Correcting Unsynchronized Zoom in 3D Video

Colin Doutre[1], Mahsa T. Pourazad[1], Alexis Tourapis[2], Panos Nasiopoulos[1] and Rabab K. Ward[1]

[1]Department of Electrical & Computer Engineering
University of British Columbia
Vancouver, Canada

[2]Image Technology Research
Dolby Laboratories
Burbank, CA

***Abstract*—When capturing 3D video with a stereoscopic camera setup, it is important for the cameras to be precisely aligned and synchronized. This is particularly difficult in transitions such as zooming where the camera parameters must be changed in unison, or else the perceived 3D effect will be degraded. In this paper we study the problem of unsynchronized zooming in 3D video. First, we present a subjective study that shows that the perceived quality of stereo video is greatly reduced if the two views are zoomed by different amounts. Next, we present a method for correcting zoom mismatch by applying cropping and scaling to ones of the views. Our method involves finding matching points between the left and right views, and performing least-squares regressions to estimate the amount of scaling and cropping required to make the views consistent. Experiments were performed on videos with digitally introduced zoom mismatch and videos with optical unsynchronized zoom. In both cases the results show that our method is highly accurate and produces videos without size differences or vertical parallax between the two views.**

I. INTRODUCTION

The majority of 3D content is produced using a dual-camera configuration, generating a stereo pair with the left-eye and the right-eye views being separately recorded from slightly different perspectives. Capturing visually pleasing stereoscopic video requires that both the director and the camera operator are highly skilled regarding 3D geometry and camera calibration. One of the shooting parameters that can degrade the perceived 3D quality is unsynchronized zooming of dual cameras. It is difficult to synchronize the optical zooming of two identical cameras precisely. If the two cameras have different zoom factors, objects will have different size in the left and right views, and vertical parallax will be introduced which causes eye-strain and interferes with fusion of the two images [1].

The problem of the two views being zoomed by different amounts can be solved by applying digital cropping and scaling to one of the views to make it match the other. Doing that requires accurate estimation of the relative amount of zoom between the two videos. Several methods have been proposed for estimating zoom in monoscopic (single-view) video [2]-[4], which all involve relating the optical flow (i.e., the motion field) to camera parameters, such as zoom (focal length), translation and rotation. A common problem with these methods is separating the optical flow caused by changing camera parameters from the optical flow caused by object motion. Determining which parts of the motion field are affected by object motion is an ill-posed problem, and it affects the accuracy of zoom estimation methods. To correct unsynchronized zoom in stereo video, the zoom ratio between the two views needs to be estimated, but no previous work has addressed this problem.

In this paper, first, we present a subjective study on the impact of zoom mismatch on the perceived quality of 3D video. The results show that unsynchronized zoom severely degrades 3D video quality. Next, we present a method for correcting zoom mismatch by applying digital cropping and scaling to one of the views. In the case of a zooming in, the view that is originally zoomed in less is scaled to match the one that is zoomed in more, and in the case of zooming out, the view originally zoomed out more is scaled to match the one which is originally zoomed out less. To avoid the effect of object motion on zoom estimation, we use only vertical coordinates for estimating the zoom ratio between the views. This takes advantage of the constraint that there should be no vertical parallax between images captured with parallel cameras, which holds regardless of object motion of depth.

The rest of this paper is organized as follows. A subjective study showing the impact of zoom mismatch in stereo videos is presented in Section II. Our proposed method for correcting unsynchronized zoom is described in Section III. Experimental results are presented and discussed in Section IV. Finally, conclusions are given in Section V.

II. IMPACT OF ZOOM MISMATCH ON SUBJECTIVE 3D QUALITY

Consider the stereo geometry of two parallel cameras shown in Figure 1. A point with world coordinates (X,Y,Z) is projected onto the left and right image planes at image

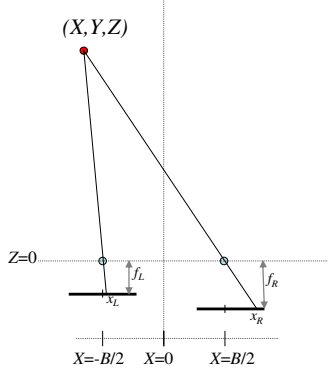


Figure 1: Stereo geometry with parallel cameras

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the British Columbia Innovation Council (BCIC).

coordinates $(x_{L,O},y_{L,O})$ and $(x_{R,O},y_{R,O})$. The subscripts L and R indicate the left and right images respectively, and the O subscript indicates that the coordinates are measured relative to the optical center of the camera. The cameras are separated by a baseline distance B and have focal lengths f_L and f_R. In a real camera the focal length is a property of the optical system (a higher focal length meaning more optical magnification), but even synthetic images are usually rendered with a virtual focal length in projecting a 3D model on to a virtual image plane. Using similar triangles, simple equations for the image coordinates can be found as follows:

$$x_{L,O} = f_L \frac{X + B/2}{Z} \qquad x_{R,O} = f_R \frac{X - B/2}{Z} \qquad (1)$$
$$y_{L,O} = f_L \frac{Y}{Z} \qquad y_{R,O} = f_R \frac{Y}{Z}$$

Ideally, the focal lengths of both cameras should be the same, i.e., $f_L = f_R$. In that case, $y_{L,O} = y_{R,O}$, so the images will have no vertical parallax, which is a very important requirement for 3D video. Vertical parallax in stereo images causes eyestrain, and if it is too large the human visual system will not be able to fuse the images at all, resulting in a loss or distortion of the 3D effect [1]. If the two initially synchronized cameras are zoomed by a different amount, it means that their focal lengths will be different. Consequently, there will be vertical parallax between the images, and objects will have a different size in each image; both results are highly undesirable in 3D video.

To evaluate this effect on viewers, we performed a subjective experiment on the impact of zoom mismatch in stereo video. In the test videos, the two views were zoomed in or zoomed out linearly and with the right view always having a larger scaling factor than the left (the difference was a constant value). To prepare the data set, the stereo images "Reindeer", "Baby1", "Aloe", and "Art" provided by [5] were used. To synthesize left-view streams with zoom-in/out effect, the left-eye image was digitally scaled (about its center) with the scaling factor being increased from 1X up to 1.8X of its original size. In each frame the scaling factor was increased 0.01X from that of the previous frame. Then the videos were zoomed back out to original size. The right-view had the same zoom pattern as the left view, only with the scaling factor higher by a constant amount so that unsynchronized zooming is simulated.

The test videos were shown to seventeen subjects on a 20" widescreen monitor. Most of the subjects had not participated in stereoscopic experiments before, and all viewers were naive to the underlying purpose of the experiment. Viewers rated the test stereoscopic video streams using the double-stimulus continuous-quality scale method based on the ITU-R Recommendation 500 [6], in which viewers wore Red/Blue anaglyph glasses and graded two versions of the same video stream (synchronized and unsynchronized zooming effect) announced as "A" and "B", from "Bad" to "Excellent". The subjects were not informed which one of the video streams had unsynchronized zooming effect. A number of zoom factor differences were tested, ranging from 0.05 to 0.2 (i.e., 5% to 20% size difference). For analysis, the ratings were digitized to range between 0 and 100 units. Figure 2 shows the results

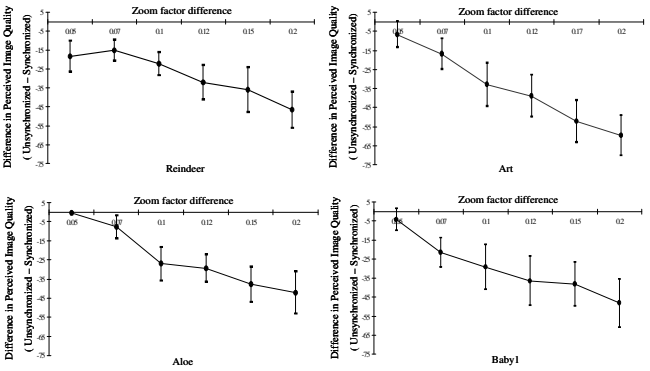


Figure 2: Subjective results: Rating is expressed as a difference between ratings for the unsynchronized and synchronized zoomed stereo videos. The error bars denote the 95% confidence intervals.

of the subjective test. The quality of the perceived 3D image degrades dramatically when the difference between the zooming factor of right and left view streams increases. These results show that zoom mismatch causes a large drop in perceived 3D quality, and should be corrected.

III. PROPOSED ZOOM CORRECTION METHOD

The problem of unsynchronized zooming in stereo video can be fixed by applying cropping and scaling to one of the views. Referring to the camera model of Fig. 1, the view with a shorter focal length has to be scaled to match the view with a longer focal length. If the cameras are zooming in, the view that has been zoomed less will have a lower focal length and should be scaled. If the cameras are zooming out, the view that has been zoomed out more will have a shorter focal length and should be scaled. For each frame in the videos, we estimate the amount of scaling that needs to be applied based on the y coordinates of matching points found between the two views.

To simplify notation, let us assume the right video has higher focal length, i.e. f_R is greater then f_L. According to equation (1), the left image could simply be scaled by a factor f_R/f_L, and the corresponding coordinates would be exactly the same as if the left image was captured with the same focal length as the right image. However, trying to implement this in practice is challenging. The image would have to be scaled about its optical center, which in general is not the same as the geometric center of the image [7]. The optical center of a camera changes as the camera zooms, so it is difficult to find the precise optical center of an image for every frame during zooming [7].

Instead of trying to directly estimate the optical center of each image and then apply scaling about that, we work in a coordinate system with the top left corner of the image defined as the origin. In this system, the optical center of the image is defined as (u,v) relative to the top left corner of the image. The coordinates of a point relative to the corner, which we will denote (x_L, y_L), are found simply by adding (u,v) to the coordinates in (1) that are relative to the optical center.

$$x_L = x_{L,O} + u_L \qquad x_R = x_{R,O} + u_R \qquad (2)$$
$$y_L = y_{L,O} + v_L \qquad y_R = y_{R,O} + v_R$$

Combining (2) and (1), we can find an expression that relates the y coordinates of the left and right images, with all the coordinates expressed relative to the image corners:

$$y_R = \frac{f_R}{f_L} y_L + v_R - \frac{f_R}{f_L} v_L \tag{3}$$

Equation (3) shows that we can apply a simple linear transform of the form:

$$y_L^{'} = s y_L + t_y \tag{4}$$

that will make the y coordinates of the left image match those of the right image (i.e., $y_L^{'} = y_R$), where s is a scaling factor $s = f_R / f_L$ and t_y is the amount of vertical translation $t_y = v_R - f_R v_L / f_L$. Estimating the parameters s and t_y is sufficient for scaling one image so that the images will have no vertical parallax. Therefore, we do not need to explicitly calculate the focal lengths or optical centers.

In order to estimate the parameters s and t_y, we find a number of matching points between the left and right images. There are many methods for finding matching points between images; we choose the Scale Invariant Feature Transform (SIFT) [8], as it is one of the most popular and reliable matching methods. Using SIFT feature matching provides a number of points (x_L,y_L) and (x_R,y_R) that match between the left and right images. A matrix equation relating these matching points to the scaling and translation parameters can be written as:

$$\begin{bmatrix} y_R \\ \vdots \end{bmatrix} = \begin{bmatrix} y_L & 1 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} s \\ t_y \end{bmatrix} \tag{5}$$

where each row in the matrices contains the data for one matching point. A standard linear least squares regression can be used to estimate s and t_y based on equation (5). With the parameters estimated, a simple scaling transform can be applied to the left image to make it match the right image:

$$\begin{bmatrix} x_L^{'} \\ y_L^{'} \end{bmatrix} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} x_L \\ y_L \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{6}$$

Applying equation (6) simply requires re-sampling the image, which can easily be done with standard methods such as bilinear or bicubic interpolation. One additional parameter is required in (6), the translation along the x-axis. This parameter is much less important than t_y, as any choice of t_x will give images without vertical parallax. We choose t_x so that equal amounts of cropping will be applied to the left and right of the image during scaling. To achieve this, t_x is chosen as:

$$t_x = \frac{W}{2}(1 - s) \tag{7}$$

where W is the width of the image.

The way we have calculated the parameter s, its value reflects the amount of scaling required if the left view were scaled to make it match the right view. Of course it is possible that instead the right view should be scaled to make it match the left view. We always want to scale up one of the images, because scaling down an image would result in there being missing data around the edges of the scaled image. Scaling up simply requires some image data be cropped from around the edges.

If the least squares regression gives a value of s less than one, it means we would have to scale down the left image to make it match the right one. In that case we actually want to scale up the right image. Simple rearranging of equation (4) shows that the appropriate scaling and translation values for modifying the right image are:

$$s' = \frac{1}{s} \qquad t_y^{'} = -\frac{t_y}{s} \tag{8}$$

where s and t_y are the parameters as estimated with the least squares regression based on (5), and s' and t_y' are the modified values that should be used for scaling the right image.

Note that the value of s is related to which camera has been zoomed more, depending on whether the cameras are zooming in or zooming out. If the cameras are zooming in, a value of $s<1$ indicates the left view has been zoomed in more (and hence objects appear bigger in it), and if $s>1$ then the right camera has been zoomed in more. If the cameras are zooming out, a value of $s<1$ would indicate the right camera has been zoomed out more, and hence objects appear bigger in the left view. A value of $s>1$ indicates the left camera has been zoomed out more, and objects appear larger in the right view.

Our complete algorithm can be summarized as follows. SIFT is used to find matching points between the left and right images, and a linear least squares regression is performed based on equation (5). If the regression produces $s>1$, then the left image is scaled and cropped with equation (6). If $s<1$, then equation (6) is applied to the right image with the modified parameter values of (8). This entire process is repeated for each temporal frame in the stereo video.

IV. Experimental Results

A. *Objective Results on Digitally Zoomed Videos*

In order to objectively measure the performance of the proposed method, we applied digital zooming to two standard test stereo videos, "soccer2" and "puppy". Both videos have resolution 720x480 pixels. We applied a different digital zooming pattern to the left and right view of each video, starting at original size and zooming in to a maximum zoom factor of 2X. The profile of the left and right zoom is

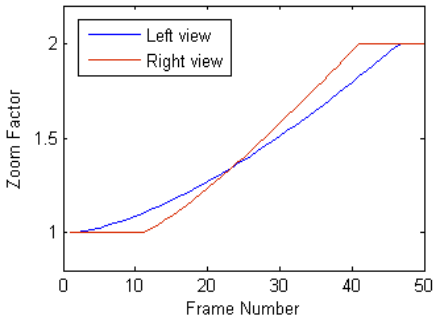
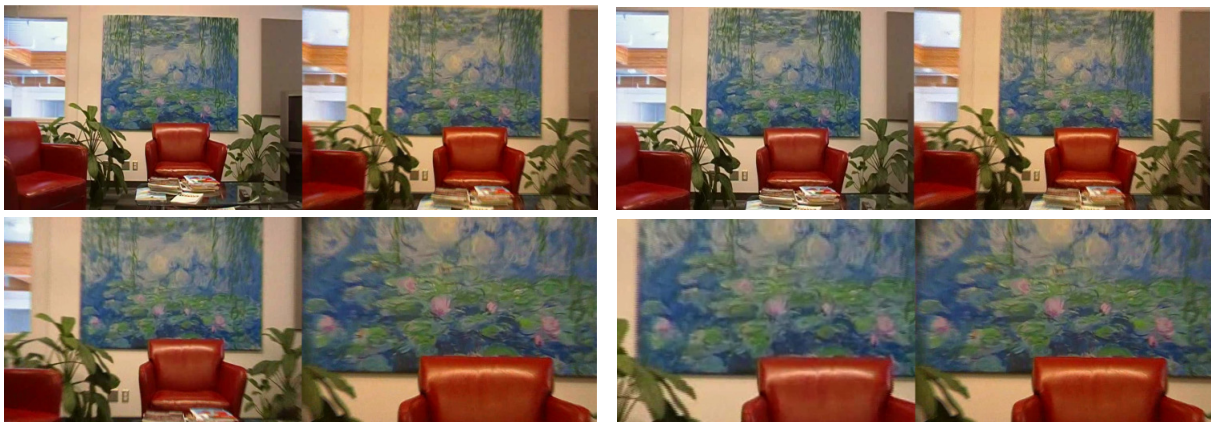


Figure 3: Digital zooming pattern applied to the left and right views for the objective tests



(a) (b)

Figure 4: Sample frames of the test video with unsynchronized optical zoom. (a) Captured left-right stereo pair (b) Corrected with proposed method.

illustrated in Figure 3. Since we applied digital zooming to these videos, the ground truth values for s and t_y are known.

We corrected the videos with our proposed algorithm, and measured the mean absolute difference between the estimated parameters and the known ground truth parameters. We also report the maximum vertical parallax that is introduced in each frame due to the error in the estimated parameters, which is calculated as:

$$\Delta y_{\max} = \max_{y\in[1,H]} \left| s_{gt}\, y + t_{y,gt} - \left(s_{est}\, y + t_{y,est} \right) \right| \qquad (9)$$

In (9) the parameters with the subscript '*gt*' are the ground truth parameters used in our experiment, and the parameters with an '*est*' subscript are those estimated with our method. The range of the argument '*y*' in (9) is from 1 to *H* (the image height), but the maximum of (9) will always occur at one of the endpoints. Therefore only the y=1 and y=*H* cases need to be evaluated to find the maximum.

In Table 1, we report the absolute difference of the estimated correction parameters, as well as the maximum vertical parallax. All values are averaged over the 50 frames of the zoom-in. From Table 1, we can see the proposed method is very accurate at estimating the correction parameters. The maximum vertical parallax introduced is well below one pixel for both test videos, which is below the limit of what is noticeable by the human visual system.

TABLE I. ACCURACY OF ESTIMATED CORRECTION PARAMETERS

Video	Mean Absolute Difference		Max Vertical Parallax, Δy_{max} (pixels)
	s	t_y	
soccer2	0.00038	0.127	0.236
puppy	0.00026	0.074	0.126

B. Results on 3D Video with Unsynchronized Optical Zoom

In the previous section, we showed that our proposed method works well on test videos where zoom mismatch was introduced synthetically through digital zooming. In order to test our method on video with optical zoom, we captured a stereo video pair with hand-controlled optical zoom. In the video, both cameras start at 1X zoom and zoom-in to a factor of just over 2X, and then zoom back out to 1X, lasting about 150 frames. The zooming was controlled by hand with no attempt to synchronize the two views. In Fig. 4a, the captured video shows clear zoom mismatch between the left and right views, and the 3D effect is lost during most of the zooming due to excessive size differences and vertical parallax. After correction with our method, the left and right views show no noticeable size differences or vertical disparity (Fig. 4b), and 3D effect is perceived during the zooming.

V. CONCLUSIONS

In this paper we propose a method for correcting unsynchronized zoom in 3D videos. For each frame, a set of matching points is found between the left and right views with the SIFT algorithm. A least squares regression is performed on the y coordinates of these matching points to determine which view needs to be scaled and to estimate the amount of scaling and translation needed to align the views. Experimental results show our method produces videos with negligible scale difference and vertical parallax.

REFERENCES

[1] A.Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," Proc. SPIE, vol. 1915, pp. 36–48, 1993.

[2] Yap-Peng Tan, Sanjeev R. Kulkarni and Peter J. Ramadge, "A New Method for Camera Motion Parameter Estimation," 1995 IEEE International Conference on Image Processing, Vol. 1, pp. 406-409, Oct 1995.

[3] I. Grinias and G. Tziritas. "Robust pan, tilt and zoom estimation," Int. Conf. on Digital Signal Processing, 2002.

[4] Y. P. Tan, S. R. Kulkarni, and P. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 133–146, Feb. 2000.

[5] D. Scharstein and R. Szeliski, Middlebury stereo vision page, http://vision.middlebury.edu/stereo.

[6] ITU-R Recommendation BT.500-10, Methodology for the subjective assessment of the quality of television pictures, 2000.

[7] R.G.Willson and S.A.Shafer, "What is the Center of the Image?," Journal of the Optical Society of America A, Vol.11, no.11, pp. 2946-2955, Nov. 1994.

[8] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, Nov. 2004.