r}_k \tag{2}$$

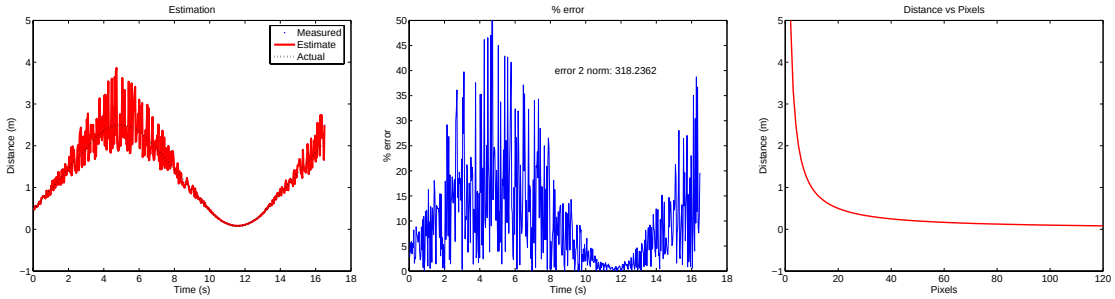
where $h'_{m,k}$ is the measured height, h'_k is the "actual" height, and r_k is a random process representing the effects of noise. For simplicity r_k is chosen as a normally distributed, zero mean random process with covariance $E\left[\mathrm{r}_k \mathrm{r}_k^\top\right] = \mathtt{R}_k$. A common measure of the amount of noise in a signal is the signal–to–noise ratio (SNR) which in this case is

$$\mathrm{SNR} = \frac{h'_k}{\mathrm{r}_k} \tag{3}$$

A large signal-to-noise ratio (SNR), *i.e.* $h'_k \gg \mathrm{r}_k$, means that noise effects are negligible. Note that the effect of image noise varies with resolution; large magnifications are more susceptible to noise. This manifests itself as a shaky zoom. To decrease this shakiness, or more formally, the sensitivity of magnification to noise we could design an optimal filter to smooth out the measurement estimates. In this case, magnification can be calculated as

$$m = \frac{h'_0}{\hat{h}'_k} \tag{4}$$

where the hat superscript denotes filtered estimate, that is, $\hat{h}'_k$ is the estimated image height. When the SNR is large, the raw measurements can be trusted more than when SNR is small. In the later case, the measurements should be filtered appropriately. This is a typical optimal filtering problem that can be solved, for example,



(a) Estimation (b) Percent error (c) Distance vs Pixels

Figure 2. Unfiltered digital zoom

using a Kalman filter.[5, 8, 9] This filter is designed to achieve the best possible tradeoff between clean and noisy measurement, or more precisely the dynamic variations in the SNR.

The height measurement depends not only on noise, but also upon the "state" of the system such that

$$h'_k = f_k \frac{h_k}{z_k} \tag{5}$$

where f_k is the camera focal length, z_k is the orthogonal distance from the image plane to the target object. This equation is obtained from the perspective equations based on the thin lens assumption, and since $z_k \gg z'_k$, $z'_k = f_k$.

The parameter $\mathtt{Q}$ was chosen as

$$\mathtt{Q} = 10^{-6} \times \begin{pmatrix} \frac{\Delta T^3}{3} & \frac{\Delta T^2}{2} \\ \frac{\Delta T^2}{2} & \Delta T \end{pmatrix}$$

where ΔT is the sampling time (between frames).

4.3. Simulations and filter tuning

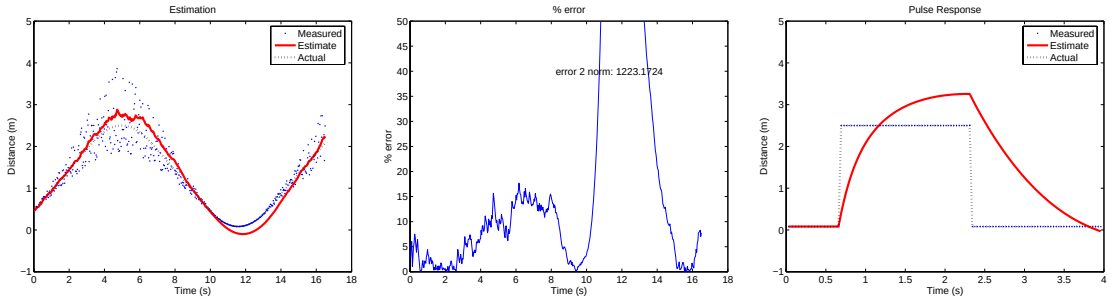
To study the effect of filter parameters at different noise levels several simulation experiments were conducted. Filters with different parameters were fed identical object motion and measurement noise and compared. The input to the filter simulates an object of height 0.35 m that located at $z = 0.5$ m that starts moving away from the camera with velocity $\dot{z} = 0.5$ m/s. Upon reaching $z = 2$ m, constant acceleration is applied to the moving object until $\dot{z} = -0.5$ m/s. The object continues to move until reaching $z = .5$ m, where it accelerates once again to $\dot{z} = 0.5$ m/s and remains at that velocity through the end of the simulation. The maximum distance is $z = 2.4$ m and the minimum is $z = 0.1$ m, which corresponds to 4 and 120 pixels, respectively, as shown in Figure 2(c). The focal length throughout simulation is constant and equal to $f = 35$ mm. The unfiltered run is shown in Figure 2(a). Notice that the noise escalates for large z.

Figure 3 shows a Kalman filter with $\mathtt{R} = 10^{-4}$. This filter relies substantially on the model, which makes it resistant to the high-noise measurements: a desirable behavior. However, when noise is low, an unwanted delay appears between actual state and estimation.

Figure 4 on the other hand has a smaller $\mathtt{R} = 10^{-8}$. It does not rely heavily upon the model for the simulated distances, which leads to less resistance to noise.

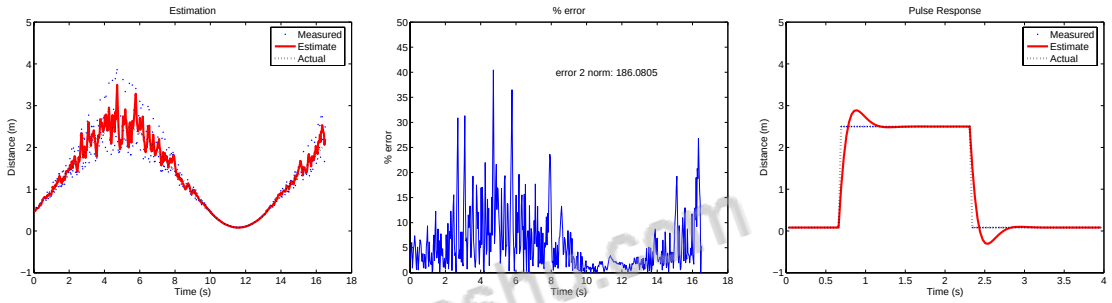
A filter which adapts itself to changing conditions is shown in Figure 5, where $\mathtt{R}_k = \frac{\mathtt{R}}{{h'_m}^2}$ and $\mathtt{R} = 10^{-8}$. The result is a filter that is resistant to high noise without adversely affecting the response when noise is low.

The results are organized in Table 1. The filter best suited for digital zooming is the one which considers varying measurement noise.



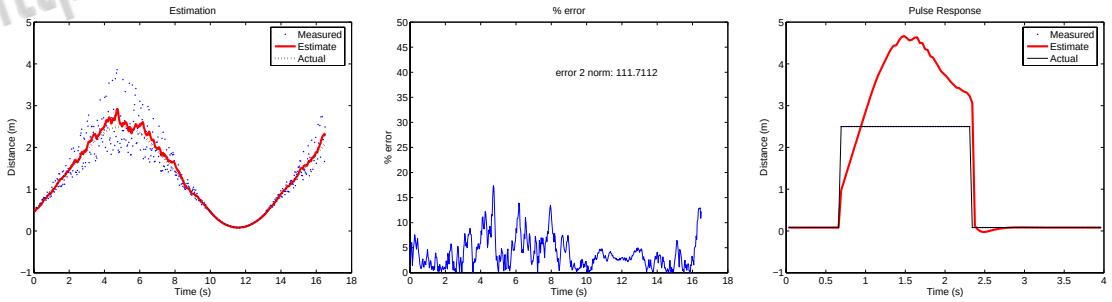
(a) Estimation (b) Percent error (c) Pulse response

Figure 3. Kalman filter of digital zoom with $\mathtt{R}_k = 10^{-4}$.



(a) Estimation (b) Percent error (c) Pulse response

Figure 4. Kalman filter of digital zoom with $\mathtt{R}_k = 10^{-8}$.



(a) Estimation (b) Percent error (c) Pulse response

Figure 5. Kalman filter of digital zoom with $\mathtt{R}_k = \frac{\mathtt{R}}{h'^{\,2}_m}$.

Table 1. Filter comparison in presence of image noise

Filter	Error (approximate 2-norm)
None	318.2
High image noise covariance	1223.2
Low image noise covariance	186.1
Variable image noise covariance	111.7

4.4. Fixating with digital zoom

Digital zooming not only allows for rapid scale changes, but also for a "movable image center" which can be adjusted to match the tracker's position measurement. This is related to the location of the center of expansion during zoom. For optical zooming, the center of expansion is fixed and often very close to the principal point of the lens which is at the intersection of the optical axis and the image plane. In an ideal case the pan/tilt motion of the camera will adjust the system so that the center of expansion is, for example, at the center of the object under tracking but this may require very accurate and fast pointing control of the camera. With digital zoom, the center of expansion can be placed anywhere on the image plane. This opens the possibility of eliminating fixation error much faster than using pan/tilt control. For more details on the reduction of fixation errors see.[10, 11]

5. THE DUAL-CAMERA HYBRID ZOOM SYSTEM

5.1. Display arbitration and mode switching

To determine when the object is in view of the zooming camera a new concept dubbed *display arbitration* is needed. Like most vision tasks, deciding which camera has captured the best image is an action naturally suited to humans, as the Cartesian Theater holds assumptions about the scene.[12] However, given the state of modern computer perception, particularly real-time perception, a single-camera has no notion of how an object *should* look; deciphering between target deformities, rotations, and fixation losses is not obvious.

In a dual-camera system, a comparison can be made between the cameras' images—if measurements do not make sense, then fixation must be lost. An assumption that camera P always has the object in view is made for simplicity; camera P is more stable, so if it loses fixation then the system is unstable.

The target object is considered in view of the zooming camera when the following conditions hold true

$$h'_P - \Delta \leq h'_Z \frac{f_P}{f_Z} \leq h'_P + \Delta \tag{6}$$

$$w'_P - \Delta \leq w'_Z \frac{f_P}{f_Z} \leq w'_P + \Delta \tag{7}$$

where Δ is the uncertainty of the tracker's results, in pixels; w' is the measured width; h' is the measured length; f is the focal length; and the subscripts Z and P denote the zooming and panoramic cameras, respectively. If both of these hold true, then the *display arbitrator* will decide that camera Z has captured the best image.

In the event of object loss on the zooming camera, the system will switch to the panoramic camera's view. Once the zooming camera regains fixation, the arbiter can return the view to the zooming camera. In the event that full scene context is preferred over high resolution—as digitally zooming out leaves an undefined region where context would normally be—the arbitration function can be altered appropriately.

6. EXPERIMENTAL RESULTS

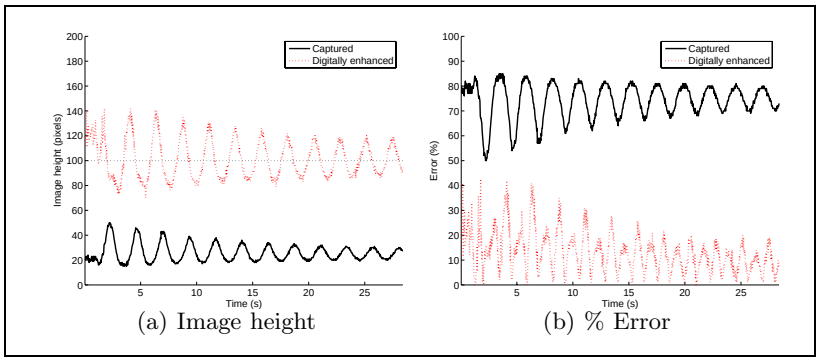
6.1. Pendulum experiment

The first experiment combining control and display arbitration is a pendulum with a red ball as its bob swinging towards and away from the cameras to test the speed of zoom. The parameters used in the test include $\Delta =$ 3 pixels, r = 0.01 (3 pixels / 320 pixels), $\gamma = 0.1$, $d = 13$ cm, $\Delta T = 33$ ms, $t_{max} = 1$ s, $\alpha = 1$ and the filter matrices

$$\begin{aligned} \mathtt{R} &= \mathrm{r}^2 = 10^{-4} \\ \mathtt{Q} &= 10^{-6} \begin{pmatrix} \frac{\Delta T^3}{3} & \frac{\Delta T^2}{2} \\ \frac{\Delta T^2}{2} & \Delta T \end{pmatrix} \end{aligned}$$

Note that the data in the following figures was generated by post-processing the captured video using color segmentation, therefore the effects of the filter may appear as inaccuracies. A better way to acquire the data would have been to use a more robust algorithm, which was not available at the time.

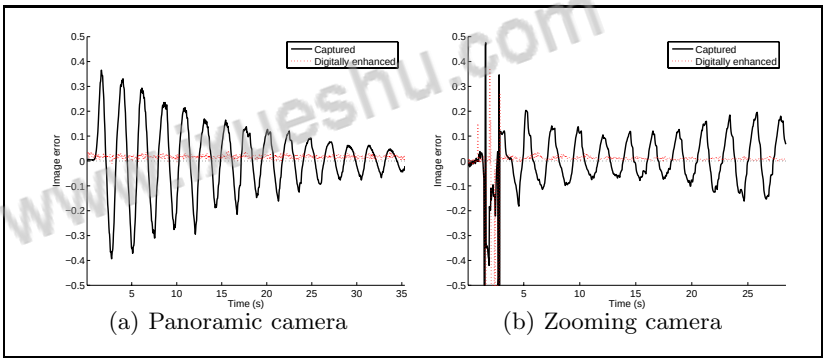
The effectiveness of digital zoom can be seen by comparing object size in the image before and after digital zoom, as shown in Figure 6. The solid curve represents image height of the original image, the dotted curve is height for the zoomed image, and the dotted horizontal line at 100 pixels represents the goal. Error is considerably decreased in the digitally zoomed image.



(a) Image height (b) % Error

Figure 6. Scale invariance for panoramic camera (pendulum)

Figure 7(a) compares the fixation error, which exhibits the image re-centering property of digital zoom. Digital zoom appears to entirely eliminate fixation error.



(a) Panoramic camera (b) Zooming camera

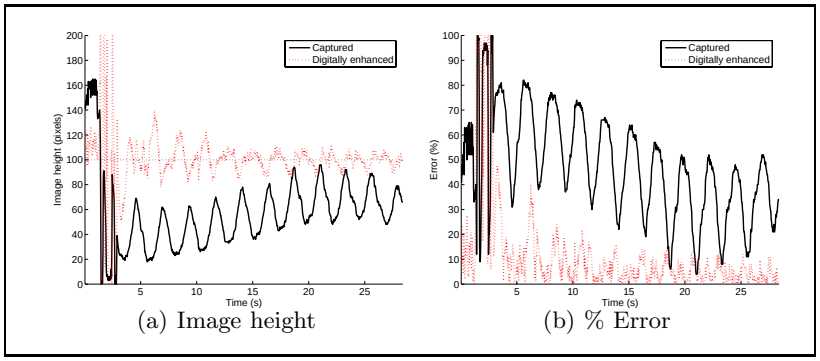
Figure 7. Digital fixation error for pendulum experiment

The combination of optical and digital zoom into hybrid zoom is shown in Figure 8. The optical zoom is closer to the target than the fixed focal length camera, but digital zoom is able to further reduce error. Note also, in Figure 7(b), that this method is able to neutralize the fixation error inherent with large focal lengths. Selected frames of one of the experimental runs are shown in Figure 12. Digital zoom is always able to fixate on the ball in the panoramic camera. However for the zooming camera, digital zoom cannot improve fixation when the ball is not in the view, particularly in frame 90 and 211. The dark border around the zooming camera's digital images in frames 30, 90, 211, 272, and 548 indicates a digital zoom out, as digital zooming is unable to recover scene context not available in the original image.

6.2. Car experiment

A second experiment involved a radio-controlled car moving sporadically. Once again, error in scale invariance, Figure 9, is considerably decreased by digital zoom, as is fixation error in Figure 10(a). Hybrid zoom is shown in Figure 11 and Figure 10(b).

Figure 13 shows selected screenshots. The digitally zoomed images appear identical between the panoramic and zooming cameras, with the exception that the zooming camera's have more detail. Frame 362 highlights this: the key feature of hybrid zoom.



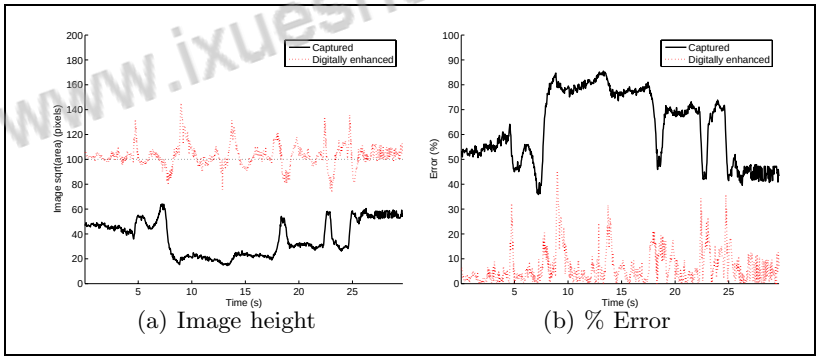
(a) Image height (b) % Error

Figure 8. Scale invariance for zooming camera (pendulum)

7. CONCLUSIONS

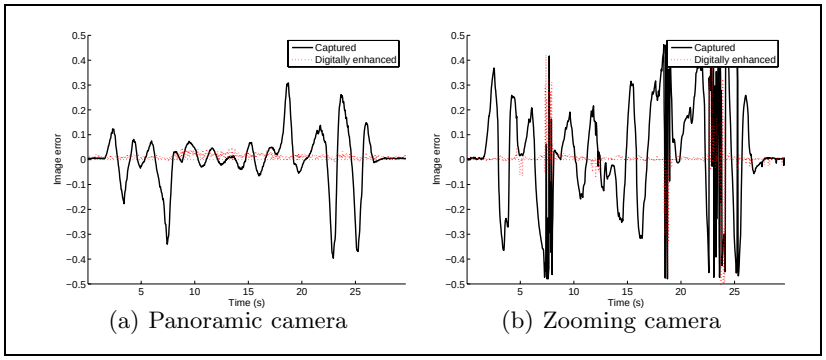
The goal of this work was to take advantage of both the high speed of digital zoom and the ability of optical zoom to alter image detail and context.

When digitally zooming, the effect of image noise increases with digital magnification. It was shown that by adjusting the static measurement noise covariance of a Kalman Filter, the effect of noise could be attenuated considerably. However, those adjustments cause the performance of digital zoom to degrade in low noise situations. This unwanted side effect was removed by allowing the measurement noise covariance to change dynamically, so that it would scale with the measured image height. The resulting filter decreased the impact of image noise



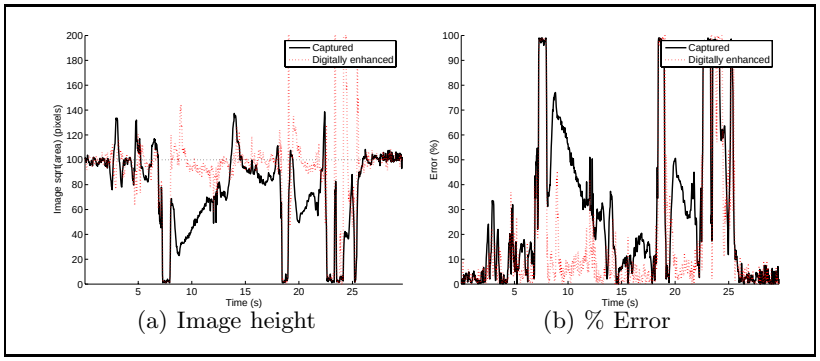
(a) Image height (b) % Error

Figure 9. Scale invariance for panoramic camera (car)



(a) Panoramic camera (b) Zooming camera

Figure 10. Fixation errors on car experiment



(a) Image height (b) % Error

Figure 11. Scale invariance of zooming camera (car)

without noticeably degrading performance.

This work showed experimentally that adding digital zoom to an active vision system decreases fixation error while improving scale invariance.

In a hybrid-zooming system, camera focal length no longer affects image size. When also coupled with a dual-camera system, the focal length does not affect system stability. Therefore, focal length control algorithms can be developed to maximize image detail: a key to image recognition.

REFERENCES

1. E. Clark, "A multicamera system for gesture tracking with three dimensional hand pose estimation," Master's thesis, Rochester Institute of Technology, 2006.
2. J. C. Rupe, "Vision-based hand shape identification for sign language recognition," Master's thesis, Rochester Institute of Technology, 2005.
3. M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(1), pp. 34–58, 2002.
4. B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," in *cvpr*, pp. 688–694, 2001.
5. B. J. Tordoff, *Active Control of Zoom for Computer Vision*. PhD thesis, University of Oxford, 2002.
6. E. D. Nelson, "Zoom techniques for achieving scale invariant object tracking in real-time active vision systems," Master's thesis, Rochester Institute of Technology, 2006.
7. B. J. Tordoff and D. W. Murray, "Reactive zoom control while tracking," Tech. Rep. OUEL 2228/00, University of Oxford, 2000.
8. Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, Academic Press, Inc., San Diego, CA, USA, 1988.
9. R. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering* **82**, pp. 35–45, 1960.
10. K. Daniilidis, C. Krauss, M. Hansen, and G. Sommer, "Real-time tracking of moving objects with an active camera," *Journal of Real Time Imaging* **1**, pp. 3–20, 1998.
11. R. Collins, O. Amidi, and T. Kanade, "An active camera system for acquiring multi-view video," in *Proceedings of the 2002 International Conference on Image Processing (ICIP '02)*, pp. 517–520, September 2002.
12. D. C. Dennett, *Freedom Evolves*, Viking Penguin, New York, NY, USA, 2003.

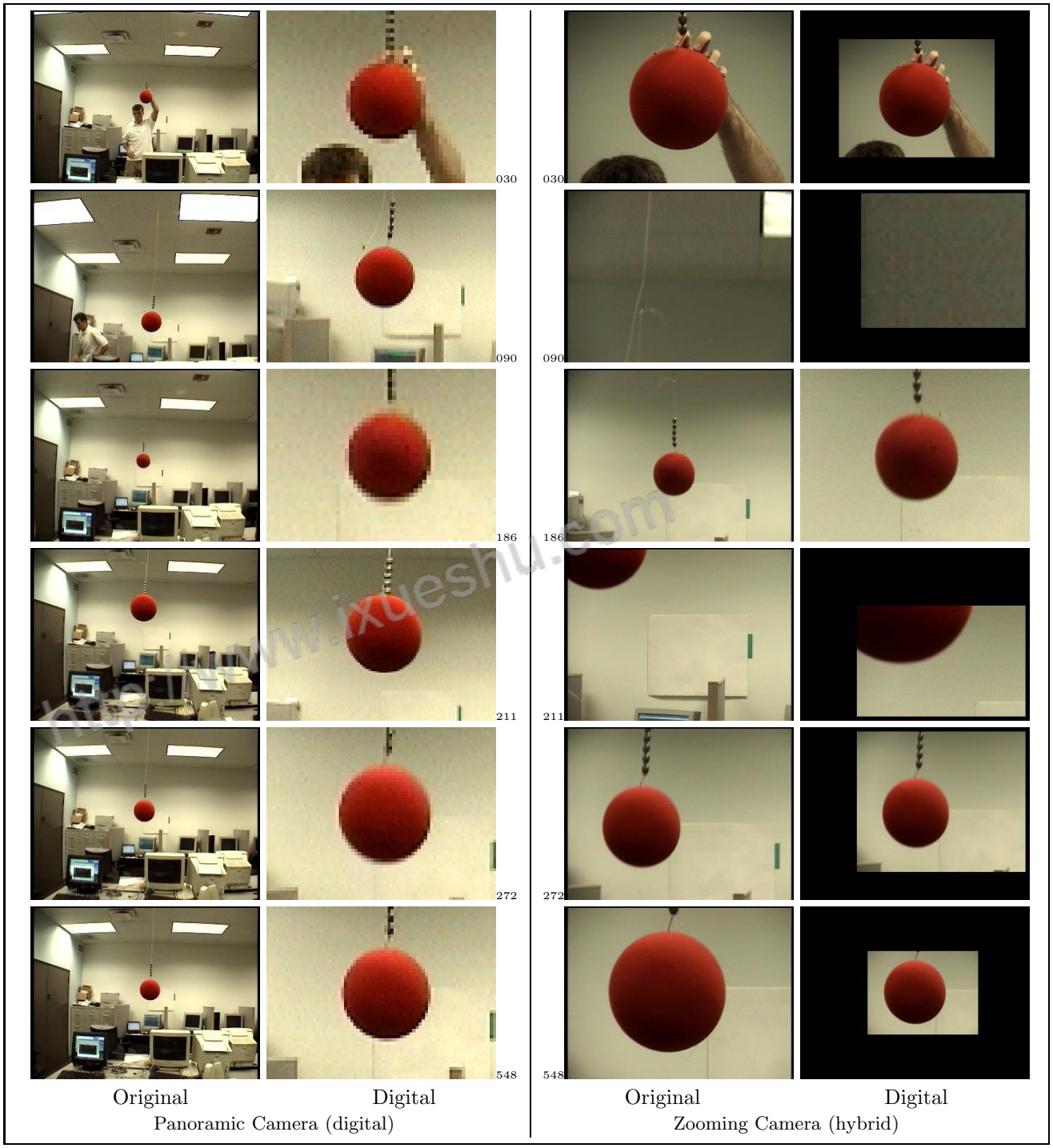


Figure 12. Display arbitration experiment (pendulum)

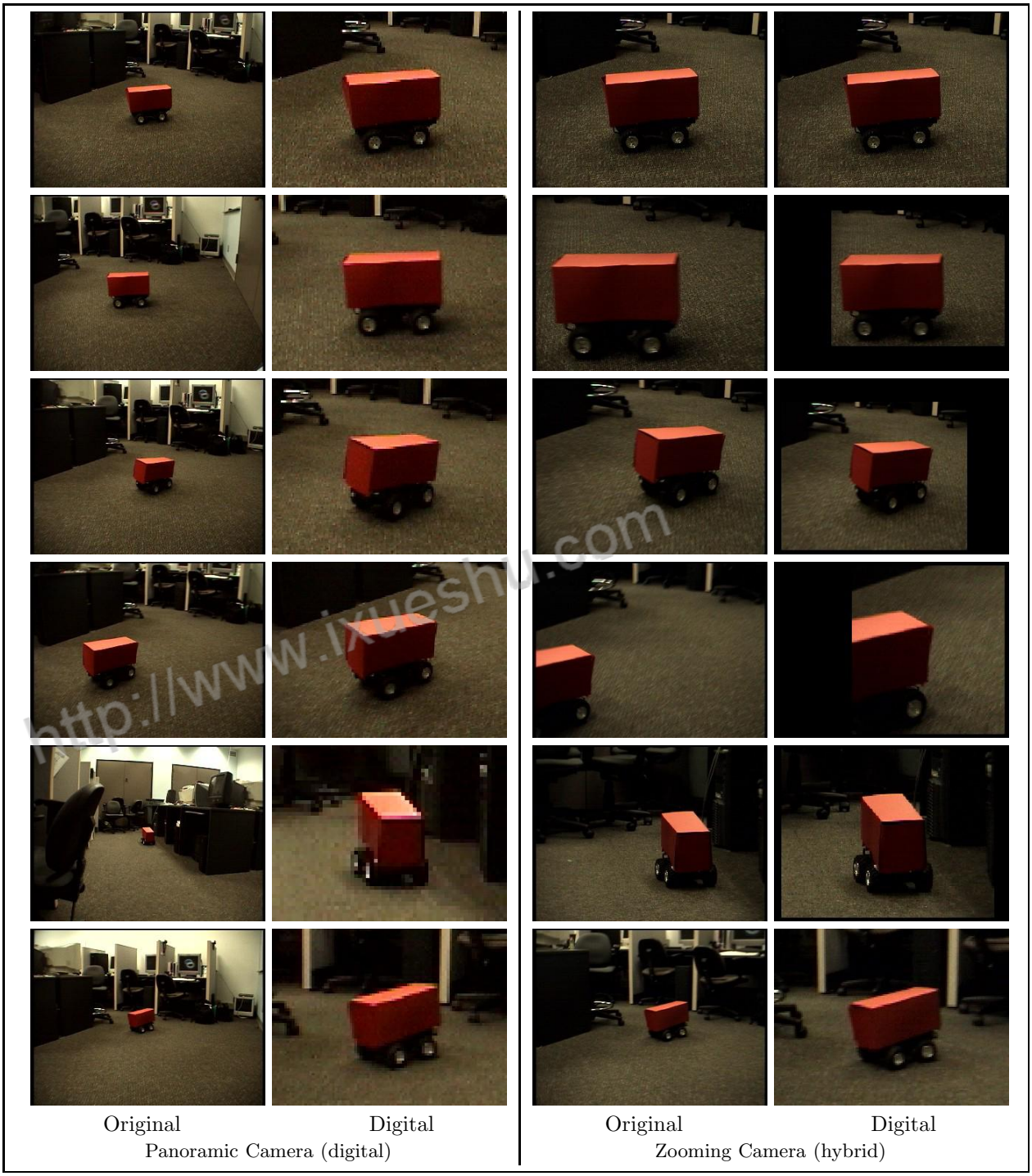


Figure 13. Display arbitration experiment (car)