CROSSTALK CANCELLATION IN 3D VIDEO WITH LOCAL CONTRAST REDUCTION

Colin Doutre and Panos Nasiopoulos

Department of Electrical and Computer Engineering, University of British Columbia
Vancouver, Canada
email: colind@ece.ubc.ca, panos@ece.ubc.ca
web: www.ece.ubc.ca/~panos

ABSTRACT

Many 3D displays suffer from noticeable crosstalk, where some of the light intended for one eye reaches the other one as well. Subtractive crosstalk cancellation, a technique where the images input to the display are modified to account for the crosstalk, can be an effective way to remove the appearance of crosstalk. However, to be effective, crosstalk cancellation requires raising the minimum image level above zero to leave enough 'foot-room' to allow for subtracting out the crosstalk from the other image. In this paper we propose a method for locally raising the image levels in regions that suffer from crosstalk. Our method involves detecting such regions in each frame and adding smooth patches of luminance around these regions. We apply temporal low pass filtering to the regions to prevent flickering, and also add fade-ins and fade-outs to regions of crosstalk that appear or disappear midway through the video to prevent sudden jumps or drops in luminance. Our method allows effective crosstalk cancellation, while maintaining better image contrast than globally scaling the image levels, and also prevents flickering that can occur with methods that operate on a frame by frame basis.

1. INTRODUCTION

Crosstalk is a critical factor that limits the quality of many 3D displays, as it can cause the viewer to perceive 'ghosting', an effect where a double image is seen. Crosstalk can severely degrade the perceived 3D image quality and can result in viewers not being able to fuse the two images. An overview of the sources of crosstalk in different 3D display technologies is presented in [1].

An effective method for reducing the appearance of crosstalk is through subtractive crosstalk cancellation. This is a technique where the image levels are lowered based on the anticipated amount of crosstalk. Therefore, after crosstalk is added during playback, the intended images will be seen by the viewers. Several methods have been proposed using this idea [2]- [4]. The same technique can be extended to multiview displays with crosstalk generated from several views [5].

A problem with crosstalk cancelation occurs if there is a high amount of crosstalk in an image region that is close to black. In this case, there may not be enough light to subtract from the image to compensate for the crosstalk. One solution to this problem is to raise the minimum image level - for example if the input images cover the entire range [0, 255], then compress the range to [50, 255]. This ensures there will always be 'foot room' for lowering the image values to compensate for crosstalk. This global approach of compressing the image range is used in several previous works [2][3][5]. A problem with this method is that it reduces the entire image contrast and therefore lowers the image quality.

Instead of globally raising the image levels, an alternative approach, which has been patented by the company realD in [4], is to raise the image levels only in local regions that suffer from noticeable crosstalk. In [4], they detect regions where conventional crosstalk cancellation will fail (i.e., where the signal is too low to be able to compensate for crosstalk) and around these regions a patch of 'disguising' luminance is added. These patches of added luminance are very smooth, and therefore likely to be less noticeable than crosstalk. Since the patches typically occupy a small percentage of the total image area, the method preserves image contrast better than globally compressing the image range. However, the method as described in [4] is operated on a frame by frame basis, without considering temporal consistency. Therefore, it may often result in flickering or sudden jumps in brightness.

In this paper we propose a method that considers temporal consistency when adding luminance to local image regions that suffer from crosstalk. Our method involves temporal filtering to remove flicker, and providing fade-ins and fade-outs for regions that appear or disappear over the course of the video. The rest of the paper is organized as follows. In section 2, we provide a description of a traditional crosstalk cancellation algorithm. In section 3 we present our proposed method. Results and discussion are presented in sections 4, and 5. Section 6 concludes the paper.

2. CROSSTALK CANCELLATION

In this section we provide a description of crosstalk cancellation. We model the crosstalk in a stereo display as follows:

$$\begin{aligned} i_{L,eye}(x,y) &= i_L(x,y) + c \cdot i_R(x,y) \\ i_{R,eye}(x,y) &= i_R(x,y) + c \cdot i_L(x,y) \end{aligned} \quad (1)$$

Here $i_L(x,y)$ and $i_R(x,y)$ are one colour channel of the input images in linear space, and c is the amount of crosstalk. The signals $i_{L,eye}$ and $i_{R,eye}$ are the ones reaching the viewer's eyes. A simple gamma transformation can convert the input images from 8-bit gamma encoded values to linear values (although an alternate display model could also be used). Throughout this paper we assume the images are represented as linear values in the range zero to one. We also assume the crosstalk is uniform throughout the image, but it is straightforward to extend the model to allow for a spatially varying crosstalk signal or different amounts of crosstalk for the different colour channels. The crosstalk can be compensated for if we input the following processed images to the display:

$$i_{L,disp}(x,y) = \frac{i_L(x,y)}{1-c^2} - \frac{c \cdot i_R(x,y)}{1-c^2}$$
$$i_{R,disp}(x,y) = \frac{i_R(x,y)}{1-c^2} - \frac{c \cdot i_L(x,y)}{1-c^2} \quad (2)$$

By substituting equation (2) into the crosstalk model of (1), it can easily be verified that the images reaching the viewer's eyes will be the original images, i_L and i_R. Note that (2) basically involves subtracting the right image from the left one and vice versa. However, a problem occurs if one of the images has some low values (i.e., values close to black) where the other image has high values. In this case, equation (2) may give negative results for one of the images. If that occurs, it means that there is not enough light to subtract from in order to compensate for the crosstalk from the other image. The following condition is necessary in order for equation (2) to work without negative values occurring:

$$i_S(x,y) \geq c \cdot i_C(x,y) \quad (3)$$

Since we treat the left and right images equally, we will refer to one image as the signal image (i_S), and the other as the crosstalk image, (i_C). A simple way to ensure that the condition in (3) is always met, is to raise the minimum image level from zero to c times the maximum level, i.e., compress the range of the image from [0,1] to [c,1]. However that decreases the image contrast, which severely degrades the image quality for even moderate amounts of crosstalk. An alternative, which is proposed in the patent described in [4] and we use here, is to raise the image levels only in local regions that suffer from severe crosstalk, rather than across the whole image.

3. PROPOSED METHOD

In order to reduce ghosting, we need to raise the levels of the image in regions where crosstalk is visible and there is insufficient luminance to subtract from using equation (2). We follow the idea from [4], where patches of "disguising" luminance are added to the image regions where crosstalk cannot be corrected and is likely to be visually disturbing. Since the patches of luminance are very smooth, they will be less noticeable and disturbing than visible crosstalk. Each image is altered by adding a smooth signal, $\alpha(x,y)$:

$$i_\alpha(x,y) = i(x,y) + \alpha(x,y) \quad (4)$$

Here, $i(x,y)$, represents any one of the red, green and blue colour channels of the image. The same signal $\alpha(x,y)$ is added to all channels in order to avoid altering the colours of the pixels too much. Two versions of $\alpha(x,y)$ need to be generated; one for the left image and one for the right image. To construct these signals we need to determine the regions of each image where crosstalk cannot be compensated for using (2), and then generate a smooth signal that will raise the luminance in those regions enough for effective crosstalk compensation. For clarity, we will first describe our algorithm for still images and then describe its extension to video.

3.1 Algorithm for still images

We can calculate the amount each pixel has to be raised, in order to meet the condition in equation (3) as follows:

$$R_K(x,y) = \max\left(0, c \cdot i_{C,K}(x,y) - i_{S,K}(x,y)\right) \quad (5)$$

Here K represents the colour channel (R, G, or B), and $R_K(x,y)$ is the amount that a colour channel needs to be raised for effective crosstalk cancellation. The result is clipped to zero because a negative value indicates the sample already has sufficient luminance. Since we will add the same signal to all three colour channels, we generate a single value for the amount each pixel has to be raised by taking the maximum over the three colours:

$$R(x,y) = \max\left(R_R(x,y), R_G(x,y), R_B(x,y)\right) \quad (6)$$

At this point, R contains all the pixels for which equation (2) would fail if it were applied to the original images, i.e., all the pixels that would still have some crosstalk even after cancellation. An example of the signal $R(x,y)$ for the left and right images of a stereo pair is shown in Fig. 1b. In many of these pixels the crosstalk may not be visually noticeable and thus raising the luminance in those areas would be unnecessary. Therefore, we apply additional processing to determine which areas of the image need luminance added to them. First we set to zero all the pixels in $R(x,y)$ that are below a threshold (we use 1% of the maximum display luminance). Then we remove small regions, which are less visually noticeable, by eroding and dilating with a circular mask of 8 pixels (as in done in [4]).

We then divide the signal $R(x,y)$ into a number of connected regions with the binary labelling algorithm presented in [6] (Fig. 1c). For each of these regions we will add a patch of smoothly varying luminance that has its maximum value at the pixels in the region and gradually decreases for the surrounding pixels based on their distance from the region. The patch for one region is calculated as:

$$\alpha_j(x,y) = M_j \cdot \max\left(0, \frac{w - d(x,y)}{w}\right) \quad (7)$$

In (7), j represents the region label, M_j is the maximum value of $R(x,y)$ of any of the pixels in the region, w is the width of the transition region, and $d(x,y)$ is the distance of

pixel (x,y) from the region. The distance can be calculated efficiently in $O(N)$ time with the algorithm described in [7]. Equation (7) uses a linear ramp for the transition region but other smoothly decreasing shapes would also work, such as a Gaussian or sigmoid. For HD resolution videos, an appropriate transition width (w) is 200 pixels. A larger width will result in the transition being less noticeable, but also a larger portion of the image having lower contrast.

Since we are adding the patch of luminance given by equation (7) to one of the images (left or right) in the stereo 3D pair, we have to add a corresponding patch to the other image to prevent retinal rivalry. To do this, we first calculate the centre-of-mass of the region and then perform block matching with a large block size (e.g., 16 pixels) to estimate the disparity at the centre of the region. Then, we create a copy of the signal $\alpha_j(x,y)$ shifted in the x-direction by the estimated disparity that will be added to the other image. An example of this is shown in Fig. 1 (d) and (e). In Fig. 1d, the patches $\alpha_j(x,y)$ are shown for the left and right images. In Fig. 1e shows the patches from the other image shifted by the estimated disparity for each region.

The entire process described here is performed twice, once considering the left image as the signal (i_S) and once considering the right image as the signal. In order to generate the final luminance that will be added to each image we take the maximum of all the individual patches, $\alpha_j(x,y)$, both the ones calculated based on the current image (left or right) and the ones from the other image that are shifted based on the disparity. An example of this is shown in Fig. 1f.

Finally, the smooth signal $\alpha(x,y)$ is added to the input images (equation (4)). Then, conventional crosstalk cancellation can be applied with equation (2), and it will be more effective since the images now have been raised in areas that suffer from noticeable crosstalk. If required (for displaying purposes), the final images can be converted from the linear space to a gamma encoded 8-bit space.

3.2 Extension to video sequences

In the previous section, we have described an algorithm for adding local patches of luminance to images to ensure there is enough 'foot room' for effective crosstalk cancellation. If this method were applied to video sequences on a frame by frame basis, annoying flickering would occur as temporal consistency is not considered. In this section of the paper, we describe an extension of the previous method to video sequences, which includes temporal filtering of the patches, removing patches of short temporal duration, and fading patches in and out when regions of uncorrectable crosstalk appear or disappear over the course of the video. We will describe a non-causal version of the algorithm that assumes the entire video is available during processing, but later we will comment on how it could be adapted for a causal real-time application.

First we detect significant regions of uncorrectable crosstalk in every frame of the video, following the same process as in the still image case. That is, equations (5) and

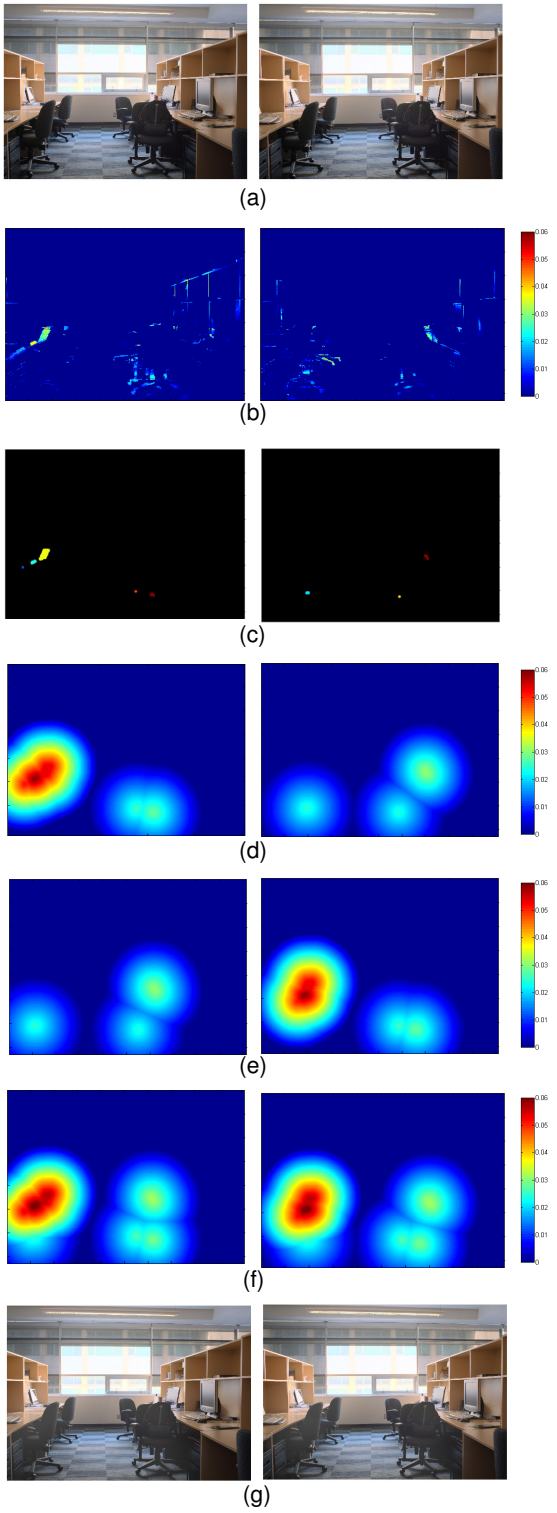

Figure 1: The steps of our method for still images. **(a)** Original images (left and right) **(b)** Amount each pixel needs to be raised, calculated with equations (4) and (5), **(c)** Labelled regions that remain after thresholding, erosion and dilation **(d)** Luminance patches for the above regions, calculated with equation (7), **(e)** Patches from the other image, shifted by the estimated disparity for each region **(f)** The final smooth signal that will be added to each image (pixel-wise maximum of the above two signals) **(g)** The output images with the added patches of luminance, calculated as images (a) plus signals (f)

(6) are used, followed by thresholding, erosion, and dilation. For each detected region, we calculate its centre-of-mass and the amount the region needs to be raised (the value of M_j in equation (7) for the region).

Next we match regions in temporally adjacent frames. Starting at the first frame and progressing through every frame in the video, we match regions with regions in the next frame. For each region in frame *N* we compare its centre to those of regions in frame *N*+1. If the distance between the centre of the region in frame *N* and the centre of a region in frame *N*+1 is less than 20 pixels, the regions are considered a match and 'linked' together. We implement this linking by having a data structure for each region that contains pointers to the matching regions in the next and previous frames. If more than one region meets the matching criterion, we choose the one with the lowest distance from centre to centre. If no match is found in frame *N*+1 for a region in frame *N*, we make a copy of the region in frame *N*+1. This copying serves two purposes; it allows us to fade out luminance patches, preventing a visually noticeable sudden drop in luminance, and it also sometimes allows us to fill in temporal gaps for regions that are not detected in one or more frames, but then later are detected again. When copying a region from frame *N* into frame *N*+1, we decrease its value for M_j by a small amount so that the region will fade out over time if it does not appear again in the video (we use 0.1% of the display luminance, which typically makes regions fade out over 2-3 seconds).

Next, we eliminate regions that have a very short temporal duration, as viewers are not likely to notice crosstalk that appears for a short amount of time. To achieve this, we count how many frames each group of regions was linked to in the previous stage (not counting any copied regions). If the count for a group of temporally linked regions is less than one second worth of frames (i.e., 30 for 30 fps video), then we delete all the regions in the group.

Afterwards, we identify regions that appear for the first time midway through the video. To prevent a sudden jump in luminance when a region first appears in time, we apply a fade-in to these regions. To achieve this, we perform a pass through the frames, checking for regions in frame *N* that do not have a backward link to a region in frame *N*-1. If any such regions are found, we copy the region into the previous frames, first into frame *N*-1, then *N*-2, and so on. Each time a region is copied, its value for M_j is decreased a small amount so that the region will fade in. As with the fade-out case discussed earlier, we use 0.1% of the display luminance for the step size of decreasing M_j, to achieve a fade transition time of 2-3 seconds.

At this stage, we have a series of regions that are linked temporally to regions in other frames. To prevent flickering, we apply temporal filtering to ensure that the amount each region is raised by (M_j) is consistent between frames and changes very slowly over time. According to [8], applying a low pass filter with cut-off frequency 0.5 Hz is sufficient to eliminate flicker in video (based on the temporal frequency response of the human visual system). Therefore, we design an 80 tap a low pass filter with cutoff frequency 0.5 Hz, and apply this filter to values of M_j for each region over time.

After filtering, we calculate a patch of luminance for each region using equation (7), only this time using the temporally filtered versions of M_j. As in the still image case, we perform a disparity search using block matching for each region, and generate a shifted version of each patch for the other image (left or right). To save computations, a smaller disparity search can be used for most frames, using the disparity of the region in the previous frame as the initial estimate (and searching for example +/- 2 pixels). The rest of the process is the same as described for the still image case. The final version of $\alpha(x,y)$ for each frame is calculated by taking the maximum of all the individual patches, and is then added to each image. After that, conventional crosstalk cancelation can be performed. If required, the image can be converted from the linear space to a gamma encoded representation.

4. RESULTS

We tested our method on several stereo videos captured with a pair of parallel cameras. The capturing setup is described in detail in [9]. The videos had a resolution of 1280x720p and were 30 fps.

In Fig. 2, we give an example of how a frame will look using crosstalk cancellation and different methods for raising the image levels. In Fig. 2a, we show the original left view of one frame of video. Fig. 2b shows what the left view will look like with 5% crosstalk from the right view, using crosstalk cancellation based on equation (2). Crosstalk results in an extremely annoying double edge appearing along the person's left side. Since the image levels are low across the person's black clothing, there is not enough light available when performing the cancellation. Raising the minimum image level globally, as shown in Fig. 2c, results in effective crosstalk cancellation, but at the expense of lowering the contrast and hence degrading the image quality. Our proposed method, where patches of luminance are added locally to regions that suffer from crosstalk, results in effective crosstalk cancellation while retaining better image contrast, as seen in Fig. 2d.

To illustrate how our method improves temporal consis-

(a) (b) (c) (d)

Figure 2: Illustration of crosstalk reduction with local and global level raising. (a) Original left image with no crosstalk (b) Image with crosstalk reduction but no level raising (c) Global raising of minimum image level (d) Proposed method with local raising

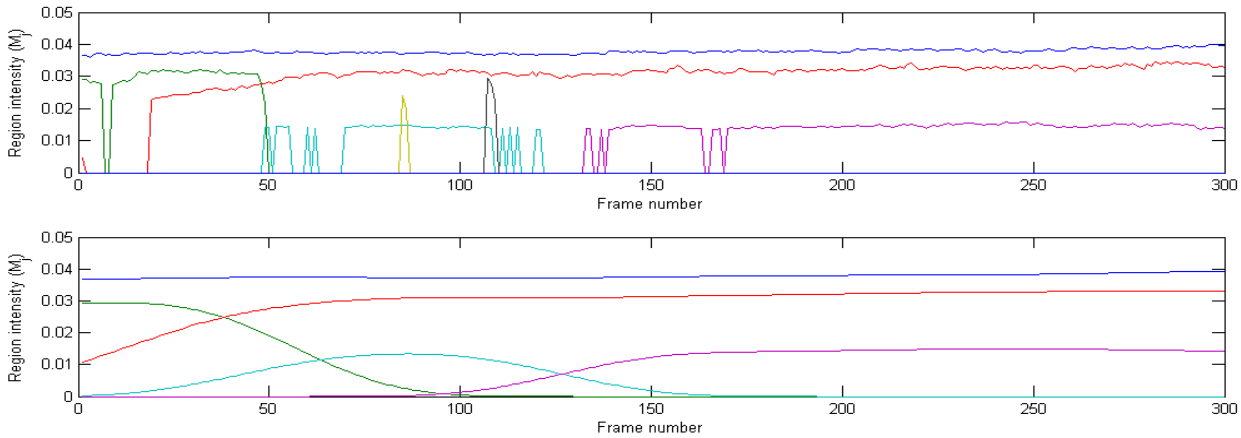

Figure 3: Regions for a video calculated frame by frame (above) and with our proposed method (below)

tency, we plot the intensity of the patches added over the course of a 10 second video when using our proposed method and compare to calculating the regions on each frame independently (Fig. 3). As seen in the top plot of Fig. 3, when the regions are calculated independently on each frame, they can be quite inconsistent from frame to frame. Sometimes a region is not detected in every frame, and a few regions are only detected in one or two frames. This results in annoying and highly noticeable flickering in the output video. Our proposed method gives much smoother temporal transitions, and avoids any flickering (Fig. 3 bottom).

5. DISCUSSION

In this paper we have described a non-causal method. Our algorithm could be modified to be real-time and causal, so that it could be applied to situations such as 3DTVs. Obviously, the filter would have to be replaced with a causal filter, and fewer taps may need to be used. When a new region is detected that was not in the previous frame, a fade-in for the new region would have to start at that frame (as opposed to fading in during the previous frames). A faster fade-in time would need to be used since crosstalk would be visible during the fade-in time. Also, deleting regions that appear only a short amount of time would not be possible, since when a new region is detected, it is not known how long it will last.

It should be noted that our algorithm could be applied using different crosstalk models than the linear one of equation (1), such as the model in [3], which is based entirely on visual measurements. To use a different crosstalk model, equation (5) would simply need to be replaced with a different function that gives the minimum amount of light required to compensate for the crosstalk from the corresponding pixel in the other image. It is also trivial to adapt our method to allow for different amounts of crosstalk in the red, green and blue channels or a spatially varying crosstalk signal.

6. CONCLUSION

We have proposed a method for locally adding patches of luminance to videos in order to improve crosstalk cancellation. Our method considered temporal consistency by applying low pass filtering to detected regions over time, removing regions of short duration, and creating fade in and fade outs for regions that appear or disappear mid way through the video. Our method allows crosstalk to be effectively corrected, while maintaining good image contrast and avoiding problems with flickering.

REFERENCES

[1] A. J. Woods, “Understanding Crosstalk in Stereoscopic Displays” (Keynote Presentation) at Three-Dimensional Systems and Applications (3DSA) conference, Tokyo, Japan, 19-21 May 2010.

[2] J. S. Lipscomb, W. L. Wooten, “Reducing crosstalk between stereoscopic views,” Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems, vol. 2177, pp. 92-96, February 1994.

[3] J. Konrad, B. Lacotte, E. Dubois, “Cancellation of image crosstalk in time-sequential displays of stereoscopic video” in IEEE Transactions on Image Processing, vol. 9, no. 5, pp. 897-908, May 2000.

[4] D. J. McKnight., “Enhanced Ghost Compensation for Stereoscopic Imagery,” U.S. Patent 0 040 280, Feb. 2010.

[5] M . Barkowsky , P. Campisi, P . Le Callet , V. Rizzo, “Crosstalk measurement and mitigation for autostereoscopic displays,” Proc. SPIE 3D Image Processing and Applications, San Jose, USA, 2010.

[6] R. M. Haralick, and L. G. Shapiro. Computer and Robot Vision, Volume I. Addison-Wesley, 1992, pp. 40-48.

[7] H. Breu, J. Gil, D. Kirkpatrick, and M. Werman, "Linear Time Euclidean Distance Transform Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 5, May 1995, pp. 529-533.

[8] R. Mantiuk, S. Daly, and L. Kerofsky, “Display adaptive tone mapping,” ACM trans. on Graphics, vol. 27, no. 3, pp. 68, July 2008.

[9] D. Xu, L. Coria, P. Nasiopoulos, "Guidelines for Capturing High Quality Stereoscopic Content Based on a Systematic Subjective Evaluation," IEEE International Conference on Electronics, Circuits, and Systems, ICECS 2010, pages 166-169, Dec. 2010.