DUAL CAMERA ZOOM CONTROL: A STUDY OF ZOOM TRACKING STABILITY

Eric D. Nelson and Dr. Juan C. Cockburn

Rochester Institute of Technology
Department of Computer Engineering
Rochester, NY

ABSTRACT

In a surveillance system, a camera operator follows an object of interest by moving the camera, then gains additional information about the object by zooming. As the active vision field advances, the ability to automate such a system is nearing fruition. One hurdle limiting the use of object recognition algorithms in real-time systems is the quality of captured imagery; recognition algorithms often have strict scale and position requirements where if those parameters are not met, the performance rapidly degrades to failure. The goal of this work is to create a system that provides scale-invariant tracking using inexpensive off-the-shelf components.

***Index Terms*—** Active vision, Object tracking, Optical 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 mere decades before we were able to surpass the capabilities of a natural vision system—but nearly half a century later, research has indicated 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 processing by either the brain or a computer. It would diminish the significance of this work to say the sensing problem has been solved, however it is the interpretation of the captured image that continues to frustrate researchers.

A particular application of interest is the ability to recognize and interpret 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 must look at the hands, face, and body language of the signer.

The starting point of this daunting task is to interpret a single hand in real-time. Preceding works present ways to track hands in real-time, such as Clark in [1], and to recognize basic signs, such as Rupe in [2]. The missing bridge between the works is that the captured image must be conditioned appropriately for the recognition algorithm to be effective. For example, the work of Yang *et al.* in [3] requires a 60 pixels $\times$ 60 pixels image. Other algorithms are more resistant to varying scale and position, 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.

The 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 (*e.g.* an object recognition algorithm).

2. OPTICAL ZOOM

The presence of zoom can improve the perceptibility of an acquired image by either magnifying detail or broadening the field of view. However, since higher focal lengths confine the periphery, both improvements cannot occur simultaneously. This compromise was addressed by maximizing resolution while bounding fixation error [4]. In the event of lost fixation—at least when there is no plan to reacquire fixation, such as in this work—an active vision system becomes unstable; observing a scene, or a specific object in a scene, is the *raison d'être* of the system, therefore if the object is out of view, the system can do nothing. Tordoff elaborates in [4] 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."

3. DUAL-CAMERA OPTICAL ZOOM

Now consider that the camera operator does indeed lose sight of the object. Recognizing that, he pulls his eye away from the camera, finds the object, and then adjusts the camera accordingly. In this case, the naked eye serves as a fixed focal length, panoramic camera. This work expands on Tordoff's work by introducing a second camera that serves as the naked eye, allowing stability to be independent of zoom action.

Similar methods are presented in the works of Greiffenhangen *et al.* in [5] and Huang *et al.* in [6]. They differ from this work in that their panoramic camera serves as a stationary overseer—it is mounted on the ceiling. In this work, both cameras rotate from nearly the same viewpoint.

When fixating with a single camera, the desired camera angles—θ and ϕ—place T' at C'. When T' is on Π', then a tracker can be used to calculate the angles, called *autonomous control*. However if T' is off the edge of Π', then *assisted control* is necessary, which requires a second camera.

Consider frame k where T' is on Π' for both cameras, then $z_{0,k}$ is the distance of the image planes to the fronto-parallel plane

$$z_{0,k} = \frac{d}{\cot(\theta_{P,k}) - \cot(\theta_{Z,k})} \tag{1}$$

Now say for some future frame $n > k$ that T'_Z is off image plane Π'_Z, but T'_P remains on Π'_P. Then θ_P and ϕ_P can be computed using assisted control. Tilt angles can be locked, $\phi_Z = \phi_P$, while θ_z can be found using Equation 7 if z_n is known. Using the basic lens equation where h is the height of the object,

$$\frac{z'}{z} = \frac{h'}{h} \tag{2}$$

then for frame n (and similarly for k)

$$z_n = \frac{h_n}{h'_n} z'_n \tag{3}$$

$$\tag{4}$$

By assuming image distance z' and object height h do not change over time, then

$$\frac{z_n}{z_k} \approx \frac{h'_k}{h'_n} \tag{5}$$

In addition, to account for object rotations about the optical axis, one may also consider measured width w', combining it with height to give measured area $A' = h'w'$. Now

$$z_n \approx z_k \sqrt{\frac{A'_k}{A'_n}} \tag{6}$$

which with Equation 7 gives the desired pan angle.

$$\theta_Z(\theta_P, d, z_0) = \tan^{-1}\left[\left(\cot(\theta_P) - \frac{d}{z_0}\right)^{-1}\right] \tag{7}$$

Like most vision tasks, switching between assisted control and autonomous control is an action naturally suited to humans. However, given the state of modern computer perception, particularly real-time perception, a single-camera has little 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

$$h'_P - \Delta \le h'_Z \frac{f_P}{f_Z} \le h'_P + \Delta \tag{8}$$

$$w'_P - \Delta \le w'_Z \frac{f_P}{f_Z} \le w'_P + \Delta \tag{9}$$

where Δ is the uncertainty of the tracker's results; w' is the measured width; h' is the measured length; f is the focal length; and subscripts Z and P represent the zooming and panoramic cameras, respectively. When both of these hold true, the *control arbitrator* may decide to allow camera Z to resume autonomous control.

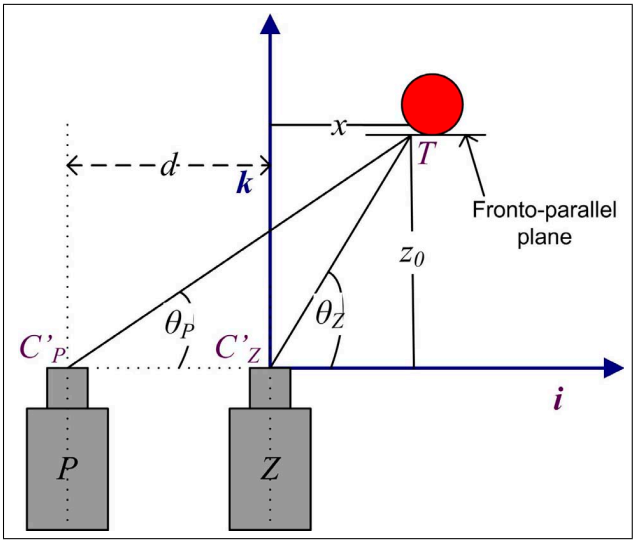

Fig. 1. Camera correspondence overhead view

4. EXPERIMENT

A system was constructed using a pair of Sony EVI-D100 cameras connected to a PC powered by two Pentium Xeon 2.8 GHz with Hyperthreading CPUs. Two Osprey 100/200 were used as a frame grabber, which have the capability of capturing 320 $\times$ 240 non-interlaced video at 30 fps. A radio-controlled car provides sporadic movements that demonstrate the advantages of having a second camera. In fact, the experiment could not be run using a single-camera method without nearly eliminating zoom.

Frames 90, and 362 of Figure 4 show large magnification in the zooming camera, while frame 225 shows a complete loss of fixation in the zooming camera, which correspond to the fast horizontal movements shown in Figure 2.

The experiment uses two cameras to guarantee stability, while using a digital zooming algorithm from [7] to produce scale invariance.

5. CONCLUSION

The goal of this work was to develop a system capable of capturing images of a specific size and position. Preceding methods used a single-camera system to implement zoom which created a tradeoff between maintaining fixation and maximizing resolution—in the event of lost fixation, the system is unstable. To split the responsibilities, a second camera with fixed focal length was introduced. The symbiotic relationship between the cameras ensured that resolution goals could be met while maintaining overall system stability.

By placing the cameras side-by-side—giving them similar views—the images captured were comparable, differing only by focal length. Using digital zoom, further magnification synchronized the images, permitting interchangeability. Separately, the cameras' original images differ in resolution and context, but together the system benefits from both.

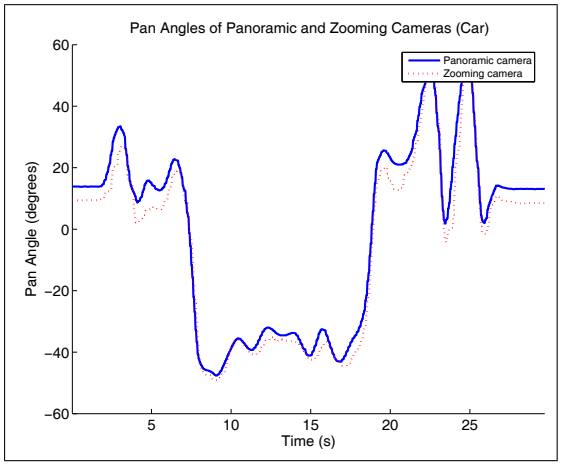

Fig. 2. Dual-camera pan angles

6. REFERENCES

[1] Evan Clark, "A multicamera system for gesture tracking with three dimensional hand pose estimation," M.S. thesis, Rochester Institute of Technology, 2006.

[2] Jonathan C. Rupe, "Vision-based hand shape identification for sign language recognition," M.S. thesis, Rochester Institute of Technology, 2005.

[3] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, 2002.

[4] B. J. Tordoff, *Active Control of Zoom for Computer Vision*, Ph.D. thesis, University of Oxford, 2002.

[5] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and H. Niemann, "Statistical modeling and performance characterization of a real-time dual camera surveillance system," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000, pp. 335–342.

[6] Qian Huang, Yuntao Cui, and Supun Samarasekera, "Content based active video data acquisition via automated cameramen.," in *ICIP (2)*, 1998, pp. 808–812.

[7] Eric D. Nelson, "Zoom techniques for achieving scale invariant object tracking in real-time active vision systems," M.S. thesis, Rochester Institute of Technology, 2006.

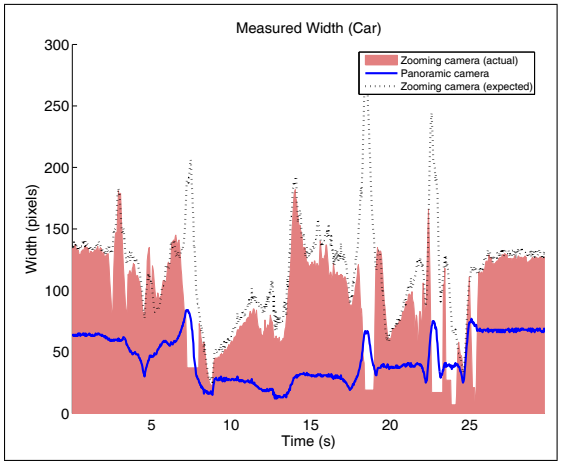

Fig. 3. **Control arbitration** When the expected and measured height do not match, fixation is lost since measured width in the panoramic camera is always accurate.

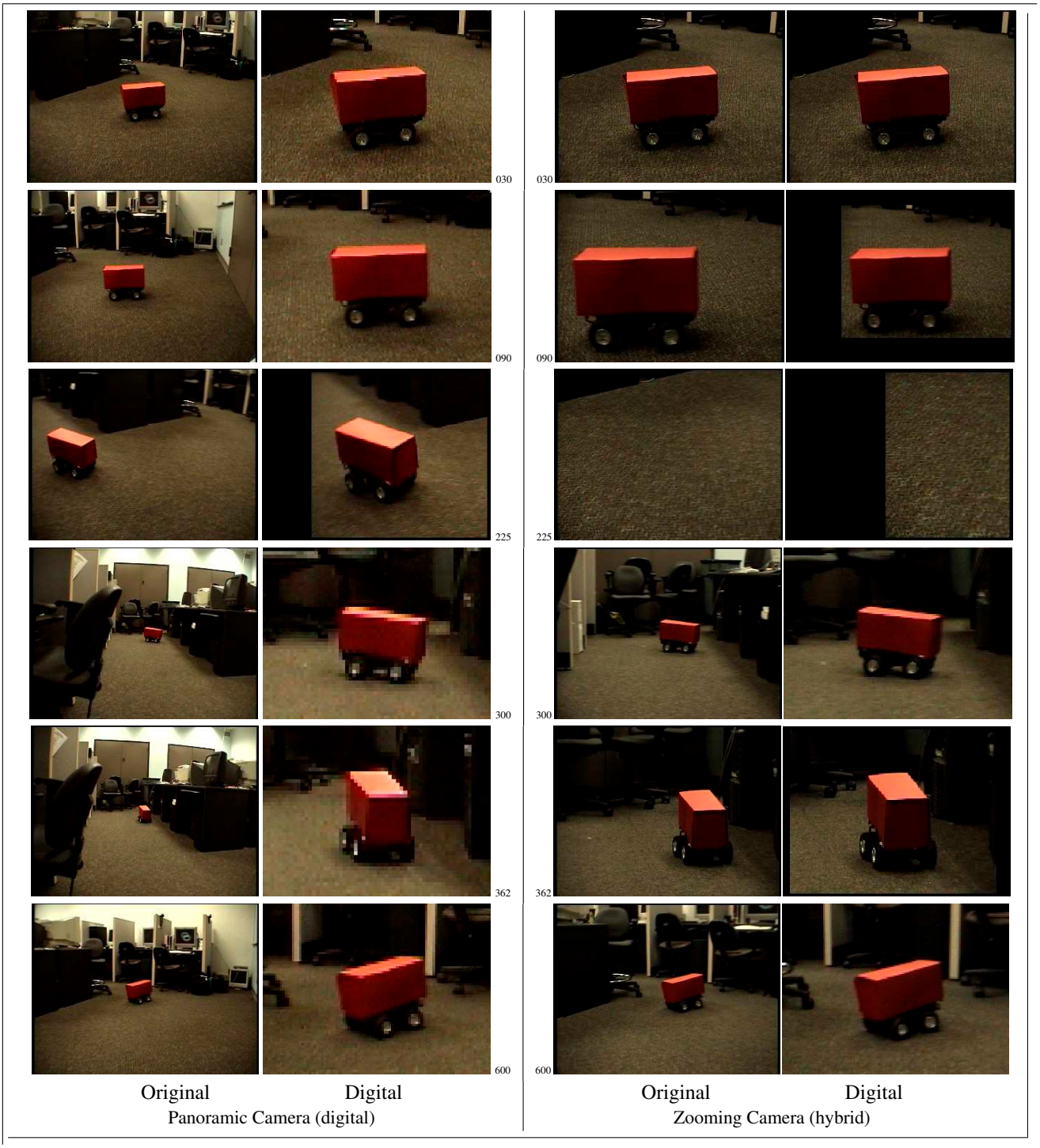

Fig. 4. **Two cameras tracking single car** The panoramic camera (left) always has the object in view, while the zooming camera (right) has improved resolution. When the object leaves the zooming camera's view, see frame 225, the panoramic camera is able to keep it in view. When the object is in view of both cameras, 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.